Supplemental materials for "Automated phonetic transcription for varieties of English: wav2vec 2.0 fine-tuned on the Buckeye Corpus"

Virginia Partridge, Joe Pater, Parth Bhangla, Ali Nirheche & Brandon Prickett
University of Massachusetts Amherst

1 Buckeye Corpus: Train, validation and test splits

We divided the Buckeye Corpus into subsets for training, validation and a held-out test set by speaker, so that success on test set requires generalizing to new speakers.

Our training set consists of 18,344 samples from 24 speakers, the validation set 5522 samples from 8 speakers, and the test set 5079 samples from 8 speakers. The ratio of speaker demographics for age and gender were maintained across each subset, resulting in the following quantities for each:

Data subset	Gender	Age group	Total utterance samples	Total audio duration (seconds)
Training	Woman	Over 40	4538	13043.31
		Under 30	3714	11658.41
	Man	Over 40	4694	13501.34
		Under 30	5398	12020.30
Validation	Woman	Over 40	1670	3500.77
		Under 30	1187	3432.34
	Man	Over 40	1382	4725.58
		Under 30	1283	3292.19
Test	Woman	Over 40	1712	4708.69
		Under 30	1132	4098.10
	Man	Over 40	982	4444.14
		Under 30	1253	4767.58

The IDs of speakers included in each split of the dataset are as follows:

- **Training:** S01, S02, S03, S04, S05, S06, S07, S08, S09, S10, S11, S12, S13, S14, S15, S16, S17, S19, S21, S22, S23, S24, S28, S30
- **Validation:** S18, S20, S26, S29, S31, S32, S33, S35
- **Test:** S25, S27, S34, S36, S37, S38, S39, S40

2 Symbol Sets Mapped to IPA and Reduced Symbol Sets

The following table shows the mappings from the original Buckeye and TIMIT symbol inventories to IPA and the shared IPA symbol set we used for evaluation. We only re-mapped the IPA symbols that appear

Proceedings of *AMP 2025*

in the first column. Any additional IPA symbols or diacritics that were output by a model were left as is in the final prediction for the purposes of calculating evaluation metrics.

Full IPA	Buckeye	TIMIT	Shared reduced IPA symbol set
ɑ	AA	AA	ɑ
æ	AE	AE	æ
ʌ	AH	AH	ə
ɔ	AO	AO	ɔ
aʊ	AW	AW	aʊ
aɪ	AY	AY	aɪ
ɛ	EH	EH	ɛ
ɹ̩	ER		ɹ̩
eɪ	EY	EY	eɪ
ɪ	IH	IH	ɪ
i	IY	IY	i
oʊ	OW	OW	oʊ
ɔɪ	OY	OY	ɔɪ
ʊ	UH	UH	ʊ
u	UW	UW	u
b	B	B	b
tʃ	CH	CH	tʃ
d	D	D	d
ð	DH	DH	ð
l̩	EL	EL	l̩
m̩	EM	EM	m̩
n̩	EN	EN	n̩
f	F	F	f
ɡ	G	G	ɡ
h	HH	HH	h
dʒ	JH	JH	dʒ
k	K	K	k
l	L	L	l
m	M	M	m
n	N	N	n
ŋ	NG	NG	ŋ
p	P	P	p
ɹ	R	R	ɹ
s	S	S	s
ʃ	SH	SH	ʃ
t	T	T	t
θ	TH	TH	θ
v	V	V	v
w	W	W	w
j	Y	Y	j

z	Z	Z	z
ʒ	ZH	ZH	ʒ
æ̃	AEN		æ
ɔ̃	AON		ɔ
ə̃	AXN		ə
ĩ	IYN		i
ẽĩ	EYN		eɪ
õʊ̃	OWN		oʊ
ɾ	DX	DX	ɾ
ãĩ	AYN		aɪ
ɑ̃	AAN		ɑ
ũ	UWN		u
ñ	NX	NX	n
ə	AX	AX	ə
ɛ̃	EHN		ɛ
ʊ̃	UHN		ʊ
ãʊ̃	AWN		aʊ
ʌ̃	AHN		ə
ʔ	TQ	Q	ʔ
ɪ̃	IHN		ɪ
ɻ̃	ERN		ɻ
ɔĩ	OYN		ɔɪ
β	BF		f
ɝ		ER	ɻ
ə̥		AX-H	ə
ɚ		AXR	ɻ
ɨ		IX	ɪ
ɦ		HV	h
ŋ̊		ENG	ŋ̊
ʉ̥		UX	u
ː			
		BCL	b
		BCL B	b
		DCL	d
		DCL D	d
		DCL JH	dʒ
		GCL	ɡ
		GCL G	ɡ
		KCL	k
		KCL K	k
		PCL	p
		PCL P	p
		TCL	t

		TCL T	t
		TCL CH	tʃ