VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis

Emily P. Ahn and **Eleanor Chodroff**
University of Washington *University of York*

Scan QR to download the **corpus** or visit osf.io/t957v

1. Motivation

> **Extend Phonetic Research**
– Study cross-linguistic phonetic systematicity & variation

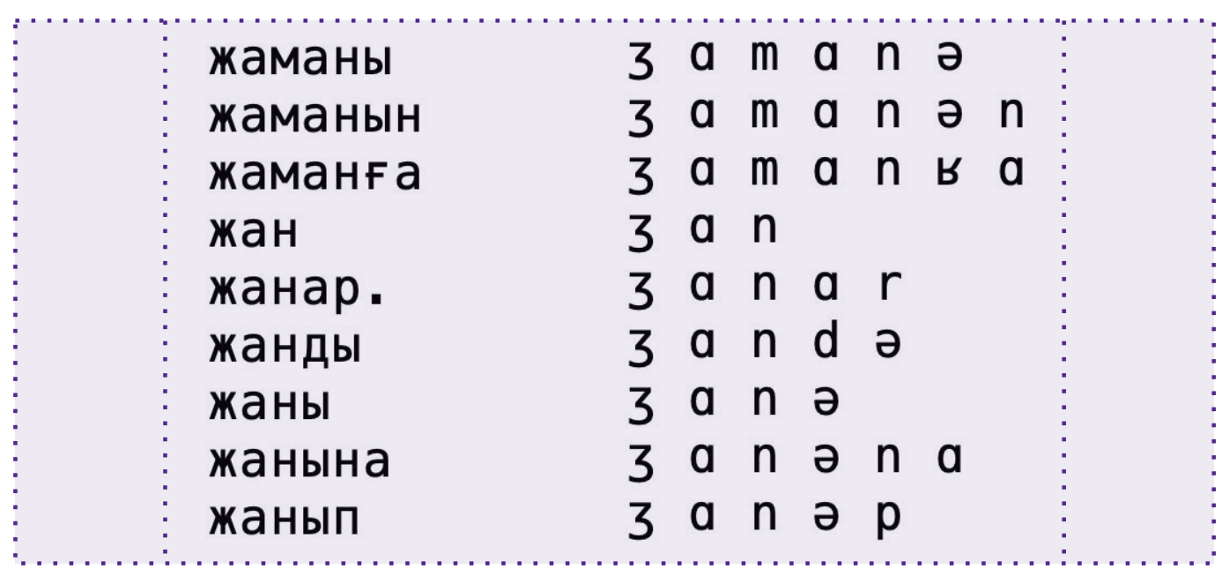
> **Improve Language Technologies**
– Increase coverage over diverse language varieties

>> **Develop Mozilla Common Voice**[1]
— Web-collected, validated read speech in 90+ languages

2. Corpus Contents

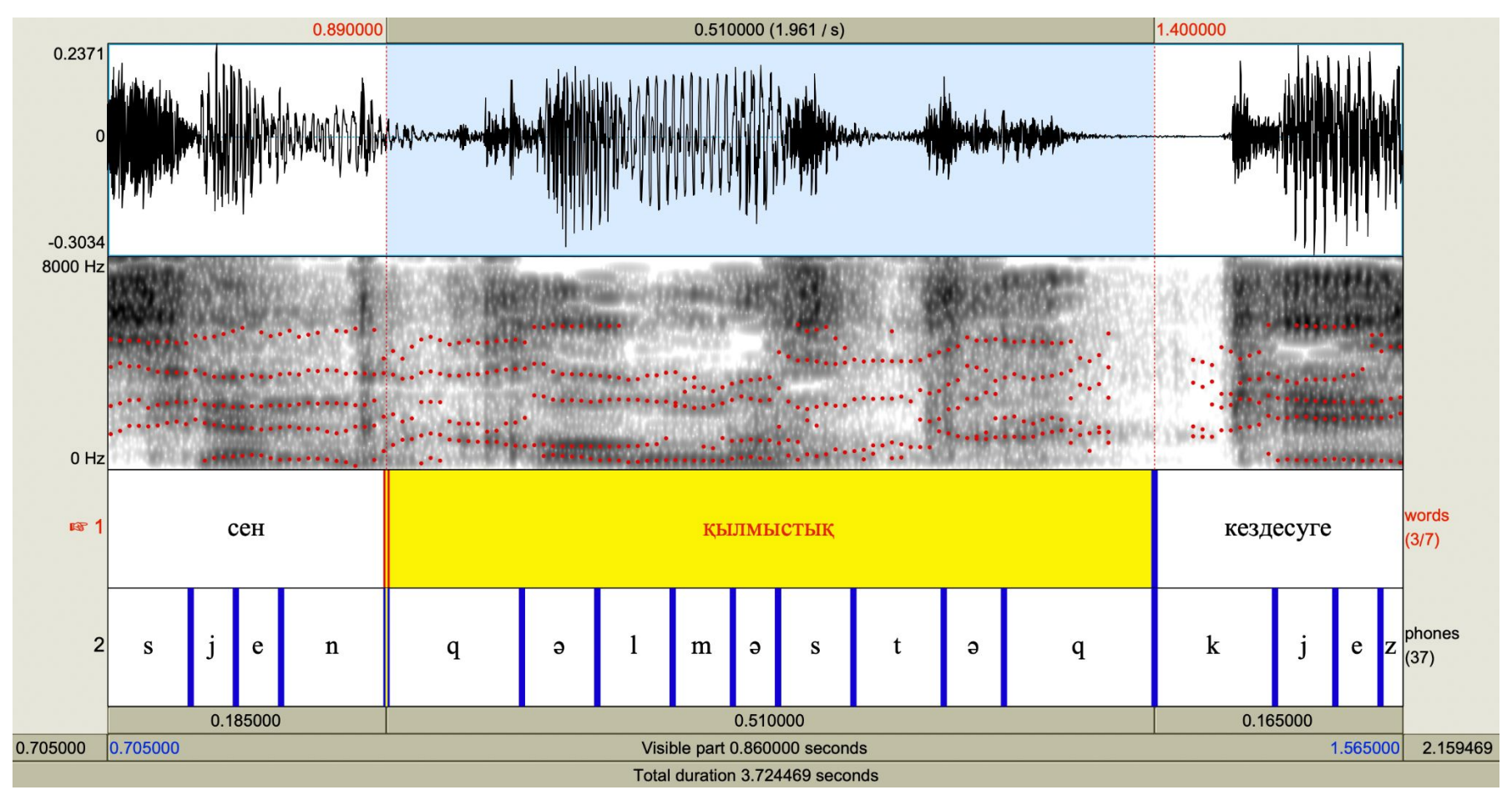
From 36 languages in the Common Voice[1] corpus (v7), ***VoxCommunis*** provides:

1. **Acoustic models***
2. **Pronunciation lexicons***



жаманы	ʒ ɑ m ɑ n ə
жаманын	ʒ ɑ m ɑ n ə n
жаманға	ʒ ɑ m ɑ n ʁ ɑ
жан	ʒ ɑ n
жанар.	ʒ ɑ n ɑ r
жанды	ʒ ɑ n d ə
жаны	ʒ ɑ n ə
жанына	ʒ ɑ n ə n ɑ
жанып	ʒ ɑ n ə p

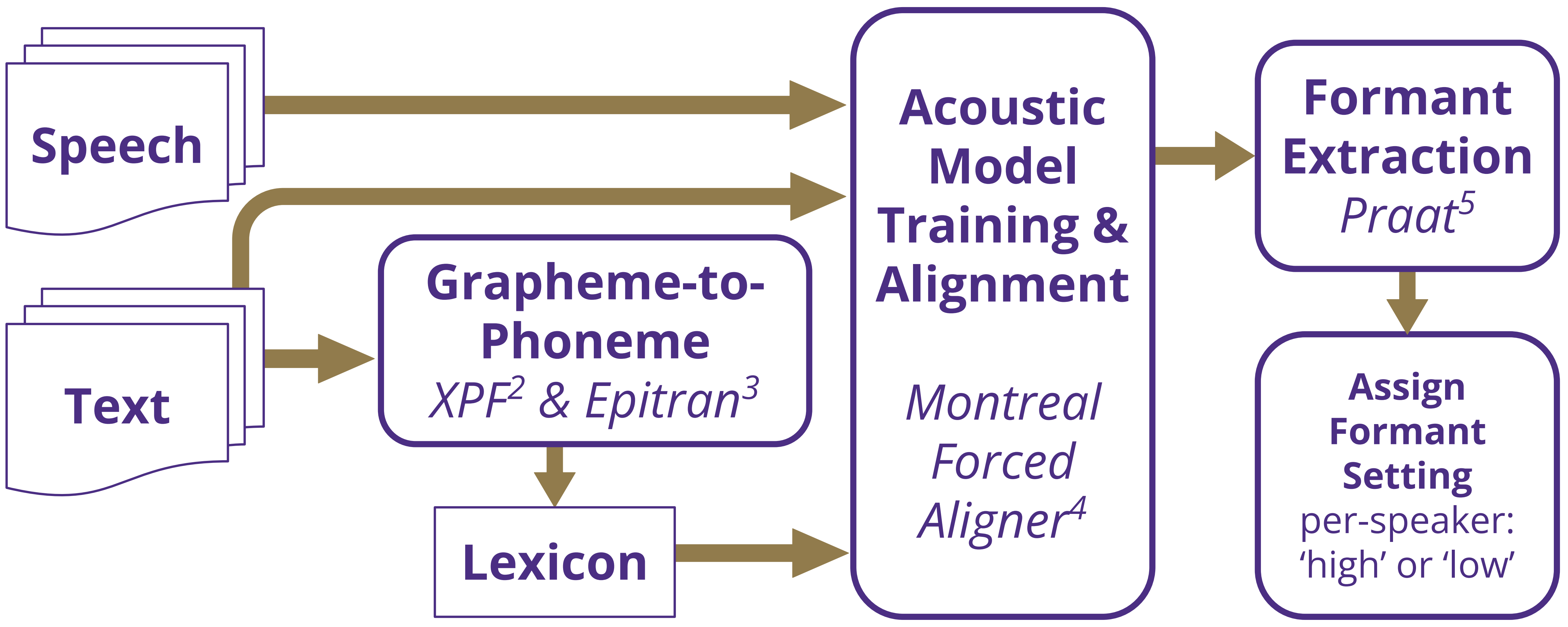
3. **Word- and phone-level alignments**



4. **Extracted vowel formants**
(F1–F4 at vowel quartiles)

**Available with Montreal Forced Aligner[4]*

3. Methodology



Speech

Text

Grapheme-to-Phoneme
XPF[2] & Epitran[3]

Lexicon

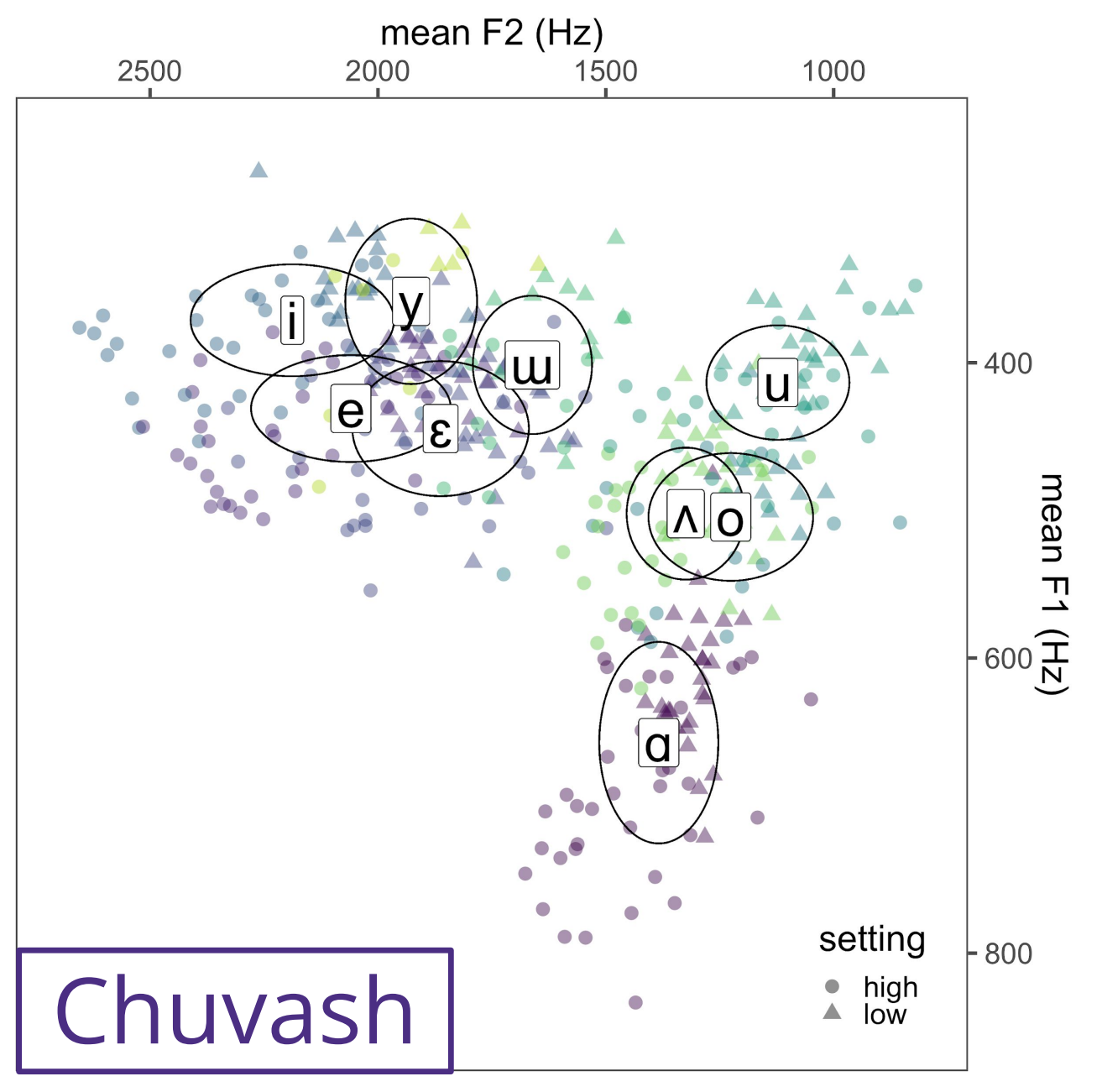
Acoustic Model Training & Alignment
Montreal Forced Aligner[4]

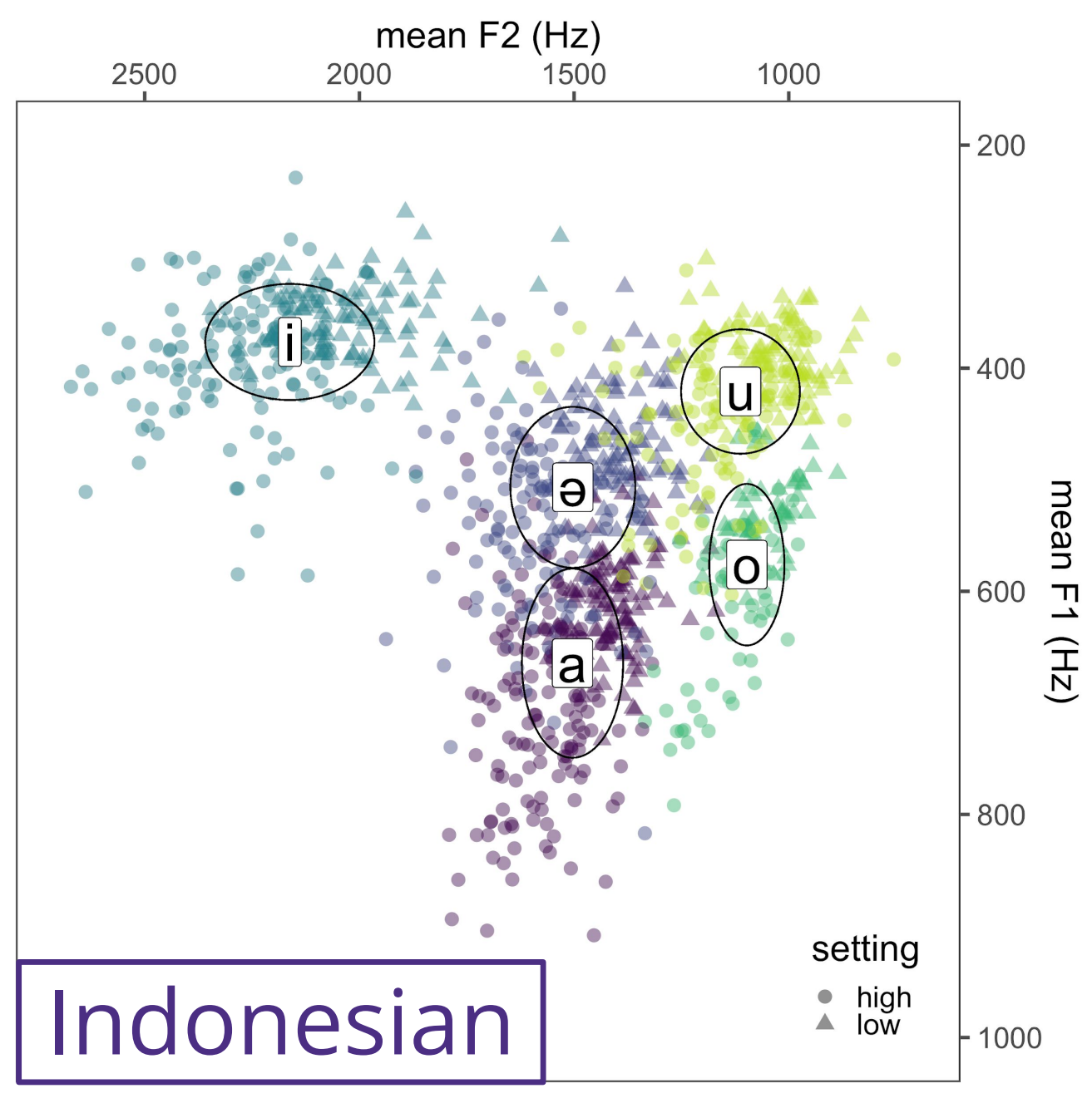
Formant Extraction
Praat[5]

Assign Formant Setting
per-speaker: 'high' or 'low'

4. Data

Language	Hours	Speakers	Utts	G2P	# V	# C	ISO 639-3	Genus	Family
Abkhaz	2	28	1166	XPF	2	55	abk	Northwest Caucasian	Northwest Caucasian
Armenian	1	22	767	XPF	6	30	hye	Armenian	Indo-European
Bashkir	247	835	200869	XPF	9	28	bak	Turkic	Altaic
Basque	91	842	63916	XPF	5	24	eus	Basque	Basque
Belarusian	91	3620	182840	XPF	5	36	bel	Slavic	Indo-European
Bulgarian	5	35	3459	XPF	6	21	bul	Slavic	Indo-European
Chuvash	5	82	3748	XPF	8	14	chv	Turkic	Altaic
Czech	49	475	41567	XPF	5	25	ces	Slavic	Indo-European
Dutch	93	1315	79153	Epi	17	23	nld	Germanic	Indo-European
Georgian	6	109	4562	XPF	5	27	kat	Kartvelian	Kartvelian
Greek	13	178	11609	XPF	5	18	ell	Greek	Indo-European
Guarani	0.53	32	432	XPF	12	17	gug	Tupi-Guaraní	Tupian
Hausa	1	13	1535	Epi	5	23	hau	West Chadic	Afro-Asiatic
Hindi	8	168	6805	Epi	12	41	hin	Indic	Indo-European
Hungarian	16	116	12529	XPF	14	25	hun	Ugric	Uralic
Indonesian	23	273	20649	Epi	5	24	ind	Malayo-Sumbawan	Austronesian
Italian	288	6125	194504	Epi	7	20	ita	Romance	Indo-European
Kazakh	0.73	57	532	Epi	10	26	kaz	Turkic	Altaic
Kurmanji Kurdish	45	258	37019	Epi	9	29	kmr	Iranian	Indo-European
Kyrgyz	37	206	29107	Epi	8	20	kir	Turkic	Altaic
Maltese	8	149	6195	Epi	6	25	mlt	Semitic	Afro-Asiatic
Polish	129	498	105585	Epi	8	28	pol	Slavic	Indo-European
Portuguese	84	1638	71155	Epi	10	25	por	Romance	Indo-European
Punjabi	1	22	1124	Epi	10	33	pan	Indic	Indo-European
Romanian	11	192	10351	XPF	7	20	ron	Romance	Indo-European
Russian	148	1609	99513	Epi	6	22	rus	Slavic	Indo-European
Sorbian (Upper)	2	18	1381	XPF	8	30	hsb	Slavic	Indo-European
Swedish	35	594	32626	Epi	17	21	swe	Germanic	Indo-European
Tamil	198	521	115193	Epi	10	24	tam	Southern Dravidian	Dravidian
Tatar	28	187	27416	XPF	10	23	tat	Turkic	Altaic
Thai	133	4537	107728	Epi	19	21	tha	Kam-Tai	Tai-Kadai
Turkish	30	850	29606	XPF	8	20	tur	Turkic	Altaic
Ukrainian	56	580	41056	XPF	6	32	ukr	Slavic	Indo-European
Uyghur	41	281	24970	Epi	8	29	uig	Turkic	Altaic
Uzbek	0.24	5	161	Epi	6	25	uzb	Turkic	Altaic
Vietnamese	3	76	2927	XPF	9	26	vie	Viet-Muong	Austro-Asiatic





5. Case Study

Are languages uniform in their realizations of vowels?

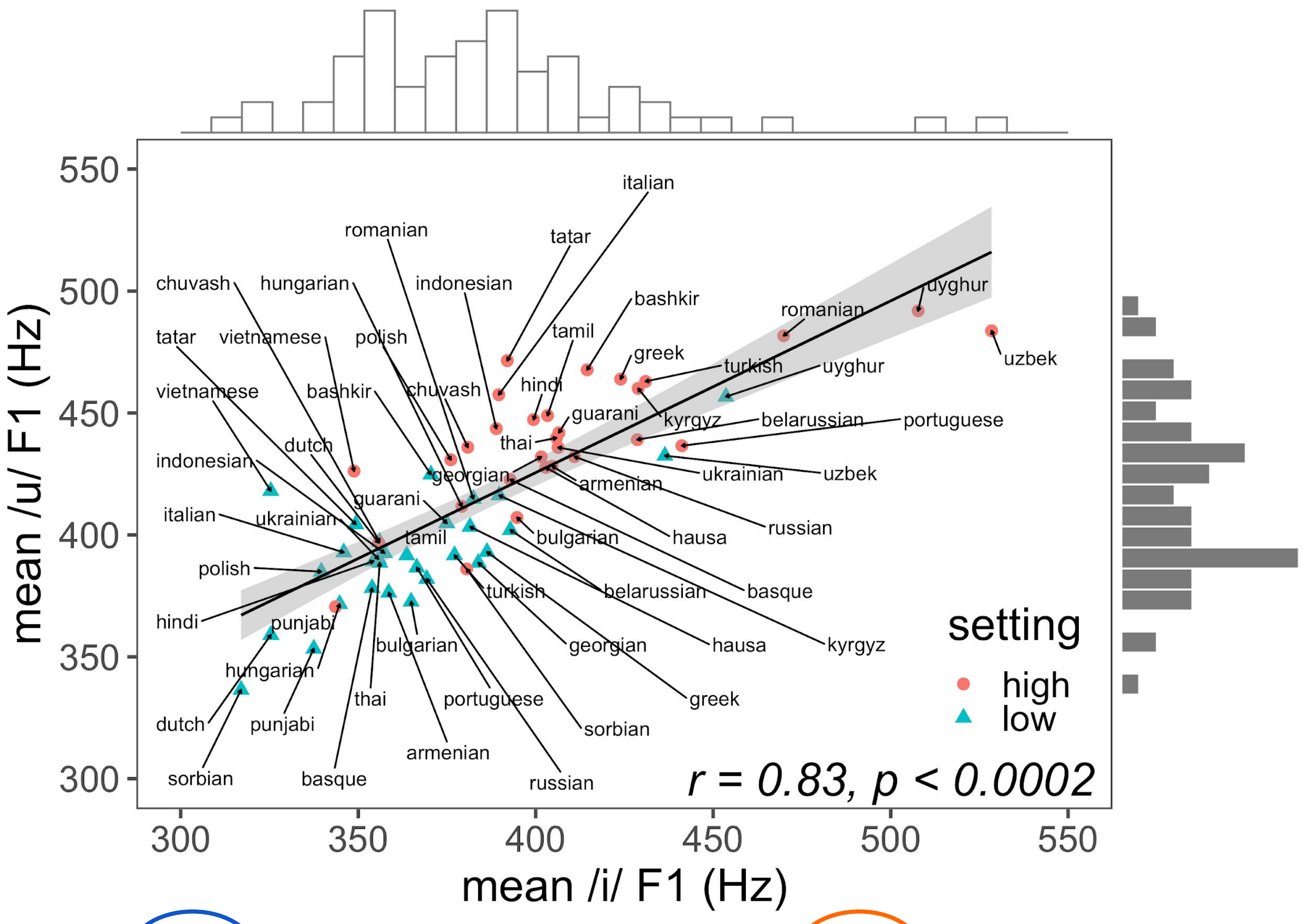
> Uniformity: the expression of a vowel feature should be the same *within* a language

Hypothesis 1: Vowel F1 is correlated for vowel segments with the same height across languages

> e.g. [i] and [u] are 'high' vowels and should be correlated along F1

Hypothesis 2: Vowel F2 is correlated for vowel segments with the same backness across languages

> e.g. [i] and [ɛ] are 'front' vowels and should be correlated along F2



V1	V2	Height	# Lang	r	p
i	u	✓	31	0.83	<0.001
e	o	✓	22	0.74	<0.001
e	a		19	0.64	<0.001
ɛ	ɔ	✓	14	0.64	<0.001
o	a		22	0.53	<0.001
u	o		28	0.46	<0.001

V1	V2	Back	# Lang	r	p
ɛ	a	✓	11	0.77	<0.001
i	ɔ		13	0.74	<0.001
i	ɑ		12	0.67	<0.001
i	a	✓	24	0.67	<0.001
e	a	✓	19	0.58	<0.001
i	ɛ	✓	18	0.57	<0.001

6. Conclusion

Future Work

1. Expand this resource
2. Improve automated tools (e.g. G2P)

Corpus Applications

> Apply acoustic models to ASR and forced alignment of phonetic data

> Test additional phonetic & phonological theories such as Dispersion Theory

References

1. Ardila et al. (2020). Common Voice: A massively-multilingual speech corpus. In *LREC*.
2. Cohen Priva et al. (2021). The Cross-linguistic Phonological Frequencies (XPF) Corpus.
3. Mortensen et al. (2018). Epitran: Precision G2P for many languages. In *LREC*.
4. McAuliffe et al. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*.
5. Boersma & Weenink. (2019). Praat: Doing phonetics by computer [computer program]. Version 6.1.08.

Acknowledgments

We are grateful to Michael McAuliffe, Richard Wright, Gina-Anne Levow, Phil Harrison, and the VoxClamantis team. This work was supported in part by NSF GRFP grant DGE-2140004.