Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5286–5294
Marseille, 20-25 June 2022

VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis

Emily P. Ahn, Eleanor Chodroff
University of Washington, University of York
Washington, USA, York, UK
eahn@uw.edu, eleanor.chodroff@york.ac.uk

Abstract

Cross-linguistic phonetic analysis has long been limited by data scarcity and insufficient computational resources. In the past few years, the availability of large-scale cross-linguistic spoken corpora has increased dramatically, but the data still require considerable computational power and processing for downstream phonetic analysis. To facilitate large-scale cross-linguistic phonetic research in the field, we release the VoxCommunis Corpus, which contains acoustic models, pronunciation lexicons, and word- and phone-level alignments, derived from the publicly available Mozilla Common Voice Corpus (Ardila et al., 2020). The current release includes data from 36 languages. The corpus also contains acoustic-phonetic measurements, which currently consist of formant frequencies (F1–F4) from all vowel quartiles. Major advantages of this corpus for phonetic analysis include the number of available languages, the large amount of speech per language, as well as the fact that most language datasets have dozens to hundreds of contributing speakers. We demonstrate the utility of this corpus for downstream phonetic research in a descriptive analysis of language-specific vowel systems, as well as an analysis of "uniformity" in vowel realization across languages. The VoxCommunis Corpus is free to download and use under a CC0 license at `https://osf.io/t957v/wiki/home/`.

Keywords: phonetics, typology, G2P, alignment, speech

1. Introduction

For a thorough understanding of cross-linguistic phonetic variation and systematicity, big data from a diverse set of languages is necessary. In 2009, it was noted that despite the advances made in speech technology and computational power, there had been "surprisingly little change in style and scale of [phonetic] research" from 1966 onwards (Liberman, 2009). Since even that period, considerable advances have been made for increased processing power of large-scale phonetic corpora, but largely within a single "high-resource" language such as English, in which relevant data for automatic speech processing approaches already exist.

Until recently, the movement towards large-scale, cross-linguistic phonetic research has been somewhat limited. Previous large-scale cross-linguistic phonetic studies have mostly been meta-analyses that rely on a standardized phonetic measure and a plethora of published cross-linguistic research (Whalen and Levitt (1995) for vowel f0, Becker-Kristal (2010) for vowel F1 and F2, Chodroff et al. (2019) for stop VOT). Prior to 2020, existing multilingual speech corpora contained maximally twenty-some languages (Harper, 2011; Schultz et al., 2013), or insufficient data for most phonetic analyses (Ladefoged and Maddieson, 2007). More recently, Salesky et al. (2020) presented the VoxClamantis Corpus for large-scale phonetic typology that provided phonetic data for over 500 languages, based on recordings and transcripts from the CMU Wilderness Corpus (Black, 2019). Only in the past few years (e.g., from 2019) have these massively multilingual corpora been made publicly available (see Section 2 for an overview). In addition, the tools necessary to process such data for phonetic analysis have been considerably improved and expanded in utility and coverage.

In the present paper, we introduce the **VoxCommunis Corpus** for large-scale cross-linguistic phonetic analysis based on the **Mozilla Common Voice Corpus**[1] (Ardila et al., 2020). The Mozilla Common Voice corpus is a publicly available, multilingual speech corpus that contains spoken utterances collected via the internet on both web and phone platforms. Version 7.0 (released July 2021) has data for approximately 75 languages. Each language contains anywhere from around 50 MB to over 100 GB of audio data, and anywhere from three to over a hundred speakers. A subset of utterances is additionally "validated" by users, indicating that at least two users have confirmed that the reading of a specific utterance is faithful to the corresponding written text. The corpus is freely available for download and academic use. It is also community-driven with active maintenance and updates, meaning that the size of the corpus is regularly increasing.

This corpus can facilitate research in language-specific phonetics and phonetic typology, which can in turn improve speech technologies. As spoken language technologies are becoming increasingly common, their effectiveness and coverage over diverse language varieties have become more and more important. Good automatic speech recognition (ASR) systems and text-to-speech (TTS) systems rely on accurate knowledge and implementation of phonetics and phonology–the studies of the production and perception of speech sounds

[1] `https://commonvoice.mozilla.org/en/datasets`

and how they are organized. Improved understanding of the universals and variation in phonetic realization can inform the development of such speech technology, particularly for low-resource languages.

2. Related Work

Aside from Common Voice, several multilingual speech corpora have been developed and some may also prove useful for downstream phonetic analysis. As mentioned in Section 1, the IARPA BABEL and GlobalPhone corpora were some of the first large, multilingual spoken corpora in research and have been popularly used for ASR system development. In 2011, the IARPA BABEL corpus was released with transcribed conversational telephone speech in 21 diverse languages (Harper, 2011) . In 2013, the GlobalPhone corpus was released with over 400 hours of read speech audio across 20 languages and around 100 speakers per language (Schultz et al., 2013). These corpora, however, are neither public nor open source.

More recently, several multilingual spoken corpora have been released that are publicly available, though may still have some minor copyright restrictions. These include the CMU Wilderness Corpus and derivative VoxClamantis Corpus, VoxPopuli, multilingual TEDx, and multilingual LibriSpeech.

Most similar in approach to the present corpus here is the VoxClamantis Corpus, derived from the massively multilingual CMU Wilderness Corpus. The CMU Wilderness corpus is a collection of audio recordings in nearly 700 languages of the New Testament, with around 20 hours of speech per language (Black, 2019). Building off of this corpus, the VoxClamantis dataset contains an initial pass of utterance- and phoneme-level alignments of the readings, along with a preliminary set of vowel formants and sibilant fricative spectral properties (Salesky et al., 2020). This Bible corpus is the largest spoken corpus in terms of range of language variation, including some severely low-resource languages. While having a wide language coverage, each language reading has very few speakers, most of whom are male. This presents a limitation for phonetic analysis as there can be confounds between speaker and language variation. In addition, some copyright restrictions limit the accessibility of the audio.[2]

The VoxPopuli Corpus contains 400,000 hours of unlabeled speech data spanning 23 languages taken from European parliament recordings (Wang et al., 2021). Around 17,000 hours of speech are transcribed, which provides a springboard for much of the processing described here. Finally, the multilingual TEDx and multilingual LibriSpeech corpora contain over 700 and 36,000 hours of speech respectively, from 8 European languages (Salesky et al., 2021; Pratap et al., 2020). Relative to the Wilderness and Common Voice corpora, these corpora have less language coverage; nevertheless, they are publicly available, mostly transcribed, and therefore will likely be very useful for future phonetic analysis.

3. Methodology

The primary goal of our data processing was to obtain word- and phone-level forced alignments for each recording to facilitate acoustic-phonetic measurement. Our processing targeted language datasets for which grapheme-to-phoneme toolkits were available, had less than 300 hours of validated data (due to space and processing power limitations), and focused only on the "validated" utterances of each dataset. Through this process, we developed language-specific pronunciation lexicons and acoustic models in addition to alignments. These resources may have independent utility for phonetic research, such as language-specific forced alignment of new speech data or improved pronunciation lexicon development.

3.1. Grapheme-to-Phoneme Conversion

A major bottleneck in cross-linguistic phonetic research is the conversion of orthographic transcripts to their corresponding phonetic (or phonemic) forms. This process is known as grapheme-to-phoneme (G2P) conversion, and can be accomplished in a variety of manners and with a variety of assumptions regarding the transcription granularity.[3] We relied on two linguist-designed, rule-based G2P systems for this conversion: Epitran and the XPF Corpus. These systems have been developed for many languages with a "transparent" orthography, in which the orthographic representation is systematically related to its phonemic form.

Epitran is a publicly available G2P toolkit that supports conversion for around 60 languages (Mortensen et al., 2018). Its architecture consists of finite state transducers, all manipulated via a Python interface. The Cross-linguistic Phonological Frequencies (XPF) Corpus is a resource of phonemic lexicons or grammars for over 200 spoken languages with a web interface for rule-based G2P conversion (Cohen Priva et al., 2021).[4]

The output of these systems is a pronunciation lexicon for each language dataset with mappings between each word form and its corresponding phonemic transcription.

Based on the intersection of languages between Common Voice, Epitran, and the XPF Corpus, we could produce pronunciation lexicons for approximately 40 language datasets. Four of these were removed due to

[2] Though it is accessible through the bible.is website, each language dataset must be downloaded individually.

[3] The G2P systems employed here appear to be using broad phonetic transcriptions that best correspond to sequences of phonemes or surface phonological segments. In the present paper, we refer to these sound segments as phonemes.

[4] https://cohenpr-xpf.github.io/XPF/Convert-to-IPA.html

poor G2P quality or processing issues in the acoustic model development. These were two Chinese language datasets, Persian, and Arabic.[5] In the resulting corpus, 18 language datasets were processed using Epitran and 18 using the XPF Corpus. Among the languages processed using Epitran, Hindi and Tamil were processed using dual G2P models to process the predominant Indic-based script and secondary English romanized script.[6] We manually updated several entries across most datasets, especially where foreign characters or loanwords were included in the input.

3.2. Acoustic Model Training

The acoustic models were developed using the generated pronunciation lexicons and the Montreal Forced Aligner (MFA). Among other applications, MFA provides a user-friendly wrapper to the Kaldi ASR toolkit (Povey et al., 2011) for acoustic model training and alignment (McAuliffe et al., 2017).

For each language with a pronunciation lexicon, we trained an acoustic model using default settings from MFA version 2.0.09b. This training used a GMM-HMM based system with various levels of speaker and channel adaptation. MFCCs were extracted from 25 ms windows with a 10 ms frame shift and then normalized using cepstral mean and variance normalization. The acoustic models were then constructed using 40 iterations of monophone training and alignment with a maximum 1000 Gaussians. These were followed by 35 iterations of LDA+MLLT and two rounds of speaker-adapted training (SAT) with fMLLR, each with a maximum 2500 leaves and 15000 Gaussians.[7] As each recording was annotated with a speaker identification code, this information was used to inform acoustic model training.

All audio files were available in 16-bit MP3 format, single channel, with a 32 kHz sampling rate.[8] Preprocessing steps included converting MP3 to WAV format and creating Praat TextGrids populated with utterances. Each recording is a standard sentence-length utterance. The word- and phone-level forced alignments were extracted directly from the final acoustic model for each language. The alignments are released as Praat TextGrids.

3.3. Formant Extraction

Formants are concentrations of high acoustic energy in the frequency spectrum, and reflect resonant frequencies in the vocal tract. Because of their relationship to the shape of the oral cavity (articulation) and corresponding perceptual quality (perception), these measures are commonly extracted for a wide variety of phonetic analyses. The first formant (F1) strongly correlates with tongue height and the second formant (F2) with tongue backness (Ladefoged and Johnson, 2014). These two dimensions are highly diagnostic for most vowel contrasts (see for example Figures 1 and 2). F3 frequently correlates with lip rounding, rhoticity, and nasality, and F4 can reflect high front vowel contrasts and aspects of voice quality (House and Stevens, 1956; Lindblom and Sundberg, 1971; Ladefoged et al., 1978; Eek and Meister, 1994).

The first four formant values were extracted from each aligned vowel quartile using the Linear Predictive Coding (Burg method) algorithm implemented in Praat (Boersma and Weenink, 2019). For each formant, values at 10 ms before and after the midpoint were also extracted, for increased accuracy in analysis (see Section 5.1). All formant values were extracted under both "high" and "low" frequency settings. Specifically, the tracker searched for five formants with a ceiling of 5500 Hz in the "high" setting, and a ceiling of 5000 Hz in the "low" setting. These are the recommended ceilings for typical female and male speech, respectively. Since only a small portion of the corpus had gender labels, we performed a simple classification algorithm to assign each speaker in the entire corpus to either the high or low setting. Using a subset of 1200 gender-labeled speakers across 11 languages, we fit two bivariate Gaussians: one distribution over average F1 and F2 values at the high formant tracking setting for speakers labeled as "female", and one distribution over average F1 and F2 values at the low formant tracking setting for speakers labeled as "male". Each speaker was classified as matching either the high or low setting depending on which of the two trained distributions their average (F1, F2) values were closer to. Distance was quantified using the Mahalanobis metric. We release both "high" and "low" sets of extracted formants over all speakers and utterances, but employ this heuristic for the case study data in Section 5.

4. Data

Table 1 lists the Common Voice language datasets that we used and processed in this work, along with several of their characteristics. Quantity of data in terms of hours of audio and number of speakers is based on the validated subset from Common Voice.[9] Table 1 further shows language-related descriptions of each dataset. Vowel and consonant inventory sizes were determined

[5] Although we processed Russian, it is important to note that Epitran may not be reliable in its G2P output for Russian.

[6] The English words in these two lexicons were mapped to English phonology, although the audio often revealed that the pronunciations were more faithful to Hindi or Tamil phonology.

[7] For more detail on the default recipe, see `https://montreal-forced-aligner.readthedocs.io/en/latest/user_guide/configuration/acoustic_modeling.html`.

[8] The compressed MP3 format of the original files may be a limitation for fine-grained acoustic analysis.

[9] Ardila et al. (2020) refer to a validated utterance as one that has a majority of upvotes from crowdsourced listeners who verified that the text transcription matched the audio.

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	Turkic
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	Turkic
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	Turkic
Kurmanji Kurdish	45	258	37019	Epi	9	29	kmr	Iranian	Indo-European
Kyrgyz	37	206	29107	Epi	8	20	kir	Turkic	Turkic
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	Turkic
Thai	133	4537	107728	Epi	19	21	tha	Kam-Tai	Tai-Kadai
Turkish	30	850	29606	XPF	8	20	tur	Turkic	Turkic
Ukrainian	56	580	41056	XPF	6	32	ukr	Slavic	Indo-European
Uyghur	41	281	24970	Epi	8	29	uig	Turkic	Turkic
Uzbek	0.24	5	161	Epi	6	25	uzb	Turkic	Turkic
Vietnamese	3	76	2927	XPF	9	26	vie	Viet-Muong	Austro-Asiatic

Table 1: This release of VoxCommunis includes datasets from 36 languages with hours of speech ranging from 0.24 to 288. Half of these languages were processed with Epitran ("Epi"), and half were processed with XPF G2P methods. Vowel and consonant inventory sizes, ISO 639-3 codes, genus, and family descriptions of each language are included as well. **While Guarani has 12 phonemic vowels, the nasal contrast was not transcribed in the output of the XPF G2P, so our data only reflects 6 vowels.*

by the mappings and rules files from Epitran and the language description pages from XPF.

As an example of the descriptive utility of the corpus for phonetic research, we present two sample vowel charts of the F1 × F2 space: one from Chuvash with an inventory of eight vowels (Figure 1) and one from Indonesian (Figure 2) with an inventory of five vowels. The Chuvash dataset contains five hours of speech from 82 speakers, and the Indonesian dataset contains 23 hours of speech from 273 speakers.

5. Case Study

The data in VoxCommunis can be a testbed for many research questions that concern phonetic and phonological theory. We focus here on cross-linguistic constraints on the phonetics–phonology interface, and specifically a uniformity constraint on phonetic realization. Phonetic realization refers to the mapping from a sound segment's phonological features to the corresponding phonetic targets.

Evidence from cross-linguistic, cross-dialectal, and cross-speaker variation implies a range of permissible phonetic realizations for each segment. Phonetic uniformity builds on a line of previous and related principles posited in the literature that emphasize reuse of phonetic targets that correspond to a phonological primitive (Maddieson, 1995; Keating, 2003; Ménard et al., 2008a; Guy and Hinskens, 2016; Chodroff and Wilson, 2017; Fruehwald, 2017). In essence, uniformity enforces economy and similarity. Similar principles to uniformity can be found in gestural economy, which requires reuse of individual gestures across multiple speech sounds (Lindblom, 1983; Lindblom and Maddieson, 1988; Maddieson, 1995), and the Maximal Use of Available Controls (MUAC) principle, which requires reuse of perceptuomotor controls in the real-

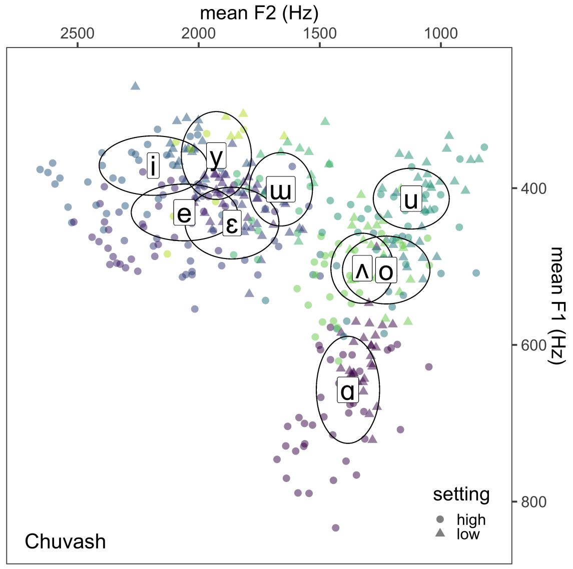

Figure 1: Chuvash vowels in F1×F2 space (Hz). Each point corresponds to a speaker-specific pair of means and is coded for the formant extraction setting. Labels are centered on the grand means for each category, and ellipses correspond to ± one standard deviation from the mean across speakers.

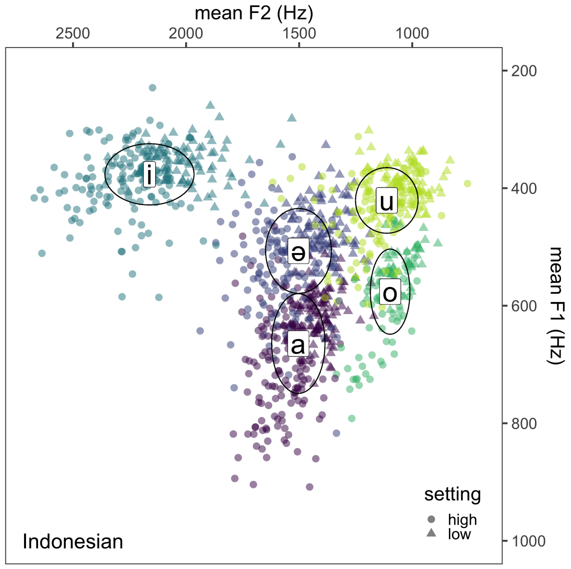

Figure 2: Indonesian vowels in F1×F2 space (Hz). Each point corresponds to a speaker-specific pair of means and is coded for the formant extraction setting. Labels are centered on the grand means for each category, and ellipses correspond to ± one standard deviation from the mean across speakers.

ization of a distinctive feature (Schwartz et al., 2012; Ménard et al., 2008a).

Chodroff and Wilson (2022) extended this notion of uniformity and considered three potential types of uniformity: pattern uniformity, target uniformity, and contrast uniformity to account for the ways in which the specification of phonetic targets may be constrained across talkers and languages. With our data, we explore the influence of target uniformity on the phonetic realization of vowels. Target uniformity requires that the mapping from a distinctive feature value to its corresponding phonetic target be uniform for all phonological surface segments specified with the feature value. We focus here on the realization of vowel height and backness features, with the corresponding phonetic targets approximated using the acoustic measures of vowel F1 and F2.

Though languages will likely differ considerably in the overall phonetic realization of a given distinctive feature, such as vowel height, the set of segments that share the featural segmentation should be strongly correlated with one another and be uniform in realization. Assuming there is underlying identity in phonetic realization, we should observe strong correlations of vowel F1 for vowel segments that are specified with the same height feature. Correspondingly, we should also observe strong correlations of vowel F2 for segments specified with the same backness feature.

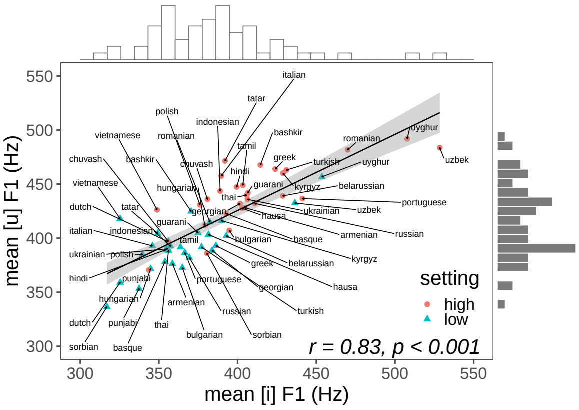

Figure 3: Correlation of mean F1 between [i] and [u] across 30 languages with the best-fit regression line. Each point represents a setting- and language-specific pair of F1 means.

Indeed, previous studies have found some support for these predictions. Vowel F1 is highly correlated between vowels with shared phonological height across speakers within a language (e.g., English: Watt (2000), French: Ménard et al. (2008a), Portuguese: Oushiro (2019), six unique languages: Schwartz and Ménard (2019)). Similar to the present study, Salesky et al. (2020) also examined the predictions of uniformity in vowel F1 and F2 across languages in pairwise correlations over approximately 10 to 40 languages. The correlations of mean F1 were generally strongest between vowels with a shared height, and correspondingly, for mean F2, correlations were generally strongest between vowels with a shared backness feature. The patterns, however, were not perfect, and along the F1 dimension, the correlations were moderate to strong for

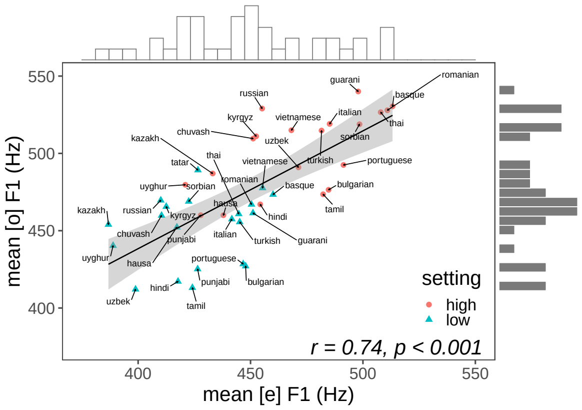

Figure 4: Correlation of mean F1 between [e] and [o] across 22 languages with the best-fit regression line. Each point represents a setting- and language-specific pair of F1 means.

many vowel pairings regardless of their phonological specification. To demonstrate the utility of this corpus for large-scale cross-linguistic phonetic analysis, we look here to replicate these previous findings.

5.1. Methods

For the analysis, we employed the set of formants described in Section 3.3 in which either the high or low formant extraction setting was used for each speaker. Focusing on F1 and F2, we used an average of three formant points at and around the midpoint for each vowel: the value estimated from the midpoint itself and the values 10 ms before and after the midpoint. We removed vowels with an F1 or F2 beyond two standard deviations from the vowel- and setting-specific means within a language, and discarded vowels whose duration was greater than 300 ms, under the assumption that these were alignment or formant-tracking errors. In addition, only vowel categories produced by at least five speakers per high or low setting in a given language were retained in the analysis. We further analyzed correlations for vowel pairings shared by at least ten languages. As there were 22 correlations in each formant analysis for a total of 44 correlations, we adjusted the significance level of $\alpha = 0.05$ to $\alpha = 0.001$ using a Bonferroni correction.

5.2. Results and Discussion

For each of the F1 and F2 analyses, six pairwise correlations reached significance, as shown in Tables 2 and 3. For F1, three of the six significant correlations were consistent with the predictions of target uniformity. We replicate previous significant correlations between [i] and [u], as well as between [e] and [o] across languages (see Figures 3 and 4). Similar to the present analysis, Salesky et al. (2020) also found significant correlations of mean F1 for [i]–[u], [e]–[o], and [e]–[a] pairings across languages. Overall, many F1 correlations were strong in magnitude even if they did not reach significance (see also Salesky et al. (2020)).

For F2, four of the six significant correlations were consistent with the predictions of target uniformity. Though we did observe several significant correlations that were consistent with the predictions of uniformity, the pattern of correlations did not replicate very well between the current VoxCommunis and previous VoxClamantis analyses. The significant correlations of mean F2 for [ɛ]–[a], [i]–[ɔ], and [i]–[a] found here did not reach significance in the VoxClamantis analysis. In fact in VoxClamantis, the correlation for [i]–[ɔ] was opposite in direction (r = -0.63) and the correlation for [i]–[a] was effectively non-existent (r = 0.06); the correlation for [ɛ]–[a] was simply weaker in magnitude at r = 0.32. These discrepancies could be related to a more idiosyncratic realization of F2 patterns across languages, which would nevertheless be insightful for phonetic theory.

Finally, many vowels were also moderately to strongly correlated with [a] along both the F1 and F2 dimensions. In this case, we speculate that the open vocal tract of [a] could be very informative of speaker anatomy, and the correlational strength could reflect anatomical similarity in the productions. That is, the same speaker contributed to the language-specific mean for both [a] and a vowel with which it is correlated (e.g., [e] for F1). Important to note though, is that anatomical similarity is unlikely to account for all strong correlations. The correlational strengths vary widely across vowel pairings, regardless of the fact that each speaker contributed to each of the means.

Overall, the current analysis replicated many of the F1 findings of previous analyses, such as the VoxClamantis analysis in Salesky et al. (2020), among others (Ménard et al., 2008b; Schwartz and Ménard, 2019; Oushiro, 2019; Watt, 2000). Though some F2 correlations were consistent with the predictions of target uniformity, so were several different ones in the VoxClamantis analysis. Moreover, some of the significant F2 correlations found here were entirely different in nature (e.g., magnitude and even direction) in the VoxClamantis analysis. There are several potential factors that may have led to this discrepancy. First, the VoxClamantis had converted formants into ERB units (logarithmic), whereas the present study assessed formant variation in the hertz (linear) space. This could have an outsized impact on the higher F2 values relative to the lower F1 values. Second and not incompatibly with the first point, these findings may simply reflect an overall more idiosyncratic realization of vowel F2 across languages. It could be that target uniformity does not apply consistently to vowel backness, or that the assumed phonological and/or phonetic specifications is ill-defined for vowel categories. At the phonological level, we had assumed the existence of a vowel backness feature, and at the phonetic level, we assumed that F2 would be a reasonable approximation of the phonetic target corresponding to the vowel backness fea-

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

Table 2: Pearson correlations (r) of **mean F1** in Hz between vowel categories. Only correlations that reached significance after Bonferroni correction ($\alpha = 0.001$) are shown here. As formalized in the present analysis, phonetic uniformity predicts strong correlations of mean F1 among vowels with a shared height specification, which is indicated by the checkmark in the table.

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

Table 3: Pearson correlations (r) of **mean F2** in Hz between vowel categories. Only correlations that reached significance after Bonferroni correction ($\alpha = 0.001$) are shown here. As formalized in the present analysis, phonetic uniformity predicts strong correlations of mean F2 among vowels with a shared backness specification, which is indicated by the checkmark in the table.

ture. These assumptions warrant additional research.

6. Conclusion

The VoxCommunis corpus aims to facilitate large-scale phonetic analyses, with the peripheral goal of improving speech technologies for a broad range of languages. We described our methods for processing 36 language datasets from the Common Voice corpus, including G2P conversion, acoustic model training, and vowel formant feature extraction. We presented our data with descriptive and quantitative measures, and highlighted the utility of VoxCommunis in a cross-linguistic case study of phonetic uniformity.

Future directions include expanding this resource in several ways. Phonetic transcriptions for the same phoneme can vary across languages depending on an individual linguist's or a particular resource's preference. Because of this ambiguity, G2P tools have the potential to improve in quality (or accuracy) and in coverage. Employing other existing pronunciation lexicons and G2P tools (e.g. WikiPron (Lee et al., 2020) and Phonetisaurus (Novak et al., 2016)) on the Common Voice corpus would be beneficial. Where overlap of language coverage between G2P systems occurs, one could also compare the quality of these methods.

Scientifically, future analyses could include testing phonetic and phonological theories like Dispersion Theory, which predicts that phonemes in a language's inventory are maximally "dispersed" across phonetic space in order to preserve perceptual distinction (Liljencrants and Lindblom, 1972). Additional research on phonetic uniformity (i.e., as it applies to different segment–target pairings) is also warranted. As VoxCommunis is made freely available to our broader scientific communities, it can inform additional typological or even language-specific studies at other linguistic levels (e.g., syntax, morphology, etc.). Phonetic insight stemming from this corpus may inform and improve automatic speech recognition, text-to-speech systems and automatic speaker adaptation processes, especially for languages with few linguistic resources.

7. Acknowledgments

We are grateful to the VoxClamantis team for setting the foundation for future phonetic corpus development. Specifically, we thank Liz Salesky, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Jason Eisner, and Alan Black. We are also incredibly grateful to Michael McAuliffe for his generous assistance with the Montreal Forced Aligner. We also thank Richard Wright, Gina-Anne Levow, and Phil Harrison for helpful discussion, particularly regarding the formant analysis. This work was supported in part by NSF GRFP grant DGE-2140004.

8. Bibliographical References

Becker-Kristal, R. (2010). *Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus*. Ph.D. thesis, University of California, Los Angeles.

Boersma, P. and Weenink, D. (2019). Praat: Doing phonetics by computer [computer program]. Version 6.1.08.

Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47.

Chodroff, E. and Wilson, C. (2022). Uniformity in phonetic realization: Evidence from sibilant place of articulation in American English. *Language*.

Chodroff, E., Golden, A., and Wilson, C. (2019). Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization. *The Journal of the Acoustical Society of America*, 145(1):EL109–EL115.

Eek, A. and Meister, E. (1994). Acoustics and perception of Estonian vowel types. *Phonetic Experimental Research*, XVIII:146–158.

Fruehwald, J. (2017). The role of phonology in phonetic change. *Annual Review of Linguistics*, 3:25–42.

Guy, G. R. and Hinskens, F. (2016). Linguistic coherence: Systems, repertoires and speech communities. *Lingua*, 172(173):1–9.

House, A. S. and Stevens, K. N. (1956). Analog studies of the nasalization of vowels. *The Journal of Speech and Hearing Disorders*, 21(2):218–232.

Keating, P. A. (2003). Phonetic and other influences on voicing contrasts. In Maria-josep Solé, et al., editors, *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 20–23, Barcelona, Spain.

Ladefoged, P. and Johnson, K. (2014). *A Course in Phonetics*. Nelson Education.

Ladefoged, P. and Maddieson, I. (2007). *The UCLA Phonetics Lab Archive*. UCLA Department of Linguistics, Los Angeles, CA.

Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L. (1978). Generating vocal tract shapes from formant frequencies. *The Journal of the Acoustical Society of America*, 64(4):1027–1035.

Liberman, M. (2009). *A New Golden Age of Phonetics*. Johns Hopkins University – Center for Language and Speech Processing, Baltimore, MD.

Liljencrants, J. and Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4):839–862.

Lindblom, B. and Maddieson, I. (1988). Phonetic universals in consonant systems. In Larry M. Hyman et al., editors, *Language, Speech, and Mind*, pages 62–78. Routledge, London.

Lindblom, B. and Sundberg, J. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, 50(4B):1166–1179.

Lindblom, B. (1983). Economy of speech gestures. In *The Production of Speech*, pages 217–245. Springer, New York.

Maddieson, I. (1995). Gestural economy. In *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, Sweden.

Ménard, L., Schwartz, J.-L., and Aubin, J. (2008a). Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, 50:14–28.

Ménard, L., Schwartz, J.-L., and Aubin, J. (2008b). Invariance and variability in the production of the height feature in French vowels. *Speech Communication*, 50(1):14–28.

Oushiro, L. (2019). Linguistic uniformity in the speech of Brazilian internal migrants in a dialect contact situation. In Sasha Calhoun, et al., editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 686–690, Melbourne, Australia. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Salesky, E., Chodroff, E., Pimentel, T., Wiesner, M., Cotterell, R., Black, A. W., and Eisner, J. (2020). A corpus for large-scale phonetic typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546, Online, July. Association for Computational Linguistics.

Schwartz, J.-L. and Ménard, L. (2019). Structured idiosyncrasies in vowel systems. *OSF Preprints*.

Schwartz, J.-L., Basirat, A., Ménard, L., and Sato, M. (2012). The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, 25(5):336–354.

Watt, D. J. L. (2000). Phonetic parallels between the close-mid vowels of Tyneside English: Are they internally or externally motivated? *Language Variation and Change*, 12(1):69–101.

Whalen, D. H. and Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics*, 23:349–366.

9. Language Resource References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*.

Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, Brighton, UK. IEEE.

Cohen Priva, Uriel and Strand, Emily and Yang, Shiying and Mizgerd, William and Creighton, Abigail and Bai, Justin and Mathew, Rebecca and Shao, Allison and Schuster, Jordan and Wiepert, Daniela. (2021). *The Cross-linguistic Phonological Frequencies (XPF) Corpus*.

Harper, M. (2011). The IARPA Babel multilingual speech database. Accessed: 2020-05-01.

Lee, J. L., Ashby, L. F., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., and Gorman, K. (2020). Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4223–4228.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech*, volume 2017, pages 498–502.

Mortensen, D. R., Dalmia, S., and Littell, P. (2018). Epitran: Precision G2P for many languages. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Paris, France, May. European Language Resources Association (ELRA).

Novak, J. R., Minematsu, N., and Hirose, K. (2016). Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6):907–938.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). MLS: A large-scale multilingual dataset for speech research. In *Proceedings of Interspeech*.

Salesky, E., Chodroff, E., Pimentel, T., Wiesner, M., Cotterell, R., Black, A. W., and Eisner, J. (2020). A corpus for large-scale phonetic typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4526–4546, Online, July. Association for Computational Linguistics.

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D. W., and Post, M. (2021). Multilingual TEDx corpus for speech recognition and translation. In *Proceedings of Interspeech*.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In *International Conference on Acoustics, Speech and Signal Processing*, pages 8126–8130. IEEE.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August. Association for Computational Linguistics.