AI-assisted analysis of phonological variation in English

Special session on Deep Phonology, AMP 2025
UC Berkeley September 27, 2025

Virginia Partridge, Joe Pater, Parth Bhangla,
Ali Nirheche and Brandon Prickett
{vcpartridge, pater}@umass.edu

UMassAmherst

These slides, and references, are available at https://websites.umass.edu/pater/handoutsslides/

Background

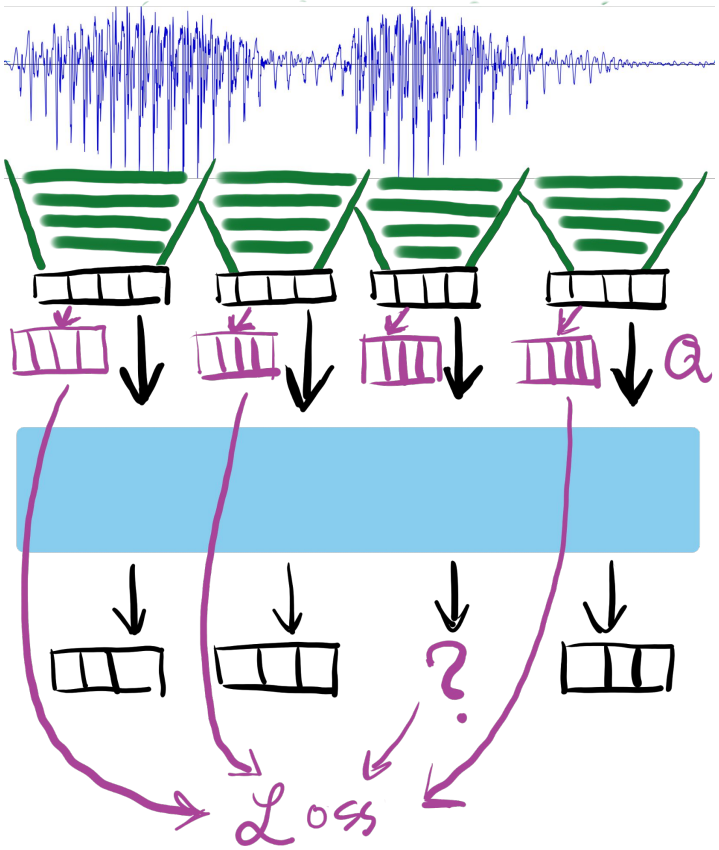
Despite the dramatic recent advances in speech recognition technology, automated phonetic transcription is not widely used in linguistics (as far as we know). This is especially surprising since the paper introducing Wav2Vec 2.0 (Baevski et al. 2020) presents SOTA results on a TIMIT phone recognition benchmark.

Our goal is to increase the usefulness of automated phonetic transcription for the study of phonological variation in English (variation within and across speakers and varieties)

In this paper, we:

- Situate our work within current research on automated transcription
- Present a Wav2Vec 2.0 model fine-tuned on Buckeye (Pitt et al. 2005), and compare it to other models, using TIMIT (Garofolo et al. 1993) as a test set
- Provide a web-based implementation, with Praat textgrid input and output
- Discuss next steps (and also issues with the standard TIMIT benchmark)

Wav2Vec 2.0 Pre-training: Sounds to vectors



Convolutional Layers: Snapshots of what's happening in audio at a particular point in time

Transformer Layer: Take into account relative positions of sounds in context

Masks + Quantization + Loss Function: Perform *self-supervised learning* by hiding sound representations that must be predicted using a probabilistic function

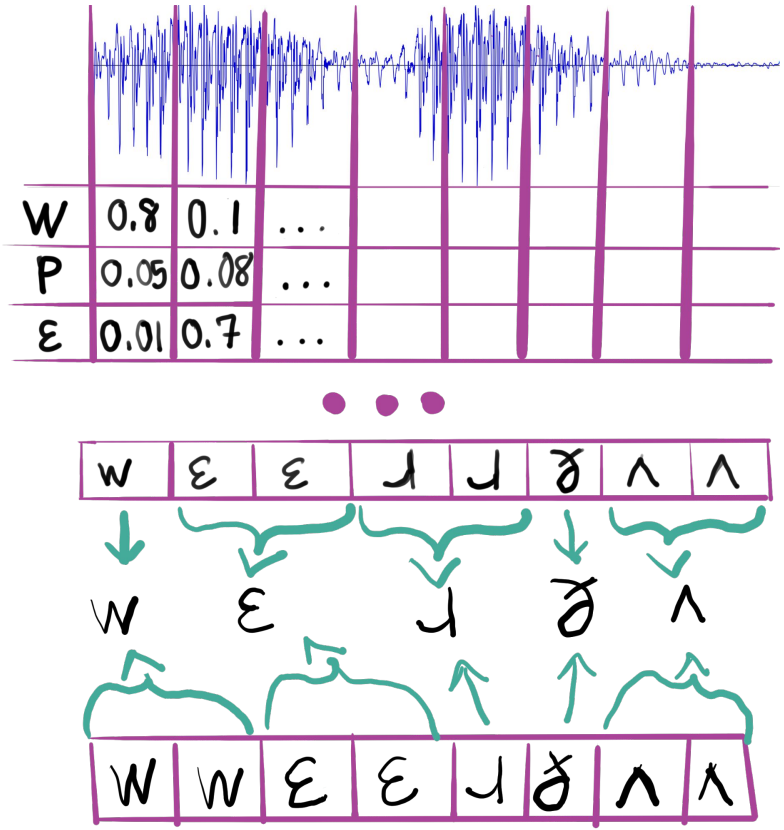
Facebook Research has open sourced many many pre-trained models for others to use.

Fine-tune: Add phone prediction

Bring your own audio with transcriptions.

Using Connectionist Temporal Classification loss, **maximize the probability of the correct output** by:

1) Predicting each symbol's **probability at each time segment**
2) **Merging** repeated symbols across neighboring time segments
3) Maximize the marginal probability of **alignments** that lead to the correct output



Multilingual automated transcription

The primary focus of contemporary research on automated phonetic transcription is the development of universal models, trained on data from multiple languages, and applicable in principle to any language.

This research includes fine-tuned Wav2Vec 2.0 models (Xu et al. 2021, Taguchi et al. 2023) as well as the quite different Allosaurus (Li et al. 2020), which emphasizes the distinction between phoneme and phone recognition.

The latest contribution to this line of research is Zhu et al. (2025). It advances the SOTA on benchmarks, but also recognizes that “error analysis reveals persistent limitations in modeling socio-phonetic diversity, underscoring challenges for future research”

Zhu et al. illustrate these limitations by testing the model on the Buckeye English corpus. We’ll be doing this for multilingual models as well, but not with theirs, since we were unfortunately unable to get it to run.

Grapheme-to-phoneme vs. actual transcriptions

Multilingual transcription leverages the availability of large amounts of orthographically transcribed speech by applying grapheme-to-phoneme conversion (e.g. Epitran; Mortensen et al. 2018) to obtain phonetic transcriptions.

These are similar to dictionary transcriptions, and are of course not equivalent to transcriptions of the individual utterances:

- The pronunciation of individual words can vary across speakers, across utterances by an individual speaker, and across phonetic and phonological contexts
- The pronunciation of phonemes can differ within words. Allosaurus also uses phoneme-to-allophone conversion, but the transcriptions are still relatively abstract

AutoIPA

We call our model “AutoIPA” (AI-assisted in our title is thus ambiguous)
It is a Wav2Vec 2.0 pre-trained model (facebook/wav2vec2-large-xlsr-53) fine-tuned on the Buckeye corpus:

- 40 speakers, balanced for gender and age (over 40 vs. under 40)
- White residents of the greater Columbus area (Northern Midland)
- 30 to 60 minutes of conversational speech for each speaker - about 20 hours total
- Phonemic transcription + vowel nasalization, flap and glottal/glottalized stop (we adopt the Seyfarth and Garellek 2020 revisions), includes syllabic sonorants
- No distinction between stressed/unstressed ʌ/ə or ɝ/ɚ (“perceived quality alone” Kiesling et al. 2006: 18; see relatedly Lindsay 2022).

As far as we know, this is the first time Buckeye has been used in the training of an automated transcriber, though Buckeye and TIMIT have been used in training phone alignment models (e.g. Kreuk et al. 2020), and Zhu et al. use it as a test set.

Processing the Buckeye data

Data from 24 speakers used for training, 8 for development, and 8 for testing

- Each demographic is represented in equal proportion across these data sets
- E.g., 6 younger women, 6 older women, 6 younger men, and 6 older men in training
- Success on test data requires generalization to new speakers

Training data divided into samples 8-12 seconds in length for Wav2Vec 2.0 fine-tuning

ARPABET transcriptions converted to IPA; the symbol set with our ARPABET to IPA mappings can be seen here. We used Phonecodes (Hasegawa-Johnson 2019): our GitHub fork, our Python package

The task is phone (rather than feature) prediction, affricates and diphthongs treated as single phones, and nasalization and syllabicity are properties of individual phones

Evaluation Metrics: Phone Error Rate (PER)

What's the **ratio of phones** we get wrong in each test sample?

1) Compute the edit distance between your prediction *p* and your reference *r* to find the total number of phone insertions, substitutions or deletions between them.
2) Normalize by the number of phones in the reference *r.*

$$per(p, r) = \frac{edit\_distance(p,r)}{length(r)}$$

per([bop], [pop]) = ⅓
per([bop], [po]) = 1
per([bip], [po]) = 1.5

LOWER IS BETTER!

Red = Deletion, Blue = Insertion, Yellow = Substitution

Evaluation Metrics: Phone Feature Error Rate (PFER)*

How many **articulatory features** are we getting wrong?

1) Compute edit distance to find phone substitutions, deletions and insertions between your prediction *p* and your reference *r*.
2) Sum up the total cost of edit distance errors between aligned phones:
   - Phone deletions and insertions each cost 1
   - For substitutions, each feature mismatch costs 1/24 (there are 24 features in the table). This is the normalized Hamming distance between the articulatory features of the phones.

pfer([bo], [po]) = 0.04166
pfer([bop], [po]) = 1.04166
pfer([pɛ], [po]) = 0.125

LOWER IS BETTER!

* Misleadingly **not really a rate**, but does obey the triangle inequality

Evaluation Metrics: Implementation

We use the PanPhon Python library (Mortensen et al. 2016) to calculate phone and feature edit distances.

For ease of use and reproducibility, we've published a wrapper around PanPhon's distance metrics that is compatible with the HuggingFace evaluate package. This makes them easy to compute when working with standard Python libraries for transformer-based speech recognition models.

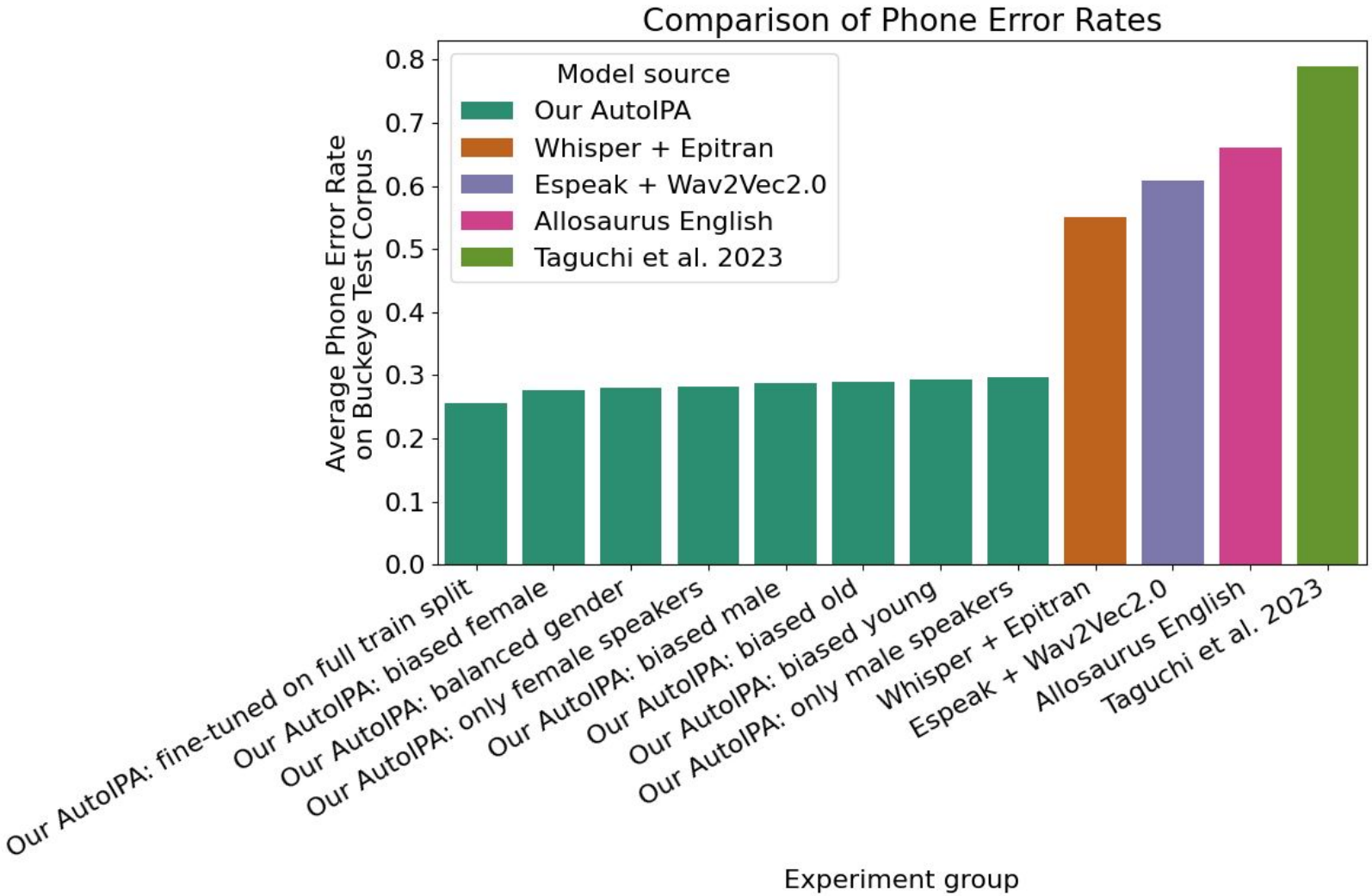
The results we report for models here are average PER and PFER across samples in the test corpus.

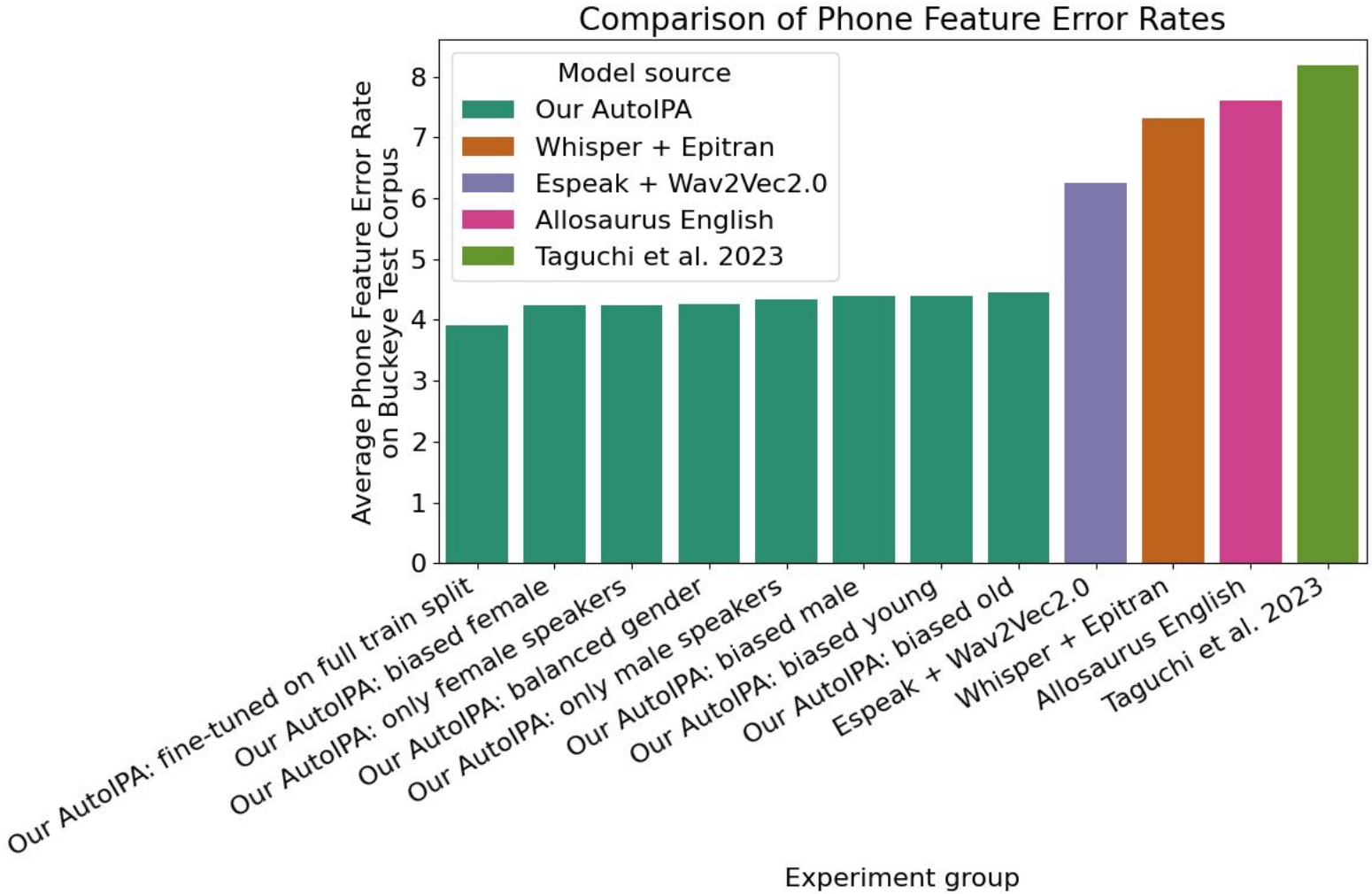
A caution on model comparison

It is completely expected that AutoIPA will outperform the multilingual models!

But we need *some* way of relativizing the performance of AutoIPA

Future work might take our results as a benchmark to improve on


Comparison of Phone Error Rates
Model source
Our AutoIPA
Whisper + Epitran
Espeak + Wav2Vec2.0
Allosaurus English
Taguchi et al. 2023
Average Phone Error Rate on Buckeye Test Corpus
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Our AutoIPA: fine-tuned on full train split
Our AutoIPA: biased female
Our AutoIPA: balanced gender
Our AutoIPA: only female speakers
Our AutoIPA: biased male
Our AutoIPA: biased old
Our AutoIPA: biased young
Our AutoIPA: only male speakers
Whisper + Epitran
Espeak + Wav2Vec2.0
Allosaurus English
Taguchi et al. 2023
Experiment group


Comparison of Phone Feature Error Rates
Model source
Our AutoIPA
Whisper + Epitran
Espeak + Wav2Vec2.0
Allosaurus English
Taguchi et al. 2023
Average Phone Feature Error Rate on Buckeye Test Corpus
0
1
2
3
4
5
6
7
8
Our AutoIPA: fine-tuned on full train split
Our AutoIPA: biased female
Our AutoIPA: only female speakers
Our AutoIPA: balanced gender
Our AutoIPA: only male speakers
Our AutoIPA: biased male
Our AutoIPA: biased young
Our AutoIPA: biased old
Espeak + Wav2Vec2.0
Whisper + Epitran
Allosaurus English
Taguchi et al. 2023
Experiment group

Some example transcriptions

The following are examples of transcriptions of utterances from the Buckeye test set by the models discussed here. The orthographic and phonetic transcriptions at the top are from Buckeye.

We have inserted spaces between words to increase readability.

s3802a_Utt33 (Older male)

where the trees were mature and they didn't really ruin our development and cut down a lot of the large trees

wɛɹ ðɪ tɹiz wɹ̩ mʌtʃuɹ ɪn neɪ dɪrn̩ ɹɛli ɹuɛ̃ ɔɹ dɪvɛlʌpmɛnt n̩ kʌt daʊr̃ ɑ lɑrʌv ðʌ lɑɹdʒ tɹis

wɛɹ ðɛ tʃɹiz wɹ̩ mʌtʃʊɹ ɛn neɪ dɪrn̩ ɹɛli ɹuʌn ɑɹ dɪvɛlpmɛʔ n̩ kʌt daʊr̃ ʌ lɑrʌ ðʌ lɑɹdʒ tʃɹiz

AutoIPA

ðə tɹiz wɹ̩ məʧʊɹ ænd ðej dɪdnti ɹɪli ɹuən awɹ̩ dɪvɛləpməntæ kʌt dawn ə lɑt ʌv ðə lɑɹʤ tɹiz

Allosaurus

wɛɹ ðə tɹiz wɹ̩ məʧʊɹ ænd ðej dɪdnt ɹɪl iɹuən awɹ̩ dɪvɛləpmənt ænd kʌt dawn ə lɑt ʌv ðə lɑɹʤ tɹiz

Whisper + Epitran

or ðɛ triz wu mɨʧɕj on ɛi dɨn wɛjvɨ wɛn aɻ dɨvɛwmɛn kɛ daːn ɔlaːra dɔ ladʂ tris

Taguchi

wɜː ðə tiːz wɜː mətʃuːɹ æn ðeɪ dɪdən ɹɪli ɹuːɪn ɑːɹ dɨvɛlpməntən kʌt daʊn ɐ lɑːrʌv ðə lɑːɹdʒ tiːz

ESpeak+Wav2Vec 2.0

s3801b_Utt80 (Older male)

yknow they benefited a lot from it now
jɪ noʊ ðeɪ bɛnʌfɪɾʌd ʌ lɑʔ fɹʌm ʌt naʊ

jɪ ɾ̃oʊ ðeɪ bɛɾ̃ʌfɪɾʌɾ ʌ lɑʔ fɹʌm ʌʔ naʊ AutoIPA

ðej bɛnəfɪtɪd ə lɑt fɹʌm naw Allosaurus

ðej bɛnəfɪtɪd ə lɑt fɹʌm ɪt Whisper + Epitran

i nɛv ðɛi benɛfedɨd ɛ aː fam nɛw Taguchi

iː nə̆ ðeɪ bɛnɪfɪɾɨd ɐ lɑːt fɹʌm naʊ ESpeak+Wav2Vec 2.0

s3902a_Utt109 (Younger female)

still just sitting down and my computer sits next to it VOCNOISE um
stɪl dʒʌs sɪrɪŋ daʊn n̩ maɪ kʌmpjudɻ̩ sɪts nɛkst tu ɪʔ ʌm

stɪl dʒʊs sɪriŋ daʊr̃ ɑm maɪ kn̩pjɪrɻ̩ sɪts nɛks tu ʌʔ ʌm	AutoIPA
stɪl d͡ʒʌst sɪtɪŋ dawn ɑn maj kəmpjutɻ̩ sɛts nɛks tə ɪt	Allosaurus
stɪl d͡ʒʌst sɪtɪŋ dawn ɑn maj kəmpjutɻ̩ sɪts nɛkst tə ɪt	Whisper + Epitran
stɔwː dwɨ sːɛrin daːn a maj kɨmpçirj sɛts nɛks tɛ wɨ am	Taguchi
stɪl dʒʌs sɪrɪŋ daʊn ʌ maɪ kəmpiːrəɾ sɪttss nɛkstuːɪt ʌm	ESpeak+Wav2Vec 2.0

Top edit distance errors on the Buckeye test split

These are the **most common mistakes** we saw from a model fine-tuned on all data in the Buckeye train split (ginic/full_dataset_train_3_wav2vec2-large-xlsr-53-buckeye-ipa).

Substitution	Percentage of all substitutions
ɪ → ʌ	5.27 %
ɪ → ɛ	4.70
ʌ → ɪ	4.02
ɾ̃ → n	2.57
ɛ → ʌ	2.32
o → ʌ	2.29
i → ɪ	2.28

Deletion	Percentage of all deletions
ɪ	14.74 %
ʊ	9.34
ʌ	9.09
ɹ	5.88
n	5.70
t	4.95
l	3.64

Insertion	Percentage of all insertions
ɪ	9.76 %
t	8.57
ʌ	7.68
n	7.57
d	4.37
ɛ	4.13
ʊ	3.95

Vowel Error Rates

For vowels *v* in the vocabulary *V*, we want to see **which are most challenging** across the entire corpus.

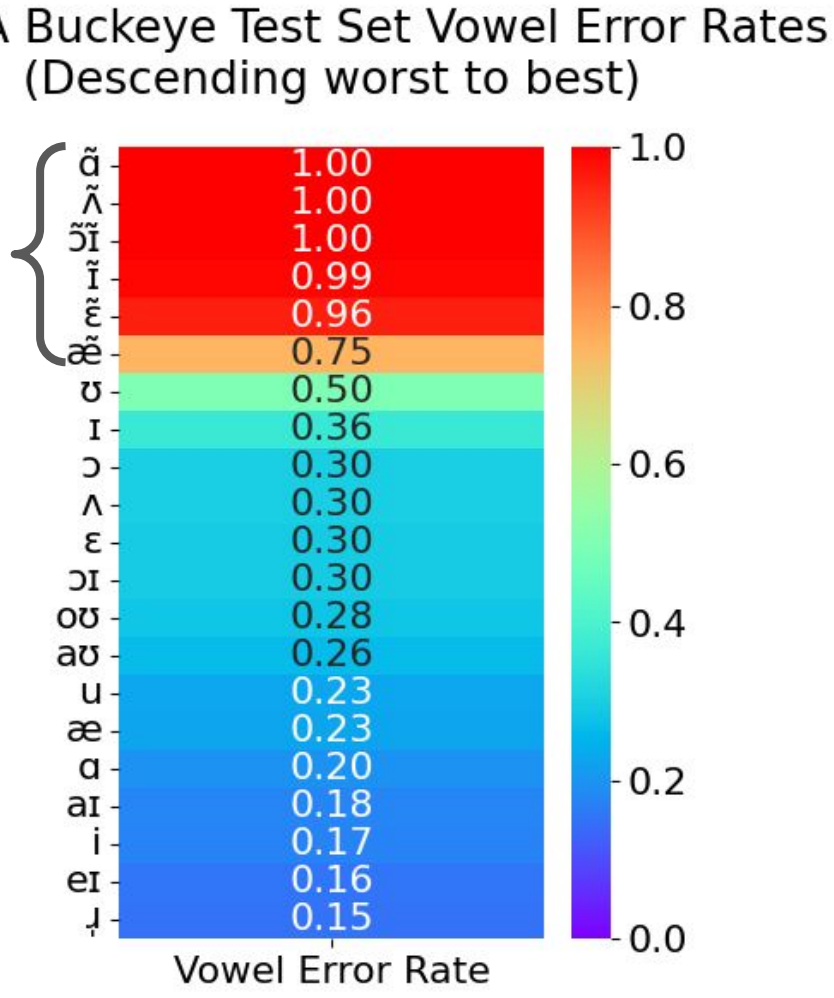
Which vowels is the model getting **wrong most frequently?**

$$count\_errors(v) = \sum_{w \in V} count\_substitutions(v, w) + count\_deletion(v)$$

$$error\_rate(v) = \frac{count\_errors(v)}{total\_count(v)}$$

We don't do well on nasalized vowels.

They are relatively rare overall, which makes them more challenging for the model to recognize. They are also apparently not well transcribed by Buckeye's humans.



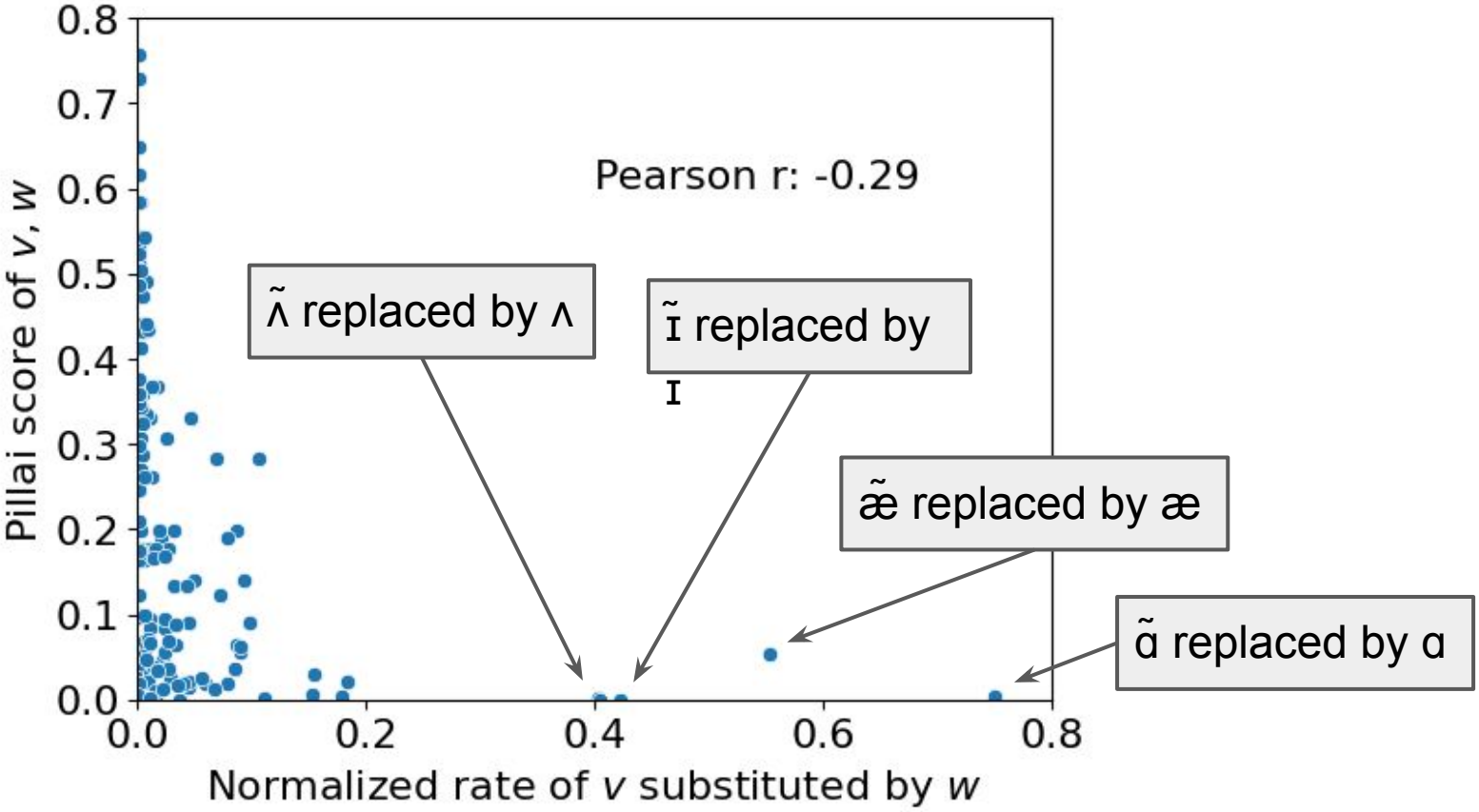
Pillai scores

Pillai scores are a measure of distributional overlap used in sociolinguistics to quantify the degree of vowel merger (Hay et al. 2006, Stanley and Sneller 2023).

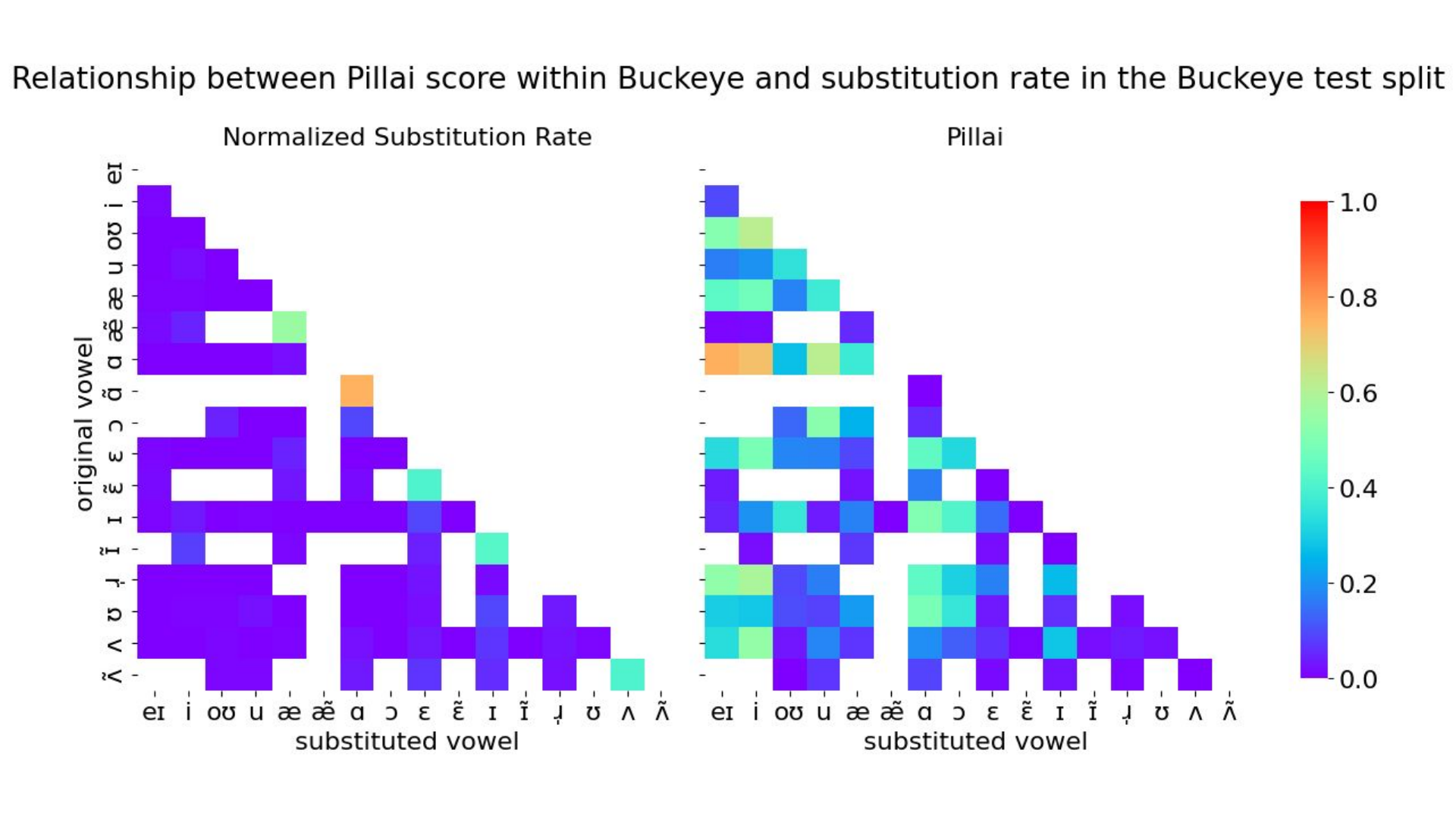
We calculated Pillai scores based on F1 and F2 measurements of the Buckeye (and TIMIT) vowels to see if they correlated with error rates

A high Pillai score indicates a low degree of overlap, or a high degree of contrast

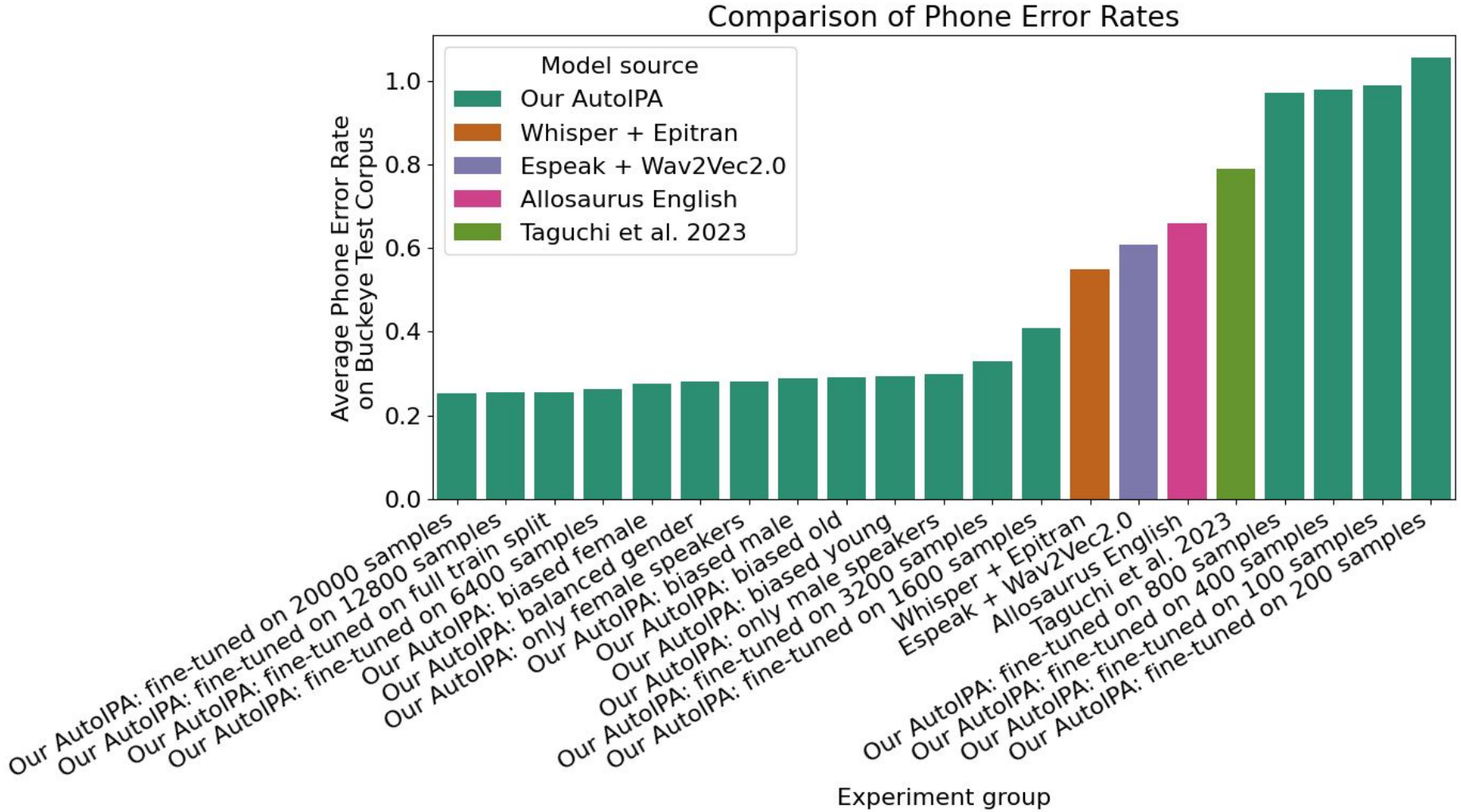
Relationship between Pillai score within Buckeye and substitution rate in the Buckeye test split

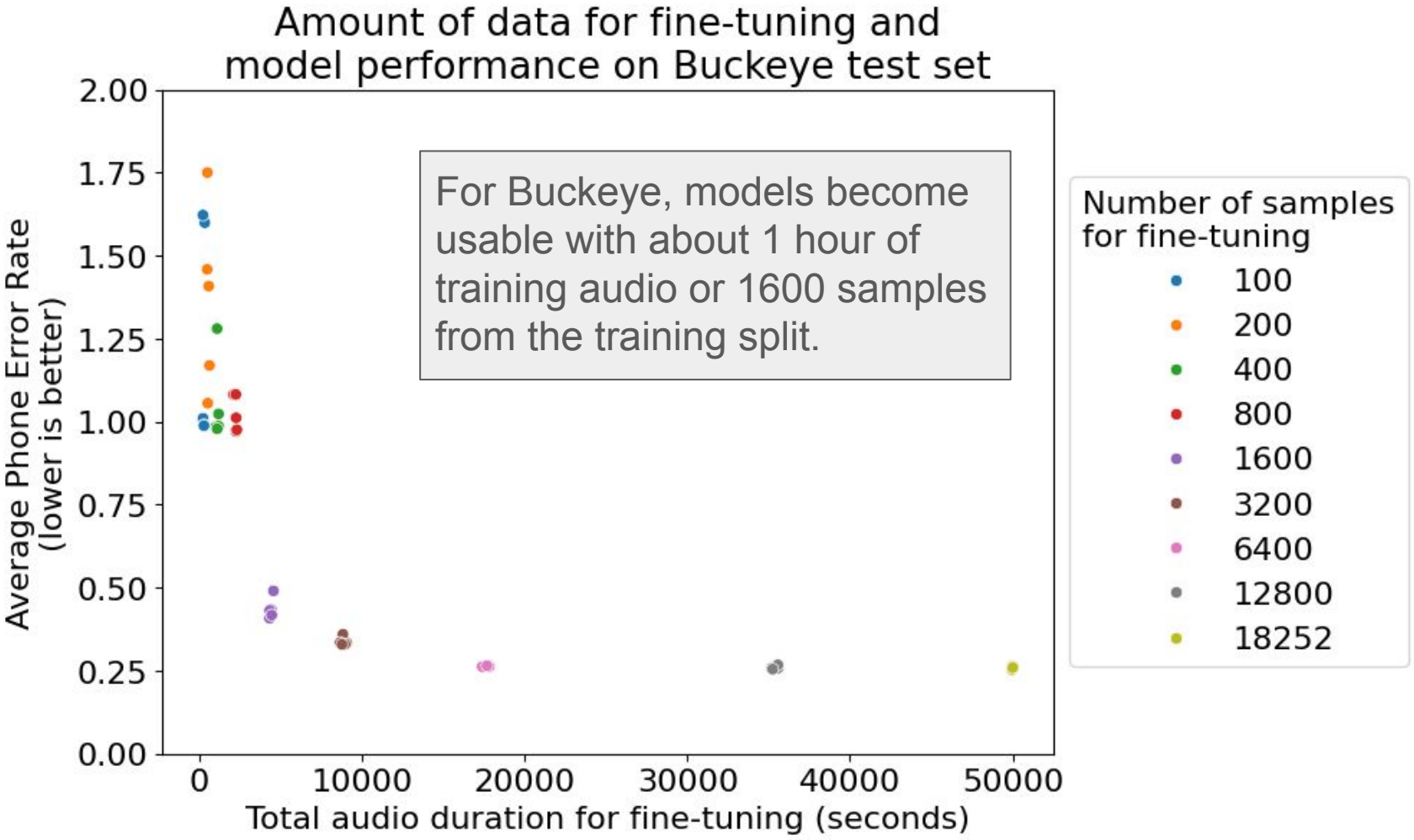


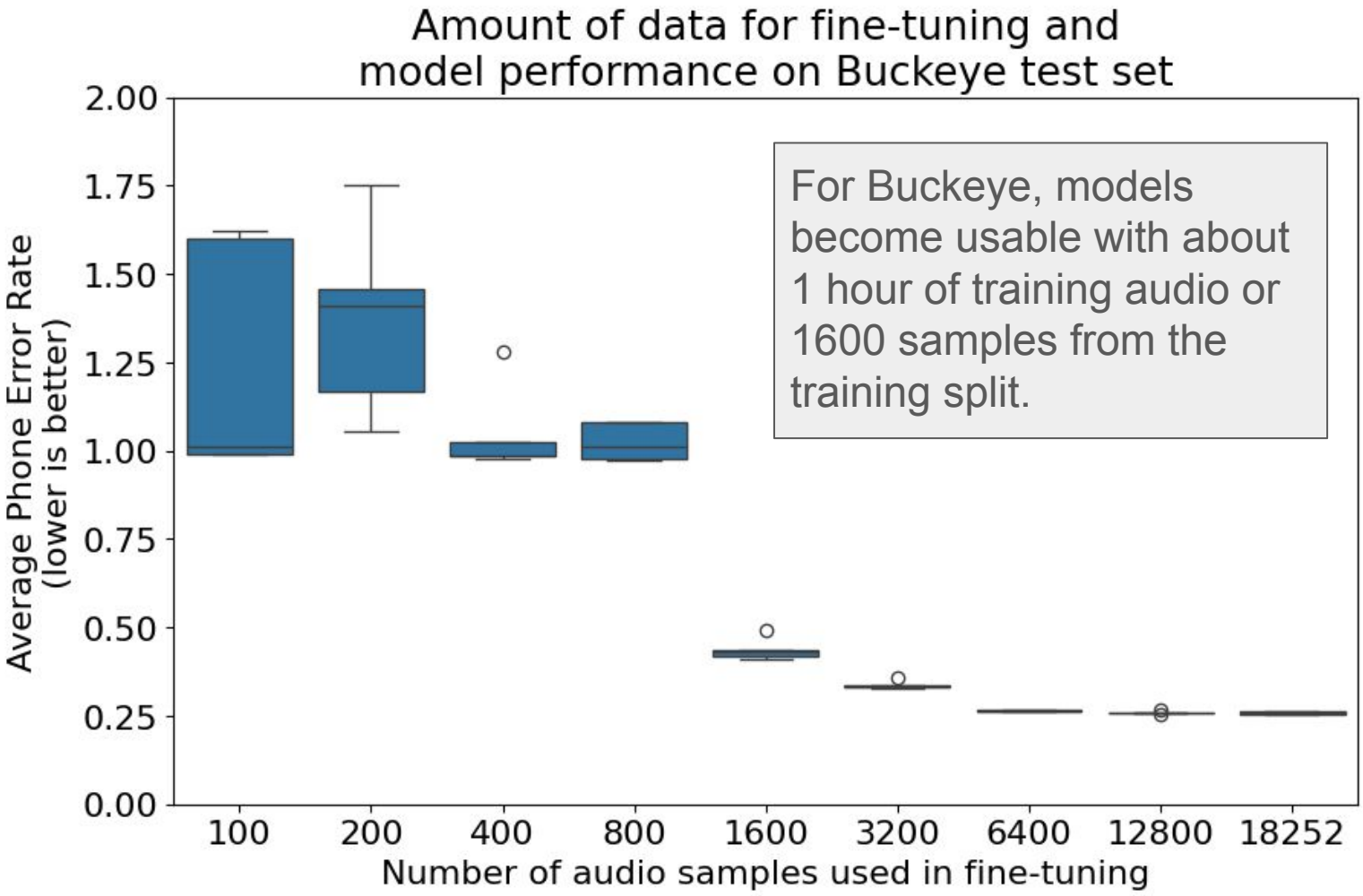
**Substitution rates, because v != w*


Relationship between Pillai score within Buckeye and substitution rate in the Buckeye test split
Normalized Substitution Rate
Pillai
original vowel
substituted vowel
substituted vowel
eɪ i oʊ u æ æ̃ ɑ ɔ ɛ ɛ̃ ɪ ɪ̃ ɻ̩ ʊ ʌ ʌ̃
1.0
0.8
0.6
0.4
0.2
0.0

Impact of the quantity of fine-tuning data


Comparison of Phone Error Rates
Model source
Our AutoIPA
Whisper + Epitran
Espeak + Wav2Vec2.0
Allosaurus English
Taguchi et al. 2023
Average Phone Error Rate on Buckeye Test Corpus
0.0
0.2
0.4
0.6
0.8
1.0
Our AutoIPA: fine-tuned on 20000 samples
Our AutoIPA: fine-tuned on 12800 samples
Our AutoIPA: fine-tuned on full train split
Our AutoIPA: fine-tuned on 6400 samples
Our AutoIPA: biased female
Our AutoIPA: balanced gender
Our AutoIPA: only female speakers
Our AutoIPA: biased male
Our AutoIPA: biased old
Our AutoIPA: biased young
Our AutoIPA: only male speakers
Our AutoIPA: fine-tuned on 3200 samples
Our AutoIPA: fine-tuned on 1600 samples
Whisper + Epitran
Espeak + Wav2Vec2.0
Allosaurus English
Taguchi et al. 2023
Our AutoIPA: fine-tuned on 800 samples
Our AutoIPA: fine-tuned on 400 samples
Our AutoIPA: fine-tuned on 100 samples
Our AutoIPA: fine-tuned on 200 samples
Experiment group


Amount of data for fine-tuning and model performance on Buckeye test set
For Buckeye, models become usable with about 1 hour of training audio or 1600 samples from the training split.
Average Phone Error Rate (lower is better)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
10000
20000
30000
40000
50000
Total audio duration for fine-tuning (seconds)
Number of samples for fine-tuning
100
200
400
800
1600
3200
6400
12800
18252


Amount of data for fine-tuning and
model performance on Buckeye test set
Average Phone Error Rate
(lower is better)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
100
200
400
800
1600
3200
6400
12800
18252
Number of audio samples used in fine-tuning
For Buckeye, models become usable with about 1 hour of training audio or 1600 samples from the training split.

Amount of fine-tuning data: Practical implications

Speaker S25 in Buckeye test split: "over by Riverside Hospital"

Original Buckeye transcription: [oʊ v ɻ b aɪ ɹ ɪ v ɻ s aɪ d h ɑ s p ɪ ɾ ˩]

Model fine-tuned on full Buckeye train split	ɪ v ɻ b aɪ i ɹ ɪ v ɻ s aɪ d h ɑ s p ɪ ɾ ˩
12800 sample fine-tuned model	ɛ v ɻ b aɪ ɹ ɪ v ɻ s aɪ d h ɑ s p ɪ ɾ ˩
6400 sample fine-tuned model	ɛ v w ɻ b aɪ ɹ ɪ v ɻ s aɪ d h ɑ s p ɪ ɾ ˩
3200 sample fine-tuned model	ɛ v ɻ b eɪ ɹ ɪ v ɻ s s aɪ eɪ d h ɑ s p ɪ ɾ oʊ
1600 sample fine-tuned model	ɪ f [redacted] b ɛ w ɪ v [redacted] s aɪ d h aʊ s p ɪ ɾ ˩
800 sample fine-tuned model	[redacted]
400 sample fine-tuned model	t [redacted]

Red = Deletion, Blue = Insertion, Yellow = Substitution

Amount of fine-tuning data: Practical implications

Speaker S38 in Buckeye test split: "oh they are they're killing on it they really are"

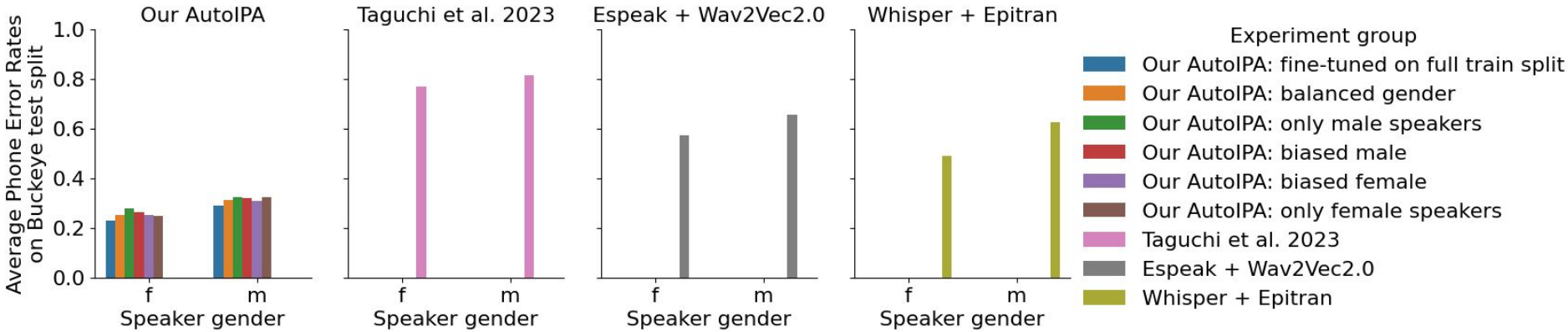
Original Buckeye transcription: [oʊ ð ɛ ɑ ɹ ð ɻ k ɪ l ɪ ŋ ɔ n ʌ t ð ɛ ɹ i l i ɔ ɹ]

Model fine-tuned on full Buckeye train split	oʊ ð eɪ ɑ ɹ ð ɛ ɹ k ɪ l n ∎ ð eɪ ɹ ɪ l i ɑ ɹ s
12800 sample fine-tuned model	oʊ ð eɪ ɑ ɹ ð ɛ ɹ k ɪ l ɳ ∎ ʌ ∎ ð eɪ ɹ ɪ l i ɑ ɹ s
6400 sample fine-tuned model	oʊ ð eɪ ɑ ɹ ð ɛ ɹ k ɪ l ʌ ɳ ɑ ∎ ʌ ∎ ð eɪ ɹ ɪ l i ɑ ɹ s
3200 sample fine-tuned model	oʊ ð eɪ ɑ ɹ ð ɛ ɹ k ɛ l ʔ ∎ ∎ n ɑ ∎ ð eɪ ɹ ɪ l i ɑ v d
1600 sample fine-tuned model	ɑ ð æ ɑ ∎ ð ɛ k ɛ l ∎ ∎ ∎ n ɑ ʌ ∎ ∎ ɛ ɹ ɪ l a ʊ ∎
800 sample fine-tuned model	∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ …
400 sample fine-tuned model	∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ ∎ t ∎ ∎ ∎ …

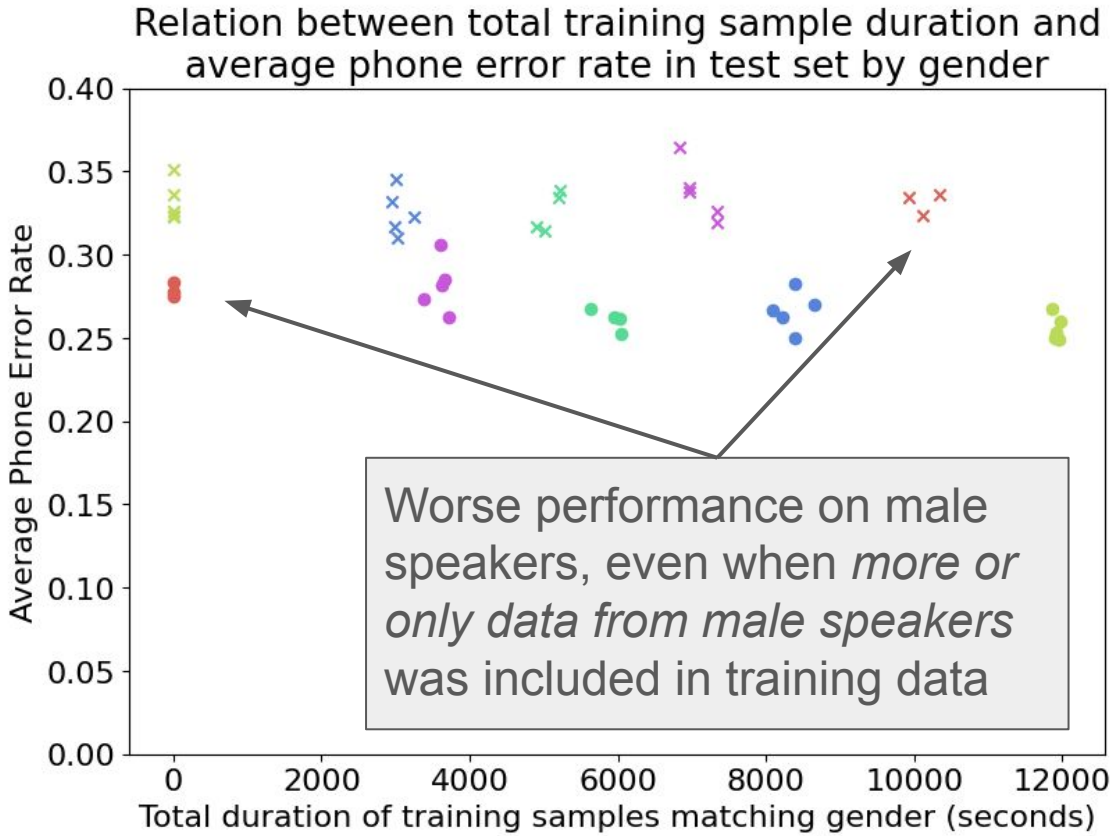
Red = Deletion, Blue = Insertion, Yellow = Substitution

Gender Effects

All models had better performance on female speakers in the Buckeye test data…



…regardless of how fine-tuning data was selected



Test sample speaker

- male speaker
- female speaker

Fine-tuning data setting

- Balanced gender
- Only male speakers
- Biased with 70% training samples male
- Biased with 70% training samples female
- Only female speakers

Yao *et al.* 2010: 108 on Buckeye data

Generally speaking, female speakers produce slightly longer vowels than male speakers (p=0.009). As expected, they also have higher formant frequencies compared to male speakers (p<0.001 for F1, p=0.002 for F2). More importantly, as can be seen from Figure 3, on average female speakers have a much larger vowel space than male speakers. This is also consistent with previous findings (Byrd, 1994). Both longer duration and more expanded vowel space are indicators of clear speech (Bradlow et al., 1996), which suggests that female speakers produce clearer speech than male speakers.

Adda-Decker & Lamel 2005: on speech recognizers performing better on female speakers

Results consistently show a lower word error rate on female speech ranging from 0.7 to 7% depending on the condition. An analysis of automatically produced pronunciations in speech training corpora (totaling 4000 hours of speech) revealed that female speakers tend to stick more consistently to standard pronunciations than male speakers. Concerning speech disfluencies, male speakers show larger proportions of filled pauses and repetitions, as compared to females.

TIMIT as a test set

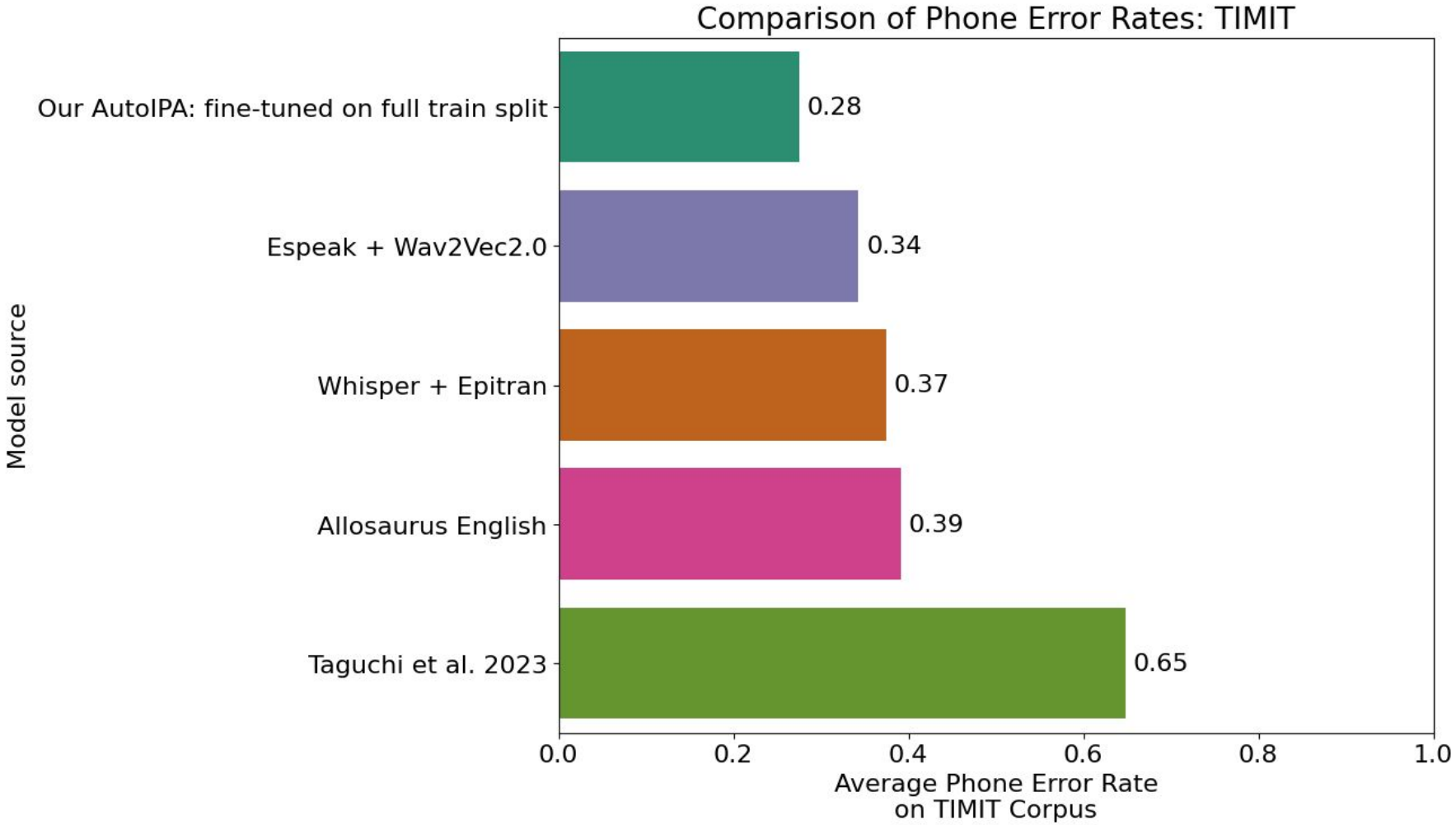
To begin to get a sense of how our model performs on speech from other varieties of English, we tested it on TIMIT

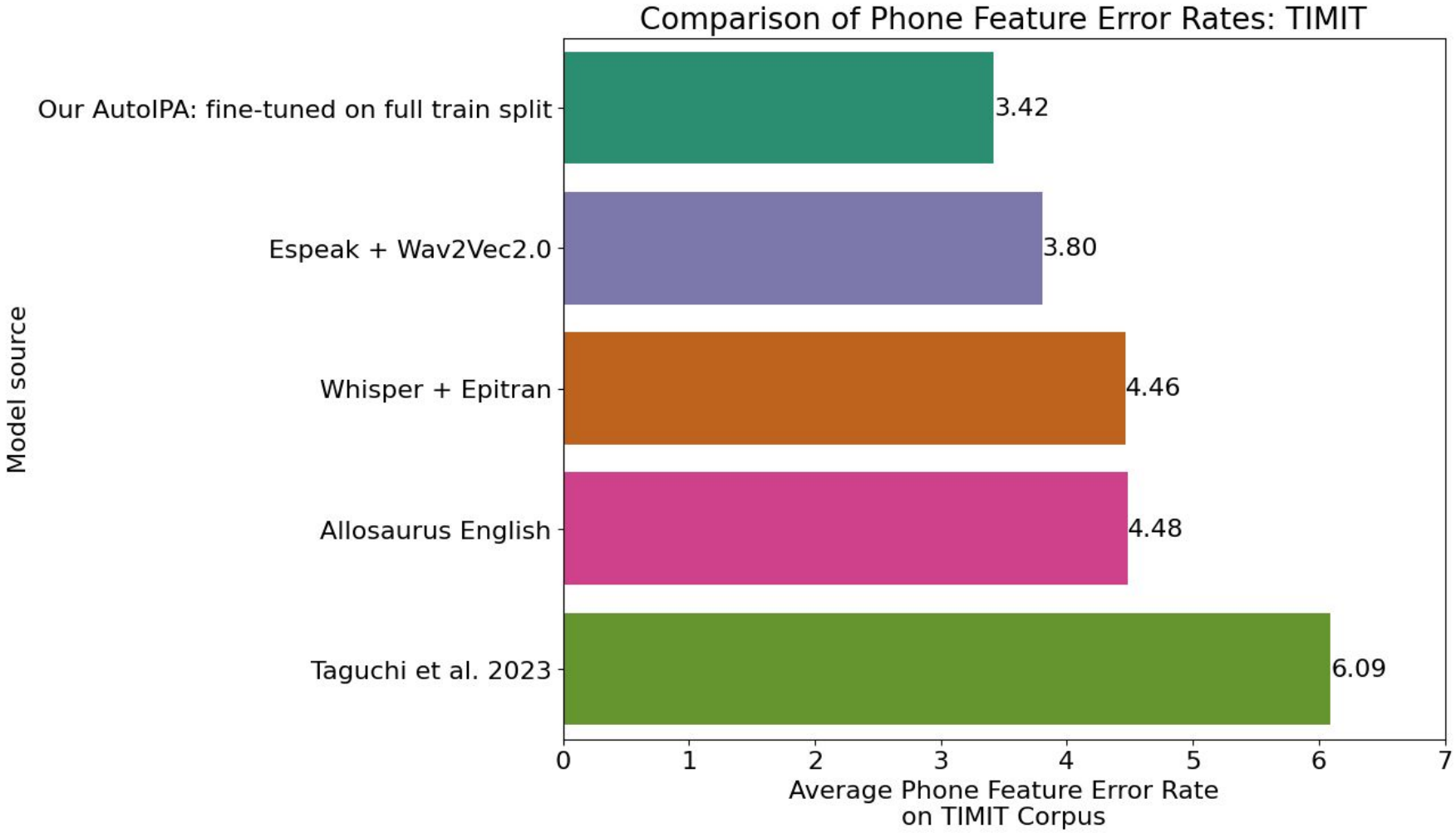
TIMIT consists of transcribed speech from 630 speakers from 8 dialect regions, each reading ten “phonetically rich” sentences. There are 5 hours of speech in total.

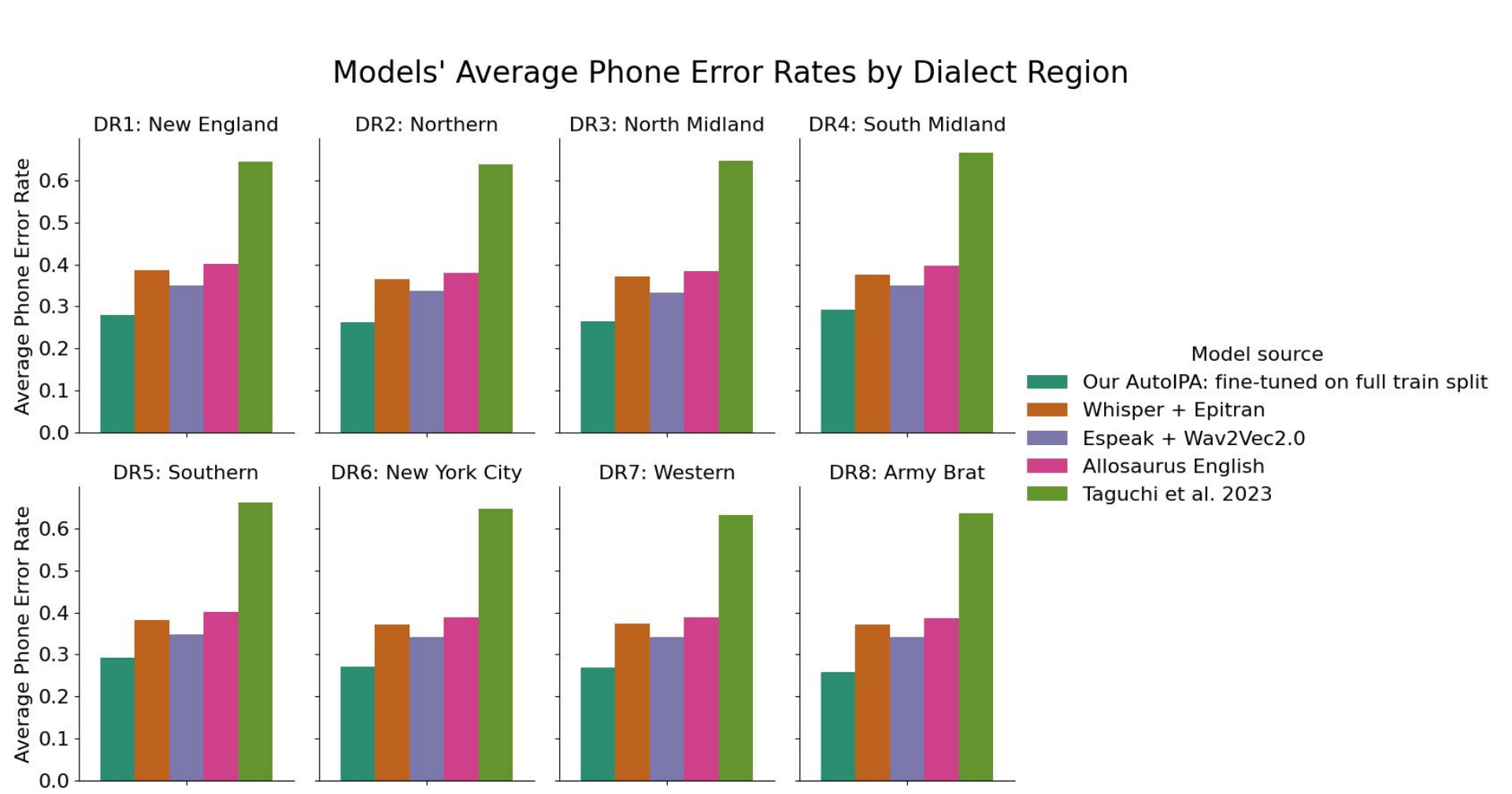
The ARPABET to IPA translation was trickier for TIMIT, because stop closure is indicated separately from release. We decided to merge all closure-release sequences for the same phone into a single segment.

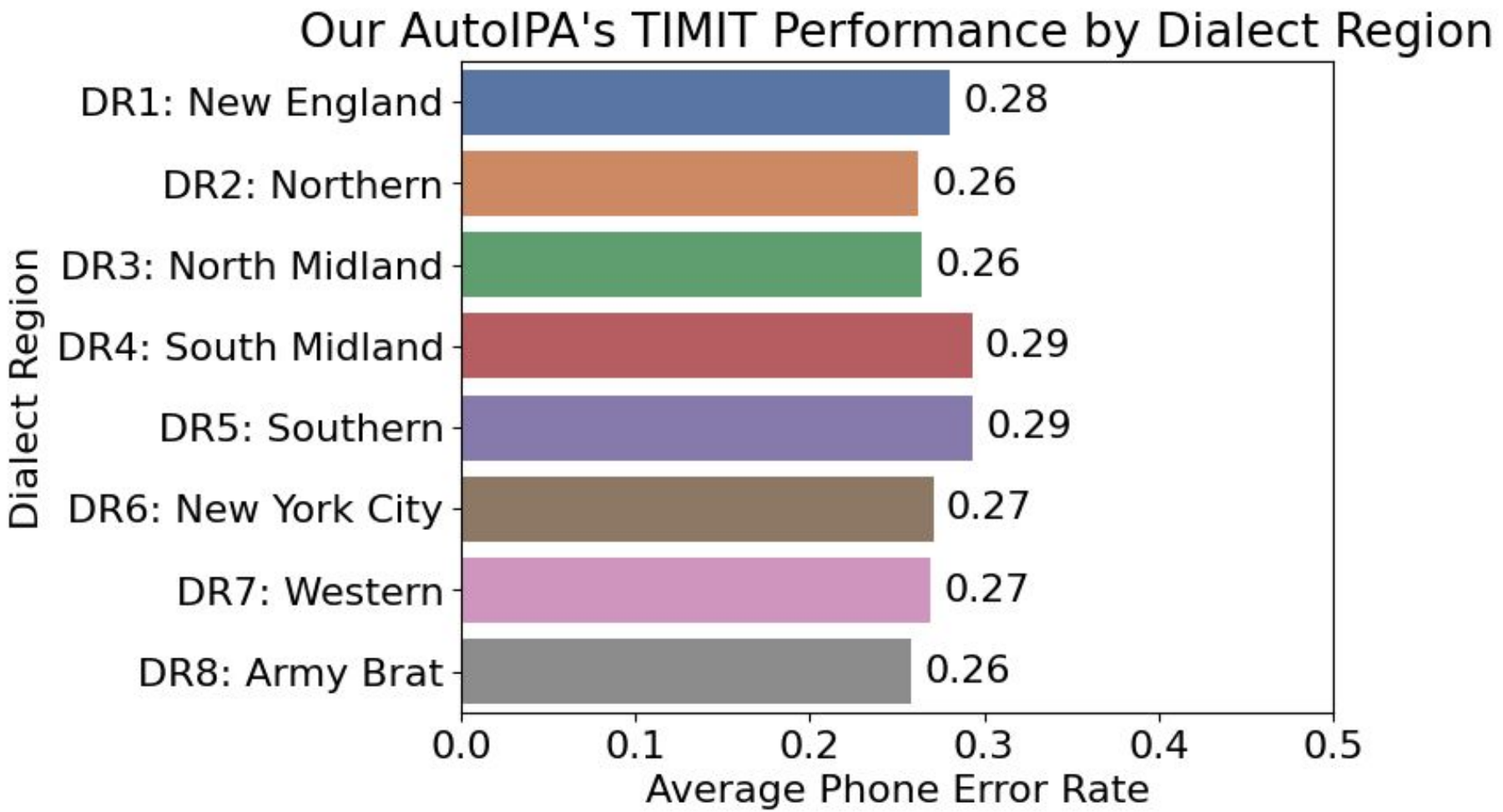
The mappings can be examined here (note that Phonecodes is greedy, so the longer input will be chosen first)

The standard approach to TIMIT in speech recognition (Lee and Hon 1989) is to merge all closures with silence. This means that all coda unreleased stops are lost (!)


Comparison of Phone Error Rates: TIMIT
Our AutoIPA: fine-tuned on full train split
Espeak + Wav2Vec2.0
Whisper + Epitran
Allosaurus English
Taguchi et al. 2023
0.28
0.34
0.37
0.39
0.65
Model source
0.0
0.2
0.4
0.6
0.8
1.0
Average Phone Error Rate
on TIMIT Corpus


Comparison of Phone Feature Error Rates: TIMIT
Our AutoIPA: fine-tuned on full train split
Espeak + Wav2Vec2.0
Whisper + Epitran
Allosaurus English
Taguchi et al. 2023
3.42
3.80
4.46
4.48
6.09
Model source
0
1
2
3
4
5
6
7
Average Phone Feature Error Rate
on TIMIT Corpus


Models' Average Phone Error Rates by Dialect Region
DR1: New England
DR2: Northern
DR3: North Midland
DR4: South Midland
DR5: Southern
DR6: New York City
DR7: Western
DR8: Army Brat
Average Phone Error Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Model source
Our AutoIPA: fine-tuned on full train split
Whisper + Epitran
Espeak + Wav2Vec2.0
Allosaurus English
Taguchi et al. 2023


Our AutoIPA's TIMIT Performance by Dialect Region
DR1: New England
DR2: Northern
DR3: North Midland
DR4: South Midland
DR5: Southern
DR6: New York City
DR7: Western
DR8: Army Brat
0.28
0.26
0.26
0.29
0.29
0.27
0.27
0.26
Dialect Region
0.0
0.1
0.2
0.3
0.4
0.5
Average Phone Error Rate

Same sentence, different regions, TIMIT vs. AutoIPA

She had your dark suit and greasy wash water all year

New England (“best New England accent so far”) (VMH0, F, b. 1960, rec. 1986)

ʃi ɦæd jɨ dɑk sʉtn̩ gɹisi wɑʃ wɑrə ʔɔl jɪə TIMIT

ʃi hæd jɪ dɑk suʔn̩ gɹisi wɑʃ wɑrʌ oʊ jɛ AutoIPA

New York (“has good NY pronunciation of ‘saw’”) (HXS0, F, b. 1941, rec. 1986)

ʃi ɦæd jʊ dɑk sʉtʔɨn gɹisi wɑʃ wɔrʌ ʔɔl jɪə˞ TIMIT

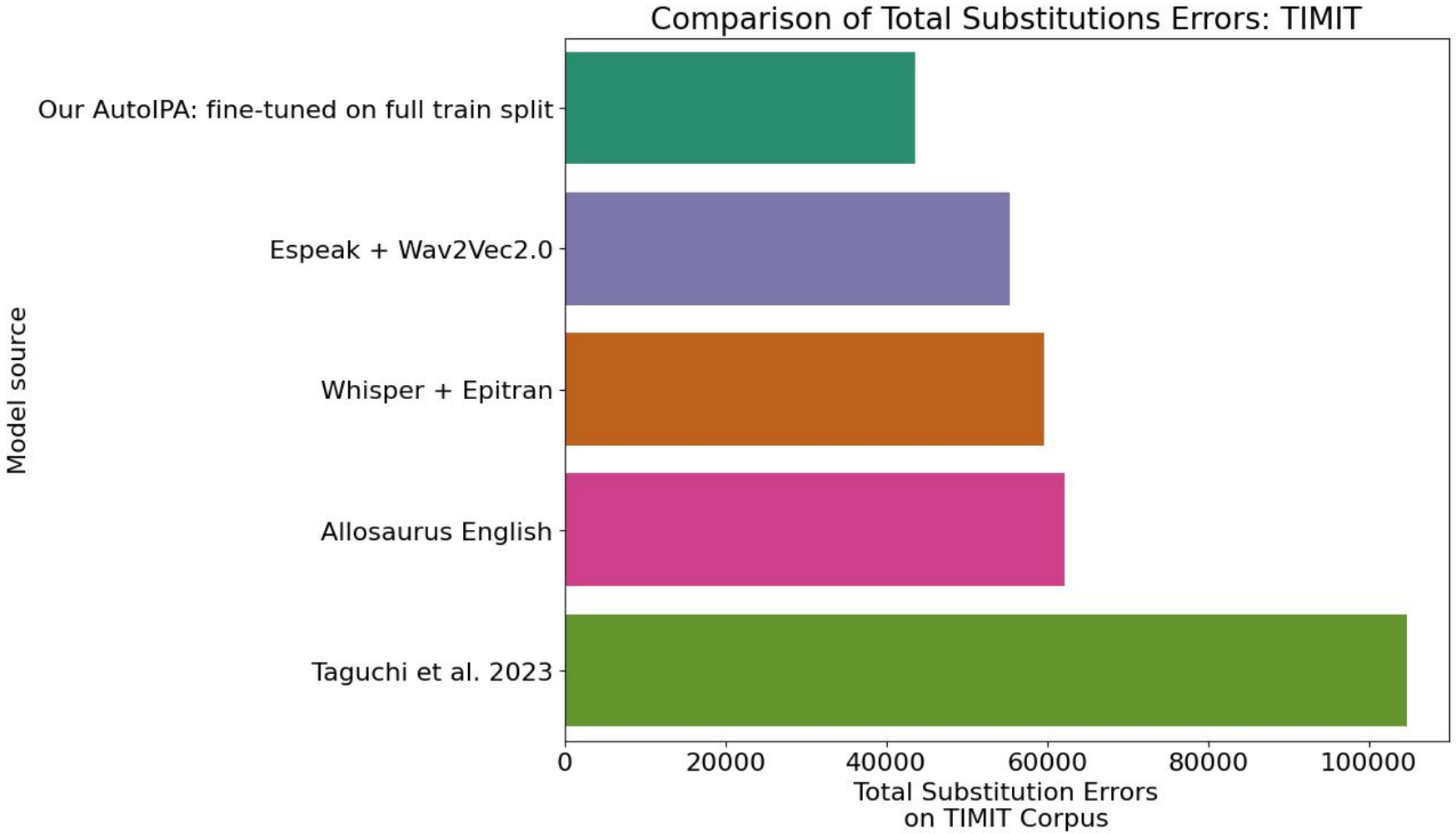
ʃi hæd juʌ dɔk sutɪn gɹisi wɑʃ wɔrɻ oʊl jiɹ AutoIPA

North Midland (DSS1, M, b. 1955, rec. 1986)

ʃi ɦæd jə˞ dɑɹk sʉrɪn gɹisi wɔɹʃ wɔrə˞ ɔl jɪə˞ TIMIT

ʃi hæd jɻ dɑɹk surɛn gɹeɪsi wɔɹʃ fʌɑrɻ ɔl jɪɹ AutoIPA

TIMIT Error Analysis


Comparison of Total Substitutions Errors: TIMIT
Our AutoIPA: fine-tuned on full train split
Espeak + Wav2Vec2.0
Whisper + Epitran
Allosaurus English
Taguchi et al. 2023
Model source
0
20000
40000
60000
80000
100000
Total Substitution Errors
on TIMIT Corpus

Top substitution errors on the TIMIT Corpus

Our AutoIPA

Substitution	Count of error
ɨ → ɪ	4704
ə → ʌ	3741
ɚ → ɹ̩	3202

Espeak + Wav2Vec2.0

Substitution	Count of error
i → iː	4075
ɨ → ɪ	3502
ɑ → ɑː	3435

Whisper + Epitran

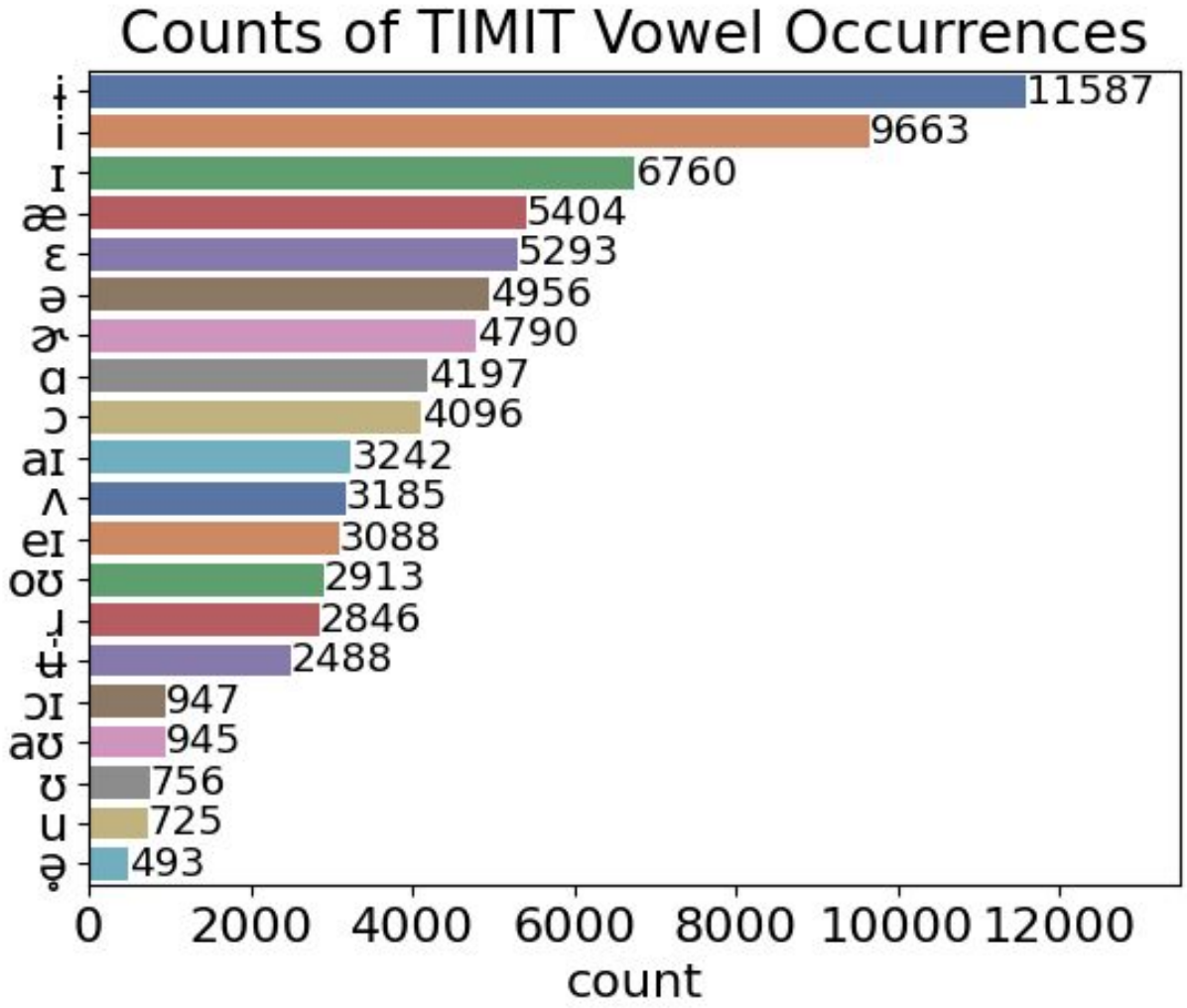
Substitution	Count of error
ɪ → j	6444
ɨ → ə	4545
ʊ → w	3049

Allosaurus English

Substitution	Count of error
ɪ → j	6189
ɨ → ə	4515
ɨ → ɪ	3122

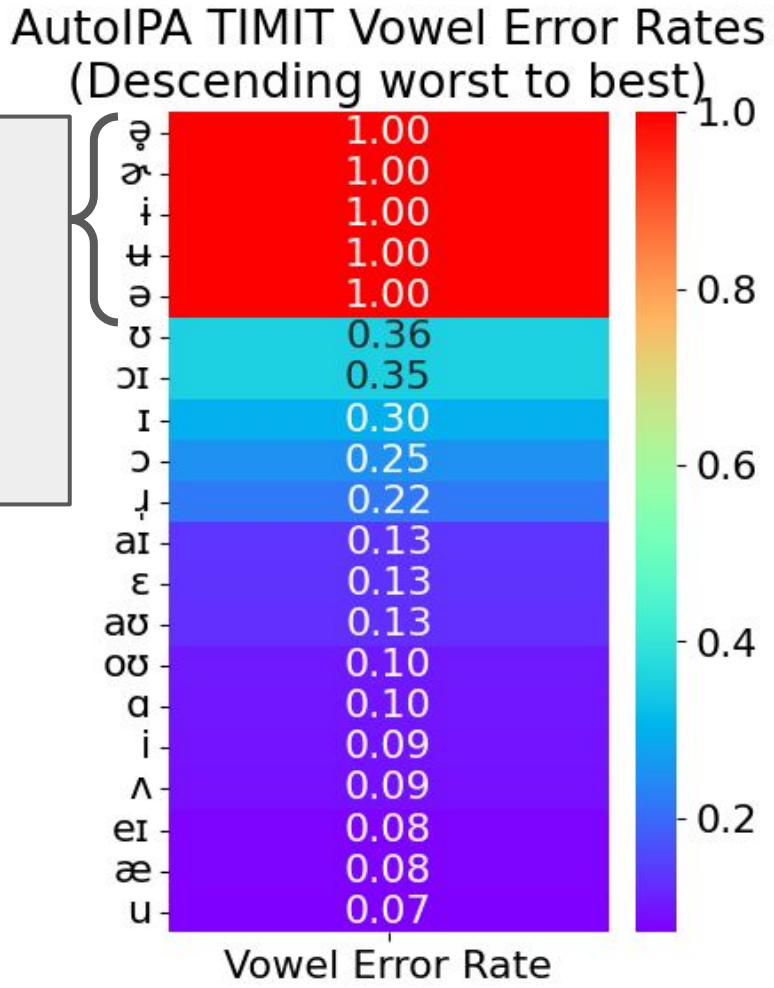
Taguchi et al. 2023

Substitution	Count of error
ɪ → i	2777
ɹ → r	1948
ɪ → j	1763


Counts of TIMIT Vowel Occurrences
ɨ 11587
i 9663
ɪ 6760
æ 5404
ɛ 5293
ə 4956
ɚ 4790
ɑ 4197
ɔ 4096
aɪ 3242
ʌ 3185
eɪ 3088
oʊ 2913
ɾ 2846
ɸ 2488
ɔɪ 947
aʊ 945
ʊ 756
u 725
ɐ 493
0 2000 4000 6000 8000 10000 12000
count

These errors reflect differences in transcription conventions between corpora.

Because these are not in the Buckeye vocabulary, our model cannot output the symbols.



Distributions of Errors for a Given Vowel

For a given vowel *v* in the TIMIT vocabulary *V*, we want to understand **how the model is getting the vowel wrong.**

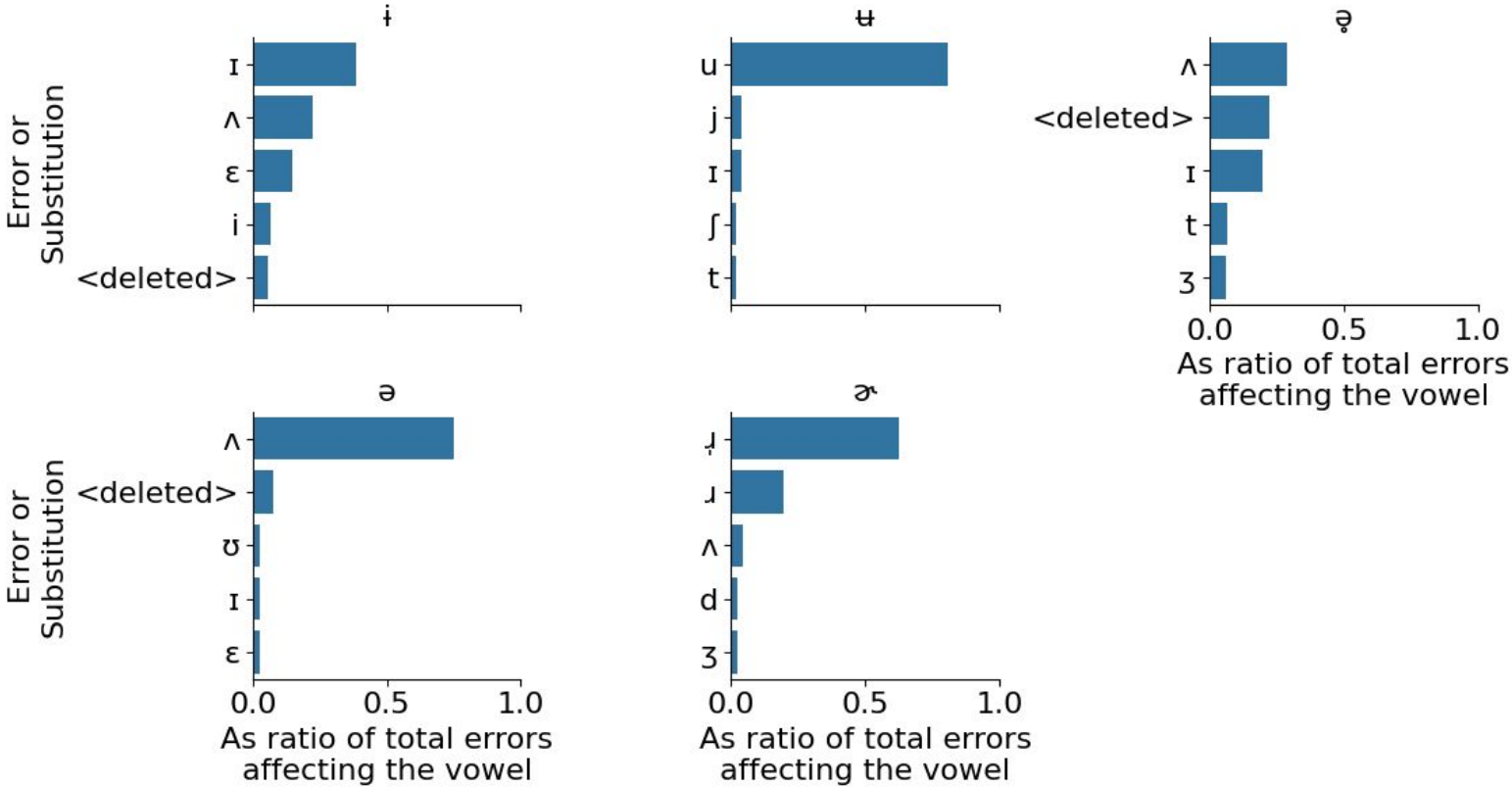
What **kinds of mistakes** and **in what distributions** does the model make for each vowel? Can this tell us anything about the corpus or its language varieties?

For a fixed vowel *v* and specific type error on *v* (deletion or substitution by *w*), normalize by the total number of errors:

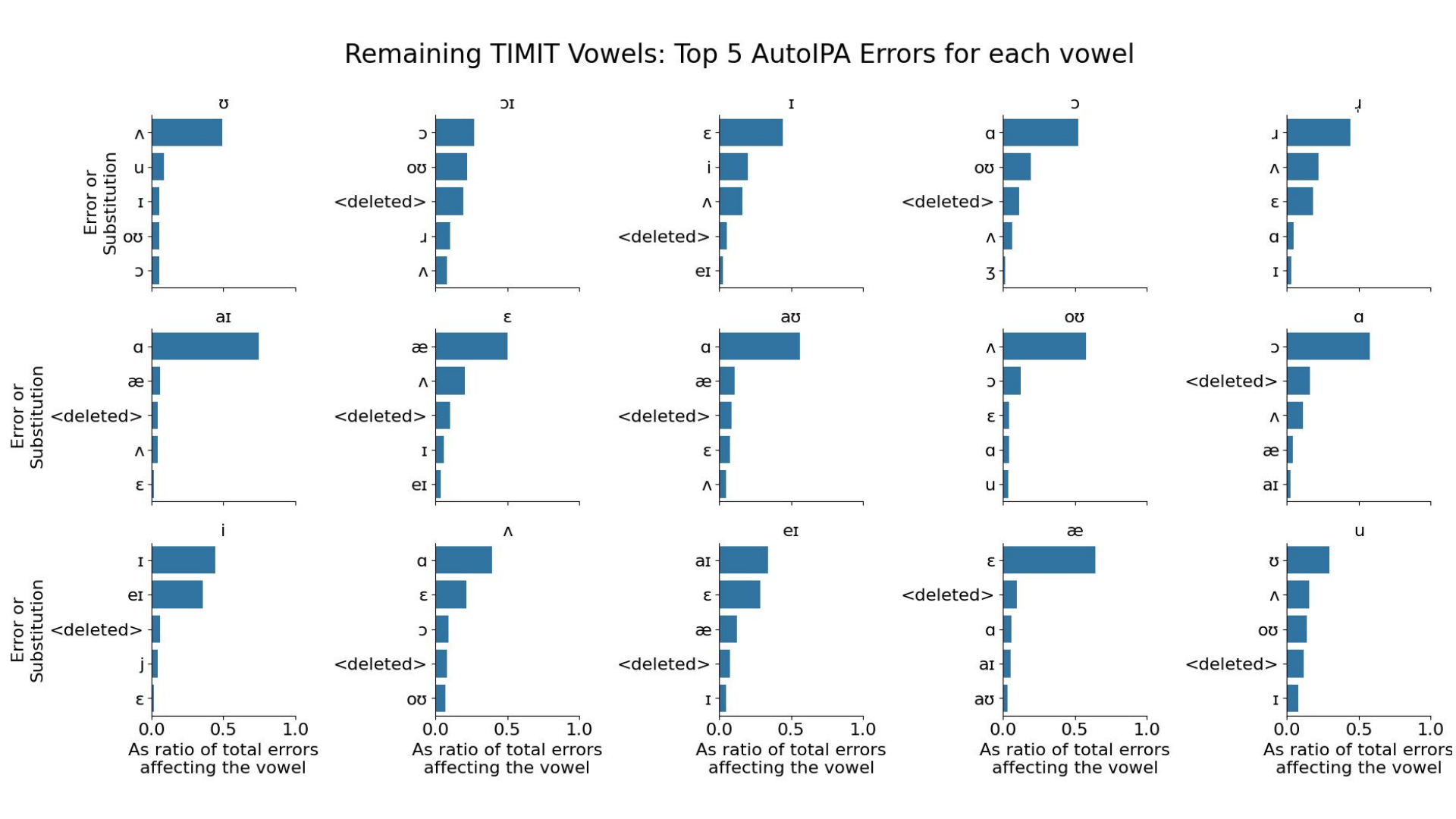
$$\frac{count\_deletion(v)}{count\_errors(v)} \qquad \frac{count\_substitutions(v,w)}{count\_errors(v)}$$

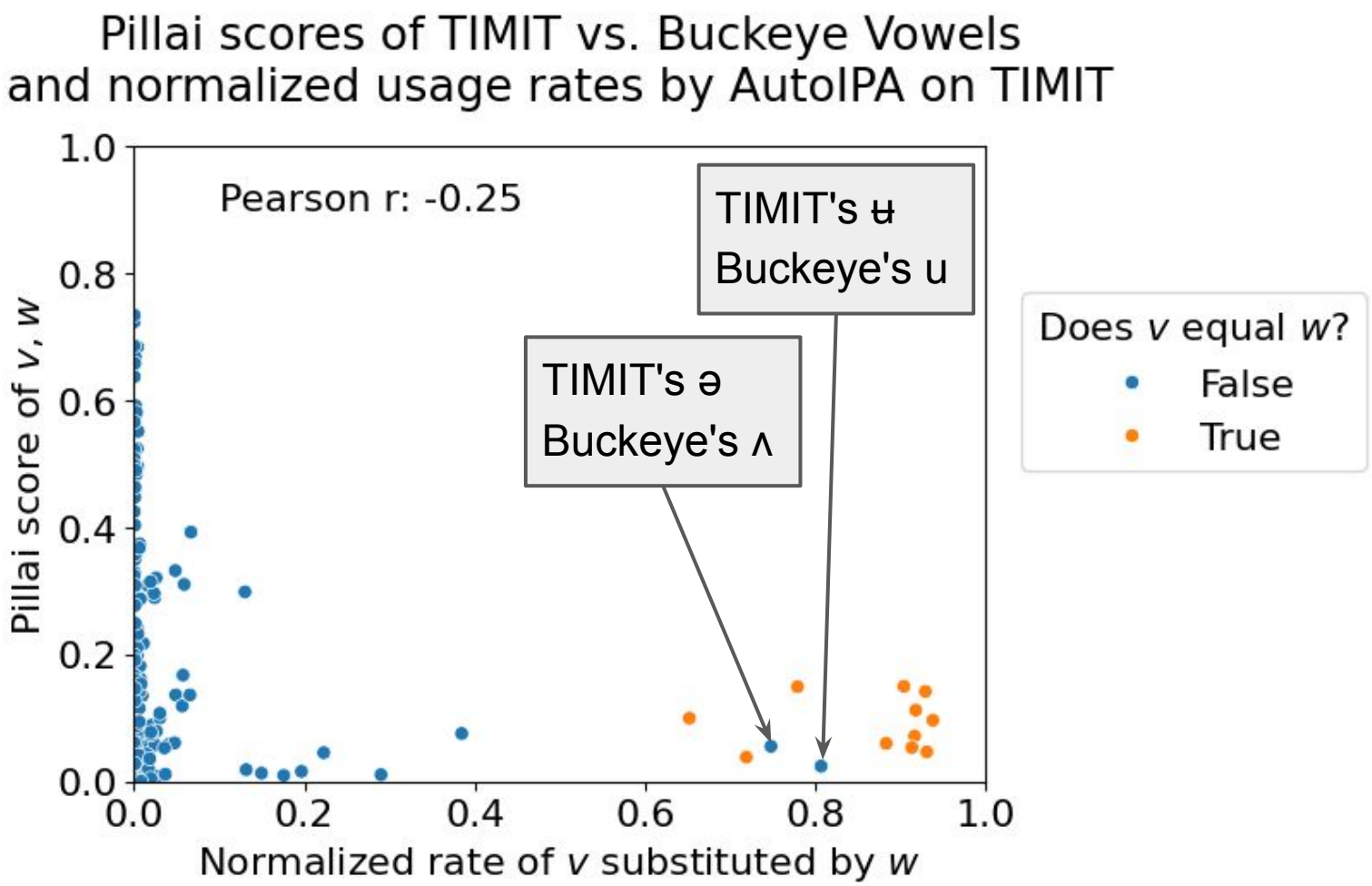
** Conditional probability of each error given that there is some error affecting v*

Top 5 errors for vowels AutoIPA always incorrectly transcribes

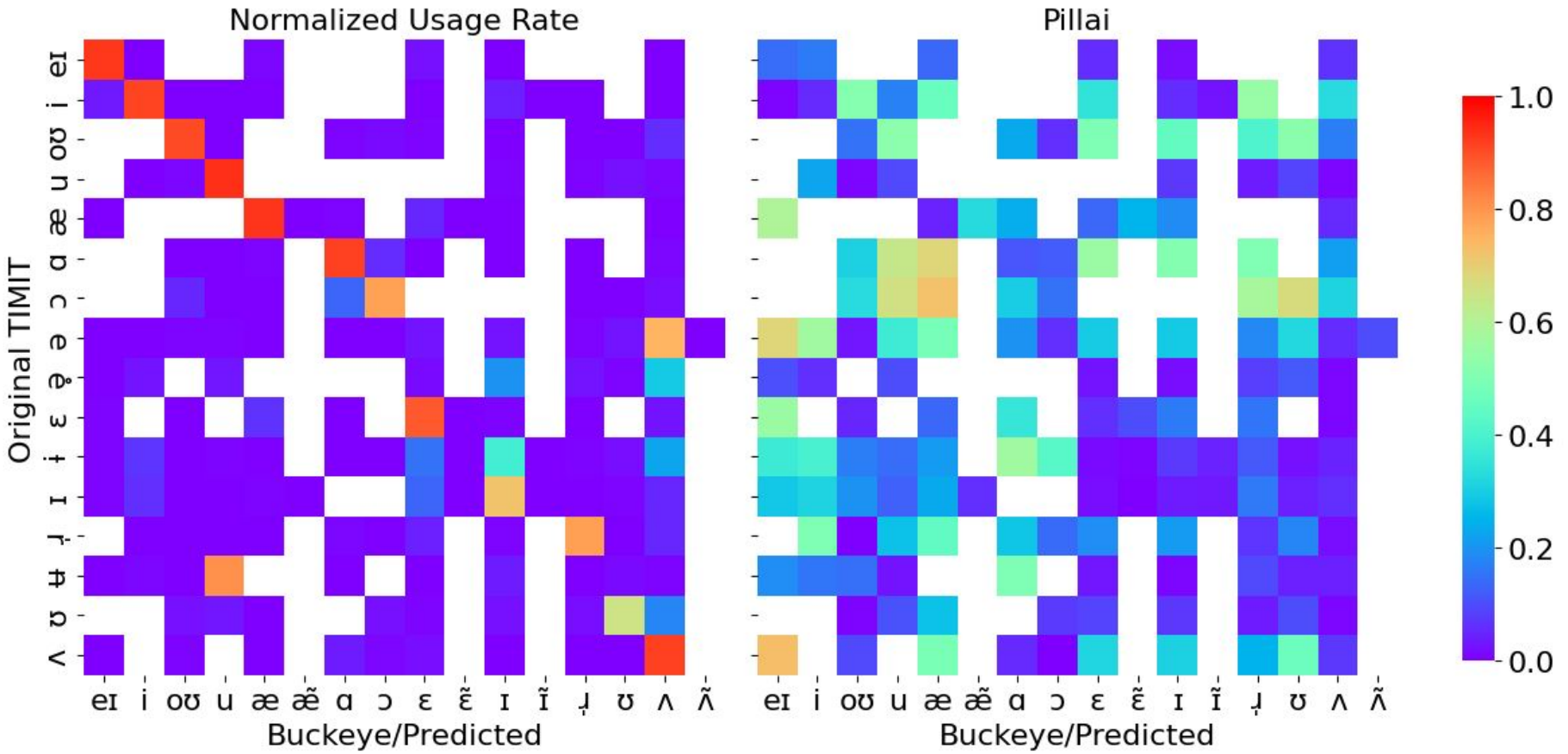


Correctable with simple replacement in pre- or post- processing


Remaining TIMIT Vowels: Top 5 AutoIPA Errors for each vowel
ʊ
ɔɪ
ɪ
ɔ
ɻ
aɪ
ɛ
aʊ
oʊ
ɑ
i
ʌ
eɪ
æ
u
Error or Substitution
ʌ
u
ɪ
oʊ
ɔ
ɔ
oʊ
<deleted>
ɻ
ʌ
ɛ
i
ʌ
<deleted>
eɪ
ɔ
ɑ
oʊ
<deleted>
ʌ
ɛ
ɻ
ɑ
ɛ
ɪ
ɹ
ʌ
ɑ
ɪ
ɑ
æ
<deleted>
ʌ
ɛ
æ
<deleted>
ɪ
eɪ
ɑ
æ
<deleted>
ɛ
ʌ
ʌ
ɔ
ɑ
ɛ
u
ɔ
<deleted>
ʌ
æ
aɪ
ɪ
eɪ
<deleted>
j
ɛ
ɑ
ɛ
ɔ
<deleted>
oʊ
aɪ
ɛ
æ
<deleted>
ɪ
ɛ
<deleted>
ɑ
aɪ
aʊ
ʊ
ʌ
oʊ
<deleted>
ɪ
0.0
0.5
1.0
As ratio of total errors affecting the vowel



*Usage rates, because it's possible that v=w


Pillai scores of TIMIT vs. Buckeye Vowels and normalized usage rates by AutoIPA on TIMIT
Normalized Usage Rate
Pillai
Original TIMIT
Buckeye/Predicted
Buckeye/Predicted
eɪ i oʊ u æ æ̃ ɑ ɔ ɛ ɛ̃ ɪ ĩ ɨ̃ ʊ ʌ ʌ̃
1.0
0.8
0.6
0.4
0.2
0.0

Web-based implementation with text grid support

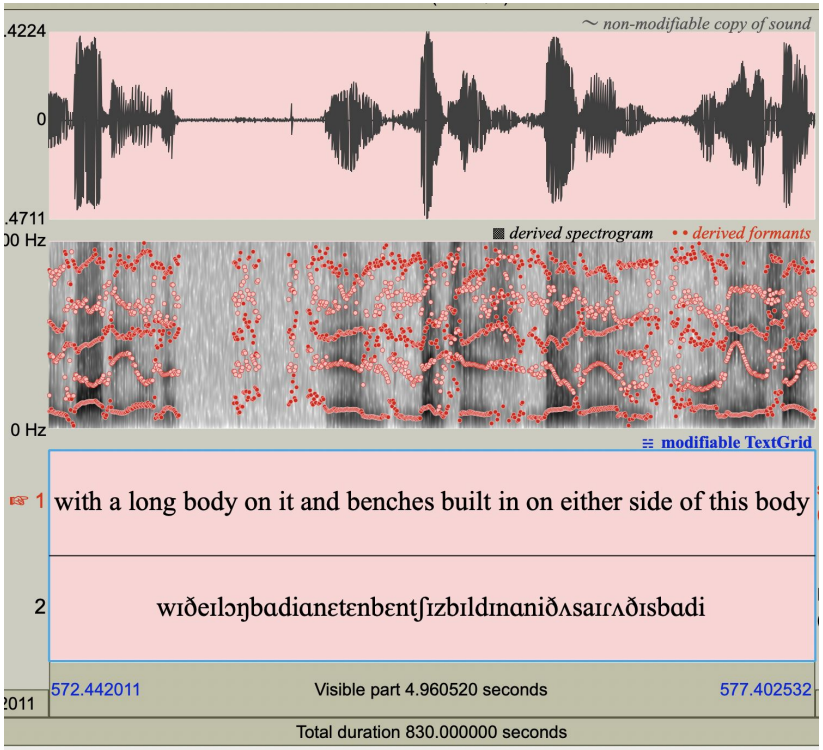
AutoIPA is now [available on hugging face](), with support for Praat text grids

We are now working on outputting phone alignments

Transcription of JP (quasi-Canadian) saying “I put my **cat** on a **cot**. I put my **cat** on a **cot**. I put my **cot** on a **cat**. I put my **cot** on a **cat**”:

aɪ pʊʔ maɪ **kɑ** ɑnʌ **kɔ** ʌ pʊʔ m **kɑ** ɑn nʌ **kɔʔ** ʌ pʊʔ maɪ **kɔt** ɑn **kɑʔ** ʌ pʊʔ maɪ **kɔʔ** ɑn nʌ **kɑp**

It uses [æ] for JP’s “ran”.



Next steps

Wav2Vec 2.0 has been used as a classifier by Kim et al. 2024 for nasalization and by Tanner et al. 2025 for stop realizations

AutoIPA will likely be useful as a pretrained model to be fine-tuned for that kind of work

It likely also has many applications out of the box (especially in the study of word-final consonant realizations)

Our next step is in applying it to the study of phonological variation at the “border” of western and eastern New England dialect regions in Western Massachusetts

We foresee an iterative approach, getting first pass transcriptions from AutoIPA, correcting them, and then fine-tuning our model with the new transcriptions

Notes on the standard TIMIT phone reduction protocol

Acknowledgments

We gratefully acknowledge the support of the Center for Data Science and Artificial Intelligence, the HFA/CICS Collaborative Seed Fund, and the Public Interest Technology Initiative, all at UMass Amherst.