Investigating the Corpus Phonetics Pipeline Applied to Diverse Speech Data

Emily Proch Ahn

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
Gina-Anne Levow, Chair
Eleanor Chodroff
Richard Wright

Program Authorized to Offer Degree:
Linguistics

University of Washington

ABSTRACT

Investigating the Corpus Phonetics Pipeline Applied to Diverse Speech Data

Emily Proch Ahn

Chair of the Supervisory Committee:
Gina-Anne Levow
Department of Linguistics

Corpus phonetics research has become increasingly large-scale as both data and automated tools have become more plentiful and available. Now that there are resources to study more kinds of data, what are some best practices in using these resources, especially when the data is diverse? This dissertation addresses the following research questions: How do we process diverse speech data, and how much can we rely on automated tools to conduct corpus phonetics research? The types of diversity covered in this work include multilingual and fieldwork corpora covering styles including read, spontaneous, and code-switched speech. Across four studies, we show that automated systems in the corpus phonetics pipeline are viable on multilingual and low-resource datasets. We first propose a pipeline that utilizes automated systems that convert orthography to phonemes, model the acoustics and align audio to those phonemes, and extract features for phonetic analysis. We apply this pipeline to a large, multilingual corpus and show both the utility and limitations of this derivative corpus in a careful study of outlying phonetic features. Then, we apply novel techniques to improve the phonetic forced alignment of low-resource field data, a challenging yet important process in language documentation. We encourage the research community to continue developing tools to aid in language documentation and cross-linguistic research. In doing so, it is important to include manual audits and to examine whether or not the tools are genuinely modeling the data.

ACKNOWLEDGMENTS

My first thanks go to those who advised me on this academic journey. I thank Gina-Anne Levow for her gentle guidance and wise perspectives in my field of research; she allowed me to explore a wide range of topics and also reel me in when I got stuck on narrow paths. My unofficial PhD advisor, Eleanor Chodroff, has inspired both broad directions and detailed content of my dissertation, and I thank her for her contagious enthusiasm and willingness to meet at odd hours between European time zones and Seattle. I thank Richard Wright for his high-level guidance of my work and for reminding me that research is fun. Additionally at UW, I thank Myriam Lapierre—for a meaningful collaboration on the Panãra forced aligner, the phonetics lab—for giving me feedback on every research project I have tackled, and Elizabeth Blankespoor—for graciously being my Graduate Student Representative during my exams. I also thank my masters and undergraduate advisors and mentors from my previous institutions: Yulia Tsvetkov, Alan Black, David Mortensen, Andrew Gordon, Sravana Reddy, Sohie Lee, and Angela Carpenter. They helped lay the foundations for me to be able to pursue interesting and important questions in the field of Computational Linguistics.

For my dissertation specifically, I thank several people who have contributed directly to this work. My undergraduate linguistics research team of Anna Batra, Sam Briggs, Emma (Miller) Rhoden, and Qianyue (Ivy) Guo helped not only with annotations but with methodological implementations and exploratory data analyses of Chapter 4. Figures 4.4 and 4.5 were created by Sam Briggs. I also had the privilege of working with Bridget Tyree and Hossep Dolatian for the work in Chapter 5. Bridget conducted preliminary annotations and modeling of the Panãra data, and Hossep inspired me to try the strategy of broadening phonetic categories. In addition, this beautiful LaTex template was provided by my mentor within the UW linguistics PhD program, Ajda Gokcen.

Lastly, my family and friends deserve my deep thanks and gratitude. First, my family has supported me every step of the way: Umma, Appa, Grandma, Grandpa, Halmoni, Haraboji, Mom,

Dad, and Ryan. My friends from Seattle, Pittsburgh, Boston, Bay Area, and beyond—they have also encouraged me to keep on this journey of following my dreams. I thank them all for reminding me of what is important in life. Finally, I thank my Creator who has most shaped who I am today.

Table of Contents

List of Figures

List of Tables

Part I

Foundation

CHAPTER 1
Introduction

1.1 Motivation

Corpus phonetics is the study of speech sounds within a large corpus, conducted with semiautomatic methods. Liberman (2019) stated that compared to standard phonetics research, corpus phonetics methods provide the benefits of increased researcher productivity (as automation speeds up manual labor), larger data coverage over a variety of speech patterns, and the ability to reproduce findings. Among his projections for what corpus phonetics may look like in the year 2030 was that there would be "robot phoneticians," if not phoneticians' "robot assistants." We support the idea of a robot that can assist the phonetician in conducting research, which is how we portray the use of automatic tools in this dissertation.

Large-scale, automated speech analysis can benefit the field in a variety of ways: a field linguist can develop a spoken corpus with accompanying transcriptions of an endangered language for use in natural language processing applications (Coto-Solano and Solórzano, 2017); a phonetician can compare the effectiveness of remote recording methods on retaining accurate acoustic measures (Oganyan et al., 2024); a sociolinguist can discover measurable vowel quality differences between groups of speakers in a large corpus (Reddy and Stanford, 2015; Coto-Solano et al., 2021); and, a typologist can test the degree to which analytic constraints may account for cross-linguistic patterns in phonetic realization (Salesky et al., 2020). However, as linguists are increasingly relying on automated tools for corpus phonetics research, one may ask how well these tools work to perform reliable analyses.

Furthermore, one may ask how well these tools work on diverse, non-standard speech data. Within a typical automated pipeline that involves grapheme-to-phoneme conversion, acoustic modeling, forced alignment, and acoustic-phonetic measurement, these automated tools may prove acceptable for some high-resource, well-studied language varieties such as Standard American En-

glish. However, when the data is more diverse (e.g., limited in a low-resource setting, or the language variety is less documented, or multiple varieties are used in a single utterance), it can be more challenging to produce reliable output for linguistics research. This dissertation explores how to utilize these automated tools in the corpus phonetics pipeline, especially in the context of diverse speech data.

1.2 Research questions and contributions

This dissertation aims to show that there is a large benefit to including automatic processes in the corpus phonetics pipeline. Using innovative methods and a careful, critical lens, these processes can aid the research of diverse speech data. We frame the larger themes of this work into two guiding research questions. We also present chapter-specific research questions, followed by the contributions of how we answered these questions.

RQ1 Methodology: What methods are viable for processing diverse speech corpora within the corpus phonetics pipeline?

(a) Chapter 3 (VoxCommunis): How do we process a large, multilingual read speech corpus to make it usable for downstream phonetics research?

We applied grapheme-to-phoneme conversion, forced alignment, and formant extraction in a novel pipeline to process read speech data from Common Voice (Ardila et al., 2020). We created a derivative corpus, VoxCommunis, for the purpose of downstream phonetic analysis. We used this corpus for a phonetic typology study that examined the uniformity of vowel realizations across 36 languages.

(b) Chapter 4 (Outliers): What does an audit of outliers from a corpus phonetics pipeline reveal about our assumptions on the quality of automation? What can this audit reveal about particular languages and data sources?

In auditing a set of outliers from two data sources across three languages, we observed that the violations of automated quality assumptions demonstrated quirks of dataset sources and revealed linguistic phenomena about Kazakh and Hausa.

(c) Chapter 5 (Panãra): What is a good level of representation of sound units when conducting language-specific and cross-language forced alignment in a low-resource, fieldwork data scenario? Do the strategies of broadening phone classes and cross-language acoustic modeling aid in the forced alignment of Panãra field data?

We tested methodologies of broadening phone categories and cross-language modeling of Panãra field data, and we recommend researchers to: (1) utilize a large, pretrained acoustic model such as the Global English MFA (McAuliffe and Sonderegger, 2023) model to align small field datasets; and (2) consider broadening phone categories to help alignment when training models from scratch.

(d) Chapter 6 (Code-switching): Is the inclusion of code-switched data viable when conducting forced alignment of fieldwork languages?

We demonstrated that the inclusion of code-switched speech and Russian pretrained models aided alignments of Urum field data.

RQ2 Impact: How much can one rely on automated processes to obtain interpretable phonetics answers?

(a) Chapter 4 (Outliers): What assumptions—about the quality of automated systems within a corpus phonetics pipeline—are violated? What are the types of "errors" that outlying vowel formants represent, how often do they occur, and why do they occur? Which types make the strongest impact?

We developed a novel outlier taxonomy for vowel formant output in phone-aligned speech, and found that errors in transcripts, alignments, and formant tracking can vary depending on the data source and the language. Our analysis also revealed that valid linguistic variation may be interpreted as erroneous.

(b) Chapter 6 (Code-switching): Does the method of acoustic modeling impact a downstream corpus phonetics investigation of code-switched speech from fieldwork data?

The method of acoustic modeling of Urum–Russian code-switched field data did impact a downstream corpus phonetics analysis. The best acoustic model yielded more

similar results to the manually fixed gold alignments than the worst acoustic model. However, without manual correction, none of the automatic alignments were suitable for reaching conclusions for this particular research question.

1.3 Outline

This dissertation is a compilation of several studies that analyze aspects of the corpus phonetics pipeline to enable the study of multilingual and low-resource language varieties. Chapter 2 discusses the background and prior literature on various aspects of language diversity and language technologies used in the corpus phonetics pipeline to process such speech corpora.

For the first half of our contribution, Chapters 3 and 4 investigate how to use large, multilingual corpora for phonetics research. In Chapter 3, we first introduce the corpus phonetics pipeline in detailing the methodology behind creating the VoxCommunis Corpus of 36 languages, a derivative of the Common Voice Corpus (Ardila et al., 2020). We then demonstrate how certain biases and assumptions throughout the pipeline impact downstream speech research in Chapter 4. We develop a taxonomy of outliers from the pipeline's output across three language datasets (Hausa, Kazakh, and Swedish) from VoxCommunis and the VoxClamantis Corpus (Salesky et al., 2020), and highlight weaknesses in the pipeline's assumptions while uncovering unique characteristics of these languages.

In the second half of our contribution, Chapters 5 and 6 test new methodologies in conducting acoustic modeling and forced alignment of low-resource field data. Instead of answering corpus phonetics questions about typology, we shift to applying corpus phonetics, specifically phonetic forced alignment, to language documentation. Chapter 5 investigates cross-language forced alignment and the manipulation of the grapheme-to-phoneme conversion in the acoustic modeling of Panãra, an Amazonian indigenous language. The final study in Chapter 6 tests the methodology of forced alignment for code-switched Urum–Russian field data. Findings from these studies not only benefit the direct research of languages such as Panãra and Urum, but more broadly demonstrate viable ways to conduct phonetics research of non-standard, diverse language data. Chapter 7 concludes this work with a summary and a discussion of broad themes from phonetics and language

technology.

CHAPTER 2
Background

Both language data and language technologies have become increasingly available in recent years. We first outline types of diverse speech data that are of interest in this dissertation, and then discuss the methodology behind automated systems within the corpus phonetics pipeline. For the reader's reference, we use the term "phone" broadly to mean an individual sound unit, whether it is a phoneme of a language—representative of one's mental representation of that sound—or an allophone, which relates to the acoustic-phonetic representation of that sound.

2.1 Diversity of speech data

Speech corpora from past decades generally came from monolingual, high-resource language varieties. Popular, well-transcribed corpora included the TIMIT Corpus (Garofolo et al., 1993) with read English sentences from monolingual US speakers, and the Buckeye Corpus (Pitt et al., 2005) covering spontaneous English conversations from speakers of Ohio, US. Now, we have access to datasets that encompass tens and hundreds of language varieties, which are commonly referred to as massively multilingual corpora. We also have online repositories of field data, in which field linguists have uploaded recordings of speakers of indigenous languages. In this dissertation, we focus on the aspect of speech diversity that covers languages or dialects that are not widely spoken, elicited in a range of styles from conversational speech to read sentences. We also study code-switching, where multilingual speakers mix their varieties interchangeably in a conversation. Diverse speech data can also include the fact that speakers come from wide backgrounds who vary in age, gender, race, ability, social status, etc, or that the dataset quality could be variable; these types of diversity are not the focus of this dissertation, but they are discussed during analysis.

Language technology that encompasses diverse speech data is still an active area of research. This section outlines three types of speech data diversity and discusses research that has been done to create such datasets and technologies.

Corpus	Reference	Languages	Speakers per lang	Hours per lang	Hours total	Hours transcribed	Speech style/type
IARPA BABEL	Harper (2011)	21	-	200	>4,000	>4,000	Conversational
GlobalPhone	Schultz et al. (2013)	20	100	400	17,000	17,000	Read sentences
CMU Wilderness	Black (2019)	699	<10	20	13,725	13,725	Read (Bible)
VoxClamantis	Salesky et al. (2020)	635	<10	20	12,700	12,700	Read (Bible)
VoxPopuli	Wang et al. (2021)	23	-	-	400,000	17,000	Parliament
mTEDx	Salesky et al. (2021)	8	-	15-189	>700	>700	Narrative
MLS	Pratap et al. (2020)	8	18-5,019	62-32,124	>36,000	>36,000	Read (audiobook)
FLEURS	Conneau et al. (2023)	102	3	12	1,400	1,400	Read (Wikipedia)
MMS (labeled)	Pratap et al. (2024)	1,107	<10	1-200	44,700	44,700	Read (Bible)
MMS (unlabeled)	Pratap et al. (2024)	3,809	-	<100	7,700	0	Misc
Common Voice (v20)	Ardila et al. (2020)	134	-	-	33,150*	33,150*	Read sentences

Table 2.1: An overview of multilingual speech corpora. For each corpus, we list the number of languages, speakers per language, hours per language, total hours, transcribed hours, and speech style or type of speech. The hyphen symbol (-) indicates that the relevant information was unavailable. *Of the 33,150 hours reported in Common Voice (v20; December 2024), 22,108 hours have been validated.

2.1.1 Massively multilingual speech corpora

Large speech corpora have been curated in recent years spanning a wide variety of languages and types of speech. Table 2.1 summarizes several prominent multilingual speech corpora described in this subsection.

From over a decade ago, the IARPA BABEL and GlobalPhone corpora were some of the first large, multilingual spoken corpora in research and have been popularly used for Automatic Speech Recognition (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.

In the last ten years, several multilingual spoken corpora have been released that are publicly available. The following corpora based on the Bible religious texts have some minor copyright restrictions. Most similar in approach to the corpus we create in Chapter 3 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 VoxCla-

mantis Corpus 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). The labeled dataset from the Massively Multilingual Speech (MMS) Corpus (Pratap et al., 2024) contains read speech from 1,107 languages of New Testament recordings derived from the same Bible corpus website as the CMU Wilderness corpus, though they reported higher quality alignment than the CMU Wilderness Corpus (Black, 2019). At the time of their releases, these Bible corpora were the largest spoken corpora 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.[1]

Other types of publicly available, read or transcribed speech corpora have also been developed. The Common Voice Corpus is a collection of read sentences voluntarily recorded and validated by internet users on web and phone platforms (Ardila et al., 2020). Version 7 (released in July 2021) of this corpus spans 75 languages and was used to create the derivative VoxCommunis Corpus in Chapter 3; the more recent Version 20 of Common Voice (released in December 2024) spans 134 languages. The multilingual Librispeech Corpus (MLS) encompasses read speech in the form of audiobooks, and it spans eight European languages and over 36,000 hours of speech (Pratap et al., 2020). The multilingual TEDx (mTEDx) Corpus contains over 700 hours of transcribed speech from TED talks in eight European languages (Salesky et al., 2021).

Lastly, FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech) has become a popular benchmark dataset for speech technology tasks (Conneau et al., 2023). It spans 102 languages with roughly 12 hours of speech per language. A set of English Wikipedia sentences had been translated into 101 other languages, and Conneau et al. (2023) recorded three native speakers per language to read these sentences. This dataset has been used for tasks such as speech recognition (Zhao et al., 2025), speech translation (Hu et al., 2025), and spoken language identification (Peri et al., 2025).

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

Aside from transcribed speech corpora, unlabeled datasets have also been created for use in training models for speech technology applications. The VoxPopuli Corpus contains 400,000 hours of unlabeled speech (and 17,000 hours of transcribed speech) data spanning 23 languages taken from European parliament recordings (Wang et al., 2021). The unlabeled set of MMS (Pratap et al., 2024) spans 3,809 languages and 7,700 hours of religious recordings from the Global Recordings Network.

Beyond speech corpora that may have transcriptions at the word-level, there also exist multilingual phonetic datasets with transcriptions at the phone-level (see for example the UCLA phonetics archive, Ladefoged and Maddieson, 1996; Li et al., 2021; Chodroff et al., 2024b; and IPAPack, Zhu et al., 2024). We do not utilize these types of datasets as starting points for our contribution studies, but such phonetic datasets can be curated by the corpus phonetics pipeline described in this work. If phone segments are automatically time-aligned or automatically transcribed, or both, one should exercise some caution in utilizing them as ground truth.

2.1.2 Low-resource, endangered language data

Among the approximate 7,000 languages spoken in the world, nearly 40% of them are believed to be dying and at risk of people no longer using the language (Campbell et al., 2022). Field linguists are actively documenting many of these languages for goals that can include language preservation, revitalization, teaching, and deepening linguistic understanding. Repositories such as the Endangered Languages Archive (ELAR)[2] allow linguists to upload speech recordings, dictionaries, and other language materials for the preservation and benefit of research over these languages. Research from Levow et al. (2021), for example, created systems for speaker diarization and recognition across eight language datasets from ELAR.

Other computational research has aggregated field corpora and developed technology to support documentation of low-resource languages. For many datasets, recordings may be plentiful while transcriptions are sparse, and this has been referred to as the "transcription bottleneck" for language research. Lane and Bird (2021) implemented the Sparse Transcription model from Bird

[2] https://www.elararchive.org/

(2021) to utilize local word discovery technology to predict phone transcription of content located between already-recognized or indexed words. Michaud et al. (2018) conducted phoneme recognition on Yongning Na speech and showed how linguists could benefit from using speech technology tools to aid their documentation.[3]

Besides developing phone recognition tools, research has also focused on alignment of text to audio at the utterance, word, and phone levels. Littell et al. (2022) created ReadAlong Studio, a toolbox for creating visualized audiobooks that have educational applications. They incorporated word-level alignment of speech to text in 30 indigenous languages, allowing learners to visualize each word they listen to in a karaoke-style display. Paschen et al. (2020) developed the DoReCo Corpus which included datasets from 44 fieldwork-based endangered languages, and they used the WebMAUS forced aligner (Kisler et al., 2017) to align all speech data at the phone-level. We utilize the DoReCo Corpus in our analysis of forced alignment of code-switched field data in Chapter 6.

2.1.3 Code-switched speech data

As many of the speakers in this world are multilingual, it is not uncommon to use more than one language to communicate. In this work, we will refer to the mixing of two or more languages in an utterance as code-switching.[4] Here is an example of a code-switched Spanish–English utterance with Spanish displayed in brackets:

> **[es su tío]** that has lived with him
>
> *"**[It's his uncle]** that has lived with him."*
>
> (From the Miami Bangor Corpus; Deuchar et al., 2014)

Much of the linguistic literature on code-switching has focused on the syntactic and sociopragmatic aspects of engaging multiple languages at once (Bullock and Toribio, 2009; Muysken, 2000).

[3] Michaud et al. (2018) also discussed perspectives on how these types of speech technology tools can help linguists psychologically and encourage them in their documentation process.

[4] While code-switching can refer to mixing languages or dialects within a whole conversation, we use it to mean switching languages within a single utterance or sentence. This finer-grained mixing is also called "code-mixing" in the literature.

With respect to the phonetics of code-switching, research has focused on how acoustic properties shift when speakers activate multiple languages in their mind. For example, Voice Onset Time (VOT) and speech rates noticeably changed near language switch boundaries between Spanish and English (Fricke et al., 2016). Using the example utterance above, that would mean that these phonetic and prosodic properties are altered when the speaker switches from the Spanish word *tío* (uncle) into the English word *that*. In another study, Seo and Olson (2024) recorded read sentences from Korean–English bilinguals to investigate vowel quality across different syntactic structures. They found that English vowels in code-switched Korean–English utterances were more Korean-like in intra-sentential rather than in inter-sentential code-switched structures. We similarly investigated vowel quality in Urum–Russian code-switched utterances for the case study in Chapter 6.

It has been observed that a language of broader communication, usually a high-resource language, is often used during the elicitation of a low-resource, target field language. In an overview of methods to bridge language documentation and speech processing technologies, Levow et al. (2017) proposed a language identification task between a high-resource language and a low-resource target language when both are present in field recordings. San et al. (2022) addressed the mixing of high- and low-resource languages by applying state-of-the-art language technologies to detect and transcribe English portions of speech in a dataset documented for the field language Muruwari. In this case, English was largely used in meta-linguistic commentary and questions, such as, *"What is the word for tree?"* This approach helped the annotation process, where authorized people could scan the meta-linguistic content and triage the recordings for later annotators who had more limited access to the corpora. These studies demonstrate that (1) language mixing is prevalent, and (2) applying technology to the higher-resource language can benefit the documentation process.

High-quality code-switched speech corpora are difficult to curate as they take extra time and resources to transcribe and annotate. Among some of these, the Miami Bangor Corpus is a set of spontaneous conversations of Spanish–English from American bilinguals living in Miami, USA (Deuchar et al., 2014). This dataset spans 35 hours and 84 speakers, with all words fully transcribed and tagged for language. Another corpus meticulously created is SpiCE, a large Cantonese–English

bilingual corpus recorded under three different elicitation styles in a controlled set-up (Johnson et al., 2020). Spanning 19 hours of speech and 34 speakers living in Vancouver, Canada, these recordings include sentence reading tasks, storyboard narration, and conversational interviews.

Developing technologies for code-switching is still a challenging area of research in the Natural Language Processing (NLP) and speech communities. Winata et al. (2023) found over 400 public research papers on code-switching from the ACL anthology and ISCA proceedings over the past few decades. These works focused on tasks ranging from language identification to sentiment analysis to automatic speech recognition (ASR). Among these papers, English mixed with a non-English variety such as Hindi, Chinese, and Spanish, was overrepresented. The authors highlighted a need for work to be done on more diverse non-English language pairs, for which our study in Chapter 6 fills a gap.

2.2 Corpus phonetics pipeline

We now give a general review of the technologies used in the corpus phonetics pipeline that we implement to process and analyze diverse speech data. The corpus phonetics pipeline can include the following processes: automatic speech recognition (ASR) or speech-to-text conversion, grapheme-to-phoneme conversion, acoustic modeling, forced alignment, and acoustic-phonetic feature extraction. Figure 2.1 displays the pipeline used throughout the contribution chapters of this work.

2.2.1 Automatic speech recognition (ASR)

Automatic speech recognition (ASR) is a system that takes in a speech audio file and outputs a transcription, typically at the word-level. This process can be done via separate processes of acoustic modeling followed by language modeling, or an end-to-end process where the ASR system is a single unit that inputs speech and outputs a string of text. Here, we discuss the former method and highlight acoustic modeling, as this is most relevant for the application of forced alignment used throughout this dissertation.

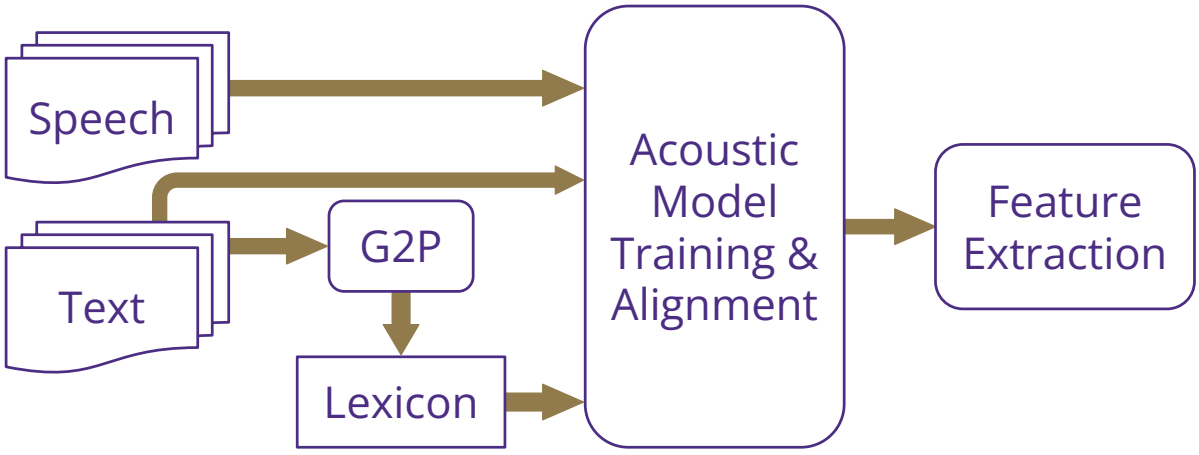

Figure 2.1: The corpus phonetics pipeline, assuming there already exists corresponding text transcriptions of the speech files. Without existing text transcriptions and utterance-level alignments, automatic speech recognition tools may be useful. G2P stands for grapheme-to-phoneme conversion, and feature extraction refers to the extraction of acoustic-phonetic features for the use of phonetic analysis.

The following process describes how the Kaldi ASR toolkit (Povey et al., 2011) in particular develops acoustic models. As we will see in Section 2.2.3, this is the underlying system used in the Montreal Forced Aligner.[5] Acoustic models are statistical representations of phonemes or phones. To train acoustic models for some speech data, Kaldi takes as input: audio files, lists of words and phones present in the transcripts, and speaker labels. It then extracts features such as Mel Frequency Cepstral Coefficients (MFCCs), which represent the human speech signal. Kaldi then trains monophone acoustic models by learning probability distributions for each phone state and their transitions between one another. A first-pass alignment is conducted, where the audio is segmented preliminarily into phones.[6] Then, Kaldi trains triphone models where not only is there a model learned per individual phone, but per each possible sequence of three phones. This allows for contextual representations of phones, as we know for example that /t/ in American English can sound different in the phonemic sequence /ɪ-t-ə/ (pronounced [ɪ-ɾ-ə] from "Italy") than in the phonemic sequence /s-t-ɹ/ (pronounced [s-t-ɹ] from "string"). Further training is combined with

[5]Details for how to use the Kaldi ASR toolkit and other corpus phonetics tools can be found in the tutorial by Chodroff (2018), also accessible at: https://www.eleanorchodroff.com/tutorial/.

[6]Segmentation is the process of identifying the start and end timestamps of each phone.

speaker normalization and adaptation. The models learn to attune to differences between phones instead of to differences between speakers or noise environments. This way, when new, "test" data is given to the system, the acoustic models would not have overfitted to only the seen data and speakers, but should generalize to the new data and speakers.

A traditional ASR system would complement acoustic modeling with language modeling, which learns the likelihoods of word sequences. The final output is a string of words that represents the highest probability transcription of the audio. In this dissertation, we do not use any full ASR system to create transcripts over speech recordings. However, it is a critical part of forced alignment, and automatic transcriptions have potential to be used within the corpus phonetics pipeline in the absence of manual transcriptions.

ASR has been utilized in linguistics research of diverse speech data. Coto-Solano et al. (2021) showed that ASR worked well to detect the sociophonetic phenomena of the US Southern and Northern Vowel Shifts, when used in conjunction with the DARLA tool (Reddy and Stanford, 2015) that also implemented forced alignment and vowel feature extraction. Popular large, pretrained ASR models such as wav2vec2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023) have been used for transcribing low-resource data as well. For example, Macaire et al. (2022) fine-tuned wav2vec2.0 on two French-based Creole languages, Gwadloupéyen and Morisien. Liu et al. (2024) explored different fine-tuning strategies of the Whisper model applied to seven different low-resource language datasets: Afrikaans, Belarusian, Icelandic, Kazakh, Marathi, Nepali, and Swahili. While research by Liu et al. (2024) was more focused on model methodologies, the best-performing strategies yielded Word Error Rates of below 20%, which is promising for future linguistics research to be conducted on these diverse languages.

Beyond standard ASR systems that produce transcriptions at the word-level, automatic phoneme recognition is another useful technology that takes speech as input and outputs a string of phonemes or phones. This technology has the potential for being more multilingual or language-agnostic, as sound systems may overlap better across languages than dictionaries of words. Michaud et al. (2018) noted that phoneme recognition systems need less training data than classic word-level ASR systems. Popular tools such as XLS-R FAIR (Xu et al., 2022) can be leveraged to conduct

zero-shot phoneme recognition on data from languages that the system had not seen before. It has the potential for usefulness in the language documentation pipeline (see Lane and Bird, 2021) and for verifying the accuracy of grapheme-to-phoneme systems (see Samir et al., 2025). However, we choose not to utilize phoneme recognition in this dissertation. It only learns surface-level acoustics and does not take into account linguistic (phonemic) contrasts. We choose instead to use linguist-curated, rule-based grapheme-to-phoneme systems. This more controlled setting is preferable when we will be conducting linguistic analyses on downstream output from the corpus phonetics pipeline.

2.2.2 Grapheme-to-phoneme (G2P) systems

Another 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, whether rule-based or statistical, and with or without audio files. The G2P systems discussed in this section generally use broad phonetic transcriptions that best correspond to sequences of phonemes or surface phonological segments. In this dissertation, we refer to these sound segments as phones.

Rule-based G2P systems

Rule-based G2P systems allow one to explicitly define the rules for mapping orthographic characters to phone segments. These are generally created by linguists and language experts, are easily controllable, and work best for languages with transparent orthography. Among popular rule-based systems, 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. Another tool that utilizes a web interface is the Cross-linguistic Phonological Frequencies (XPF) Corpus, a resource of phonemic lexicons or grammars for over 200 spoken languages (Cohen Priva et al., 2021).[7] The output of these systems is a pronunciation lexicon for each language dataset with mappings between each word form and its corresponding

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

phonemic transcription. Epitran and XPF are the main tools used in this dissertation. Additionally for application in language documentation contexts, there exists a tool named G_i2P_i (Pine et al., 2022), which is another rule-based G2P system with user-friendly features. While it already supports over 30 mainly indigenous languages, linguists can easily create mappings for new languages and implement the tool over their datasets.

Machine learning G2P systems

Other G2P tools and datasets have also been employed in research. The SIGMORPHON 2021 shared task on G2P utilized WikiPron, a pronunciation database derived from Wiktionary that spanned nearly 2 million pronunciations in 165 languages (Ashby et al., 2021; Lee et al., 2020). One baseline model for this task included Phonetisaurus (Novak et al., 2016), which paired an n-gram model to finite state machines. Other advanced models utilized neural network models such as a neural transducer (Makarov and Clematide, 2018) and bidirectional LSTMs (Hammond, 2021). Neural multilingual G2P systems have also been developed, such as the ByT5-based system by Zhu et al. (2022) that was trained on around 100 language datasets.

These commonly used G2P systems mentioned above were not evaluated based on the sounds actually produced in speech. It is understandably difficult to do so, due to variability in natural speech. More recently, there has been research that incorporates audio files in developing a G2P system. Gao et al. (2024) developed a G2P transducer that included speech units, where the model simultaneously processed audio and graphemes to produce the hypothesized phone sequence. This is a promising direction of research that is more similar in vein to phoneme recognition, and that we could see in potential applications of informing the documentation process of phone discovery.

G2P evaluation

For these models, accuracy was typically measured with Word Error Rate (WER) or Phone Error Rate (PER) based on a gold set of phone transcriptions (Ashby et al., 2021; Lee et al., 2020; Mortensen et al., 2018; Novak et al., 2016). Gold phone transcriptions, however, may be subjective or prone to errors and inconsistencies, as Gorman et al. (2019) revealed in an error analysis using data mined from Wiktionary. Furthermore, phoneme inventories for even the same language have

been reported to vary across reputable databases. Anderson et al. (2023) attributed this variation to both systematic and unpredictable differences in the transcription conventions and design choices for each database. In a careful audit performed by Samir et al. (2025), output from state-of-the-art phoneme recognizers was compared with "gold" phone sequence labels. For data partitions where annotators highly dispreferred the gold labels—when those partitions were removed, there was a noticeable improvement in the performance of a downstream phoneme recognition task. This shows that automatic labeling of phone sequences can lead to worse downstream performance of another system in the pipeline.

In Chapter 4, we create an outlier taxonomy that can attribute possible errors in downstream acoustic-phonetic output to the nature of the G2P system. Following that, in Chapter 5, we manipulate the level of granularity in defining a set of phones, to test the utility it has on alignment precision. While we do not define new metrics for measuring G2P accuracy, we show that the G2P system in a corpus phonetics pipeline has significant impact on both alignment and acoustic-phonetic measures.

2.2.3 Phonetic forced alignment

For phonetics research, it can be extremely useful to know where all phones of interest occur within a speech recording. While this can be achieved manually, having an automated process to facilitate this can greatly speed up annotation and enable analysis of substantially larger speech corpora (Labov et al., 2013). In this subsection, we discuss the various types of forced alignment systems and how they function, and the Montreal Forced Aligner (MFA) in particular, along with how these systems are evaluated and utilized on diverse speech datasets.

Types of alignment systems

Forced alignment tools take speech and its orthographic transcription and automatically produce time-aligned segmentation of the sound units. Popular forced alignment tools include the Montreal Forced Aligner (MFA, used in this work; McAuliffe et al., 2017), EasyAlign (Goldman, 2011), FAVE (Rosenfelder et al., 2011), LaBB-CAT (Fromont and Hay, 2012), Prosodylab-Aligner (Gorman et al.,

2011), and WebMAUS (Kisler et al., 2017). These aligners rely on acoustic models of a language's phone categories that have been trained on annotated speech data. Oftentimes they are built with GMM-HMMs, which apply the learned distributions of the acoustics (Gaussian Mixture Models) to the sequential output of hidden states (Hidden Markov Models). In this case, each output state is a time slice of acoustic features and each hidden state is a sound unit representation.

Beyond these popular tools, there has been recent research exploring neural network-based forced alignment. Rousso et al. (2024) compared state-of-the-art ASR systems, WhisperX and Massively Multilingual Speech Recognition (MMS), to MFA and found that MFA still outperformed these other models, when applied to English data. Kelley et al. (2024) used interpolation techniques and trained the acoustic model to allow for non-discrete, overlapping phone boundaries, providing increased precision performance over English speech data, compared to MFA. Zhu et al. (2024) developed a multilingual forced aligner that can segment unseen language data into IPA phones. While it would be interesting to utilize these latter two systems in our experiments, we leave this for future research.

Montreal Forced Aligner (MFA)

The Montreal Forced Aligner (MFA; McAuliffe et al., 2017) is the forced aligner used throughout this dissertation. Among other applications, MFA provides a user-friendly wrapper to the Kaldi ASR toolkit (Povey et al., 2011) for acoustic model training and alignment. We chose MFA over other forced aligners because it is user-friendly, it is built on a high-quality ASR system, and it performs consistently well against other aligners (Mahr et al., 2021). The tool is also easy to use for training or adapting on small datasets.

For MFA, three types of input are required: speech files, text transcripts, and a lexicon, or pronunciation dictionary. A lexicon contains entries of each orthographic word in the data mapped to its sequence of sounds, and these sounds are considered the dataset's phone inventory. If the forced aligner is trained from scratch, it will generate an acoustic model. For every possible sound that is defined in the lexicon, the acoustic model will learn the distributions of the acoustic features (typically MFCCs). If for example [a] and [i] are separate sounds in the phone inventory, the model

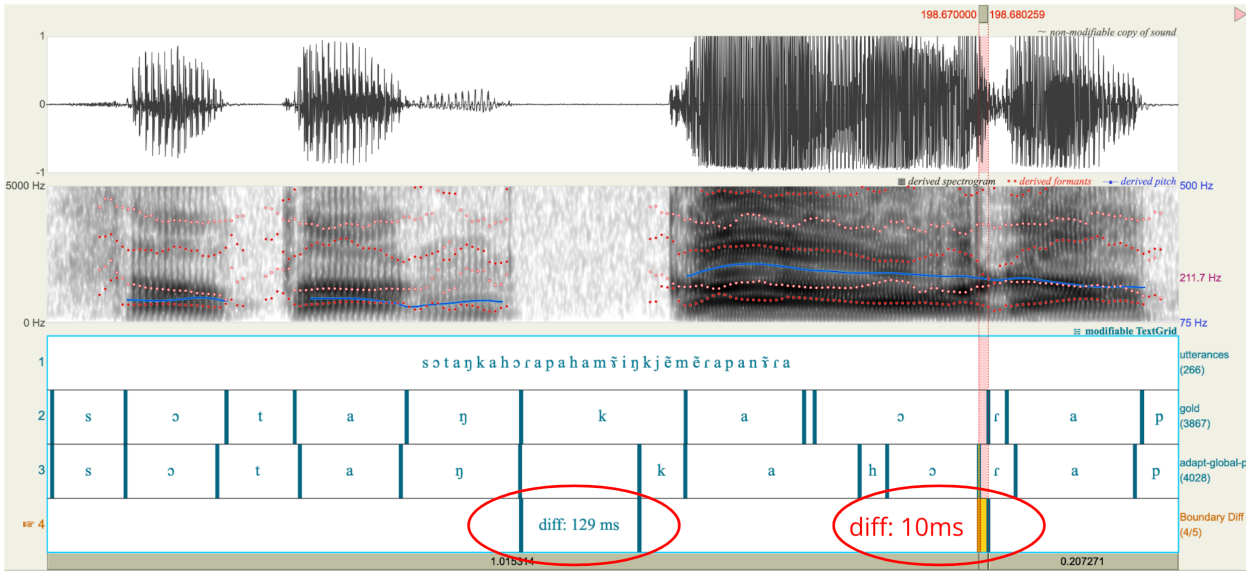

Figure 2.2: Example of boundary difference calculations in a Panãra audio file from the dataset in Chapter 5, visualized in Praat (Boersma and Weenink, 2022). The first interval tier displays the input utterance. The second tier shows the manually-aligned boundaries, and the third tier shows system-aligned boundaries. The fourth tier shows the onset boundary differences for the phones [k] and [ɾ].

will create a separate probability distribution for [a] and [i] features. In order to produce time-aligned segments of the sounds, a model is then applied to the target speech with its corresponding transcript.[8] If a user of the toolkit wants to align some data of a language for which an existing pretrained acoustic model exists, they can simply download that model and apply it to their data. When applying a pretrained model to new data, one can either apply the model directly, or one can "adapt" the model to the new data. The adaptation process requires the same three types of input as training a model—namely, speech files, text transcripts, and a lexicon. Adaptation updates the means of the pretrained acoustic models (but not the variances) based on the acoustics of the new data; this is the technique used in several experiments throughout this dissertation. Default MFA settings, employing a triphone GMM-HMM architecture with MFCCs, were used for all model training and adaptation.

[8]This alignment process is not a supervised learning task, because phone alignments are not used to train the models. The alignments are estimated based on the acoustic model.

Alignment evaluation

We utilized two main metrics for the evaluation of alignment performance in Chapters 5 and 6. Each metric requires comparing automated system alignments to the hand-corrected, "gold" alignments. **Precision** is the percent for which the model onset boundary is within a 20 ms tolerance (the selected threshold) of the manually aligned onset boundary (MacKenzie and Turton, 2020; McAuliffe et al., 2017).[9] It can also be referred to as "agreement" between a human annotator and system output. **Accuracy** is the proportion of model-aligned intervals whose midpoints lay within the respective gold intervals (Knowles et al., 2018; Mahr et al., 2021). Figure 2.2 shows an example of (1) an onset boundary that is displaced greater than 20 ms and therefore considered imprecise: the [k] is displaced 129 ms after its corresponding gold boundary, and (2) an onset boundary that is within 20 ms and is considered precise: the highlighted [ɾ] is displaced only 10 ms before its corresponding gold boundary.

Additional metrics used in prior research have included the absolute difference between system and gold offsets or segment midpoints (Coto-Solano and Solórzano, 2017). Overlap rate was also used by Gonzalez et al. (2020), who calculated a percentage based on the gold segment duration, system segment duration, and their shared duration.

Forced alignment applied to diverse data

Research has explored a range of strategies to force align non-standard and low-resource language varieties. Cross-language forced alignment is a technique that uses a pretrained acoustic model of a high-resource language to align a low-resource language. American English models have been helpful for aligning related English varieties such as British English (MacKenzie and Turton, 2020) and North Australian Kriol (Jones et al., 2019), as well as less related languages such as Cook Islands Maori (Coto-Solano et al., 2018) and Yoloxóchitl Mixtec (DiCanio et al., 2013). Meer (2020) used American English models to align Trinidadian English, representing one of the nativized, post-colonial "New Englishes" spoken worldwide. Bribri, a Chibchan language of Costa Rica, has also

[9] A tolerance of 20 ms has been commonly used as an optimal threshold for inter-rater agreement and for evaluation of forced aligners. See Williams et al. (2024) for rationale and further references.

been phone-aligned with English-based FAVE and French-based EasyAlign pretrained models (Coto-Solano and Solórzano, 2017). We explore cross-language modeling and alignment techniques in Chapters 5 and 6.

Besides cross-language or cross-variety modeling, forced alignment has been applied to the study of other diverse language settings. Mathad et al. (2021) investigated the goodness of pronunciation in a speech pathology setting through forced aligned speech of children with and without a cleft palate. Williams et al. (2024) evaluated forced alignment of MFA on L2 English, producing findings such as that speaker rate (correlated more with language proficiency) may have more effects on alignment performance than L1 language variety. Bailey (2016) used FAVE and dictionary expansion techniques to study sociophonetic phenomena of British English. These studies that span speech pathology, education, and sociology show that forced alignment plays a critical role in the pipeline of broader speech language research.

2.2.4 Acoustic-phonetic measurements

Last in the pipeline of corpus phonetics, once audio files have been segmented, is to extract the acoustic-phonetic measurements. Researchers may be interested in measures such as vowel formants, voice quality, pitch, voice onset time (VOT), spectral features, and so forth. In this dissertation, we focus on vowel formants for two reasons. First, vowel segment boundaries tend to be more accurately placed during the automatic forced alignment process. Second, vowel formants are standard measurements reported across all language variety illustrations of the Journal of the International Phonetic Association (JIPA).

Vowel formants

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. F3 frequently correlates with lip rounding, rhoticity, and nasality, and F4 can reflect high front vowel contrasts and aspects of voice quality (Eek and Meister, 1994; House and Stevens, 1956; Ladefoged et al., 1978; Lindblom and Sundberg, 1971).

A popular method for extracting formants is the Burg linear predictive coding (LPC) algorithm that relies on parameters including the maximum number of formants to search for and the maximum frequency under which to do the search (Barreda, 2021; Boersma and Weenink, 2019; Reddy and Stanford, 2015; Rosenfelder et al., 2011). To evaluate the accuracy of extracted formants, one can use the mean absolute differences between gold standard formants and automatic formant estimates (Barreda, 2021; Evanini et al., 2009). While we do not explicitly report formant measurement performance against a gold standard in this work, our taxonomy of outliers in Chapter 4 includes formant extraction errors, where annotators could manually specify that the extraction was erroneous.

It is from this last piece of the corpus phonetics pipeline that phoneticians and field linguists can discover patterns of their language varieties of study. As the data increases in size and diversity, it is important to consider the impact of the pipeline on these acoustic-phonetic measurements.

2.3 Looking ahead

We have introduced several types of diverse speech data and given background on the automated systems within the corpus phonetics pipeline. In the next part of this dissertation, we will fill gaps in the literature and extend work done in these areas. At a high level, our contributions lay foundations for conducting corpus phonetics on multilingual and low-resource speech data.

Chapter 3 takes the Common Voice Corpus (Ardila et al., 2020) through the corpus phonetics pipeline to create a derivative corpus (VoxCommunis) with 36 languages, usable by linguists to conduct phonetics research. This new corpus complements the existing massively multilingual speech corpora detailed in Section 2.1.1. Chapter 4 then analyzes the quality of VoxCommunis and VoxClamantis (Salesky et al., 2020), conducting a semi-automatic audit of outlying vowel formants in order to identify various points in the pipeline where errors occurred. We provide a

novel taxonomy and holistic framework for interpreting outliers from a multi-step process pipeline. Chapters 5 and 6 explore methods for conducting phonetic forced alignment on low-resource, field corpora. Research on cross-language forced alignment for low-resource data has been done for language varieties in the past, but our work experiments with new techniques. First, we change the level of granularity of phone classes in acoustic modeling (Chapter 5). This technique has been done for sentence-level alignment of larger, high-resource language datasets (Hoffmann and Pfister, 2013), and we extend that work by conducting phone-level alignment of Panãra field data. Second, we include code-switched utterances in the training of acoustic models (Chapter 6). Code-switching has not yet been analyzed in a computational corpus phonetics framework on fieldwork datasets, and we fill this gap by testing strategies for phone-aligning Urum–Russian field data. Each contribution chapter of this dissertation critically examines the experimental results and provides case studies or error analyses to discern linguistic patterns in the output and to better understand the effects of corpus phonetics technology on language data.

Part II

Contribution

CHAPTER 3

The Corpus Phonetics Pipeline Applied to the Common Voice Dataset

Overview

We answer RQ1 and introduce the methods of a corpus phonetics pipeline and how it can be applied to multilingual read speech. Data resulting from this processed corpus is used in Chapter 4. This chapter is largely equivalent to the paper titled *VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis* from *The Proceedings of the 13th Language Resources and Evaluation Conference* (Ahn and Chodroff, 2022).

3.1 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/.

3.2 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. 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 chapter, 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, community-driven, multilingual speech corpus that contains read sentences collected via the internet on both web and phone platforms. Sentences

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

are gathered from Wikipedia articles as well as from internet users who identify themselves with the language community. These sentences are then recorded individually by language community members. A subset of utterances is additionally "validated" by members, indicating that at least two members have confirmed that the reading of a specific utterance is faithful to the corresponding written text. 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. 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 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.

3.3 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.

3.3.1 Grapheme-to-phoneme conversion

We relied on two linguist-designed, rule-based Grapheme-to-Phoneme (G2P) systems for this conversion: Epitran (Mortensen et al., 2018) and the XPF Corpus (Cohen Priva et al., 2021). These systems have been developed for many languages with a "transparent" orthography, in which the orthographic representation is systematically related to its phonemic form.

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 poor G2P quality or processing issues in the acoustic model development. These were two Chinese language datasets, Persian, and Arabic.[2] 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.[3] We manually updated several entries across most datasets, especially where foreign characters or loanwords were included in the input.

3.3.2 Acoustic model training

The acoustic models were developed using the generated pronunciation lexicons and the Montreal Forced Aligner (MFA; 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 of 1,000 Gaussians. These were followed by 35 iterations of Linear Discriminant Analysis with Maximum Likelihood Linear Transform (LDA+MLLT), which reduced the feature space and derived speaker-specific transformations, and two rounds of

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

[3]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.

speaker-adapted training (SAT). This training was conducted with feature space Maximum Likelihood Linear Regression (fMLLR), which removed speaker-specific information from the models, each with a maximum of 2,500 leaves and 15,000 Gaussians.[4] 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.[5] Pre-processing 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.

3.3.3 Formant extraction

The first four formant values were extracted from each aligned vowel quartile using the Linear Predictive Coding (Burg method) algorithm implemented in Praat via Praat scripting (Boersma and Weenink, 2019). For each formant, values at 10 ms before and after the midpoint were also extracted, for increased stability in analysis (see Section 3.5.1). All formant values were extracted under both "high" and "low" frequency settings. Specifically, the tracker searched for five formants with a ceiling of 5,500 Hz in the "high" setting, and a ceiling of 5,000 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 1,200 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

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

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

formants over all speakers and utterances, but employ this heuristic for the case study data in Section 3.5.

3.4 Data

Table 3.1 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.[6] Table 3.1 further shows language-related descriptions of each dataset. Vowel and consonant inventory sizes were determined 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 3.1) and one from Indonesian (Figure 3.2) 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.

3.5 Case study

The data in VoxCommunis can be a testbed for many research questions that concern phonetic and phonological theory.[7] 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 (Chodroff and Wilson, 2017; Fruehwald, 2017; Guy and

[6] Ardila et al. (2020) referred to a validated utterance as one that has a majority of up-votes from crowdsourced listeners who verified that the text transcription matched the audio.

[7] This case study was largely inspired and written by co-author Eleanor Chodroff.

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 3.1: 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.*

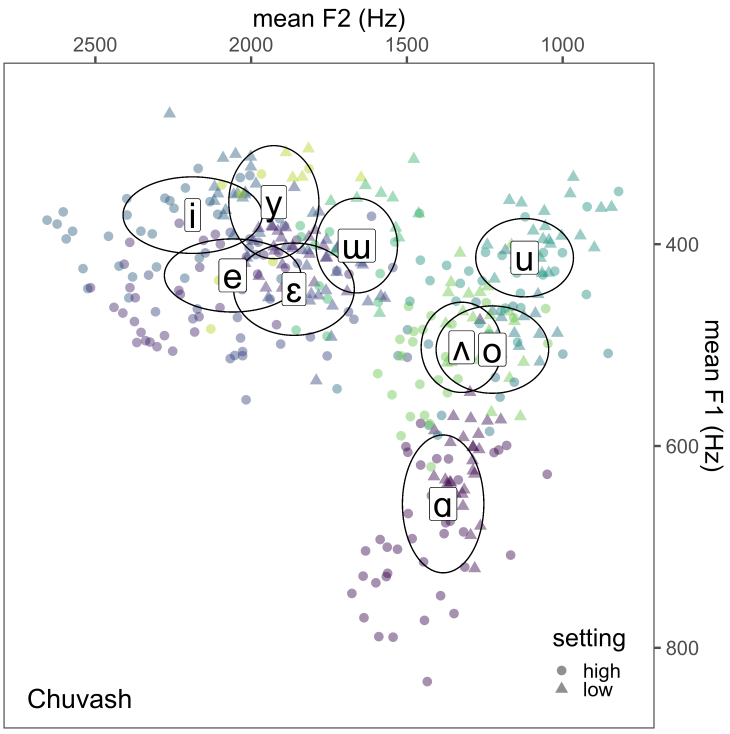

Figure 3.1: Chuvash vowels in F1×F2 space (Hz). Each point corresponds to a speaker-specific pair of means and is coded for the formant extraction setting. Labels are centered on the grand means for each category, and ellipses correspond to ± one standard deviation from the mean across speakers.

Hinskens, 2016; Maddieson, 1995; Ménard et al., 2008; Keating, 2003). 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 realization of a distinctive feature Ménard et al. (2008); Schwartz et al. (2012).

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.

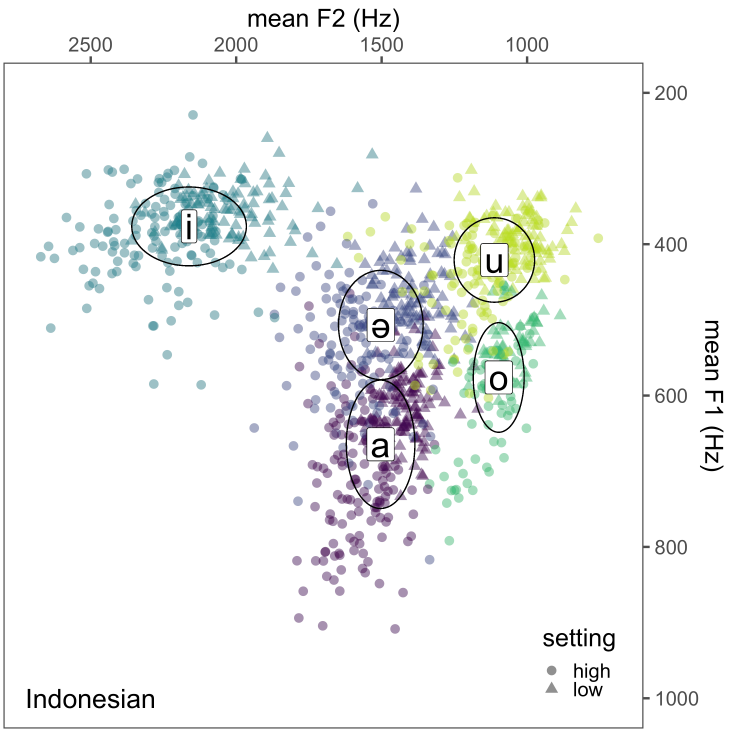

Figure 3.2: Indonesian vowels in F1×F2 space (Hz). Each point corresponds to a speaker-specific pair of means and is coded for the formant extraction setting. Labels are centered on the grand means for each category, and ellipses correspond to ± one standard deviation from the mean across speakers.

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.

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., 2008, Portuguese: Oushiro, 2019, six unique languages: Schwartz and Ménard, 2019). Similar to the present study, Salesky et al. (2020) also

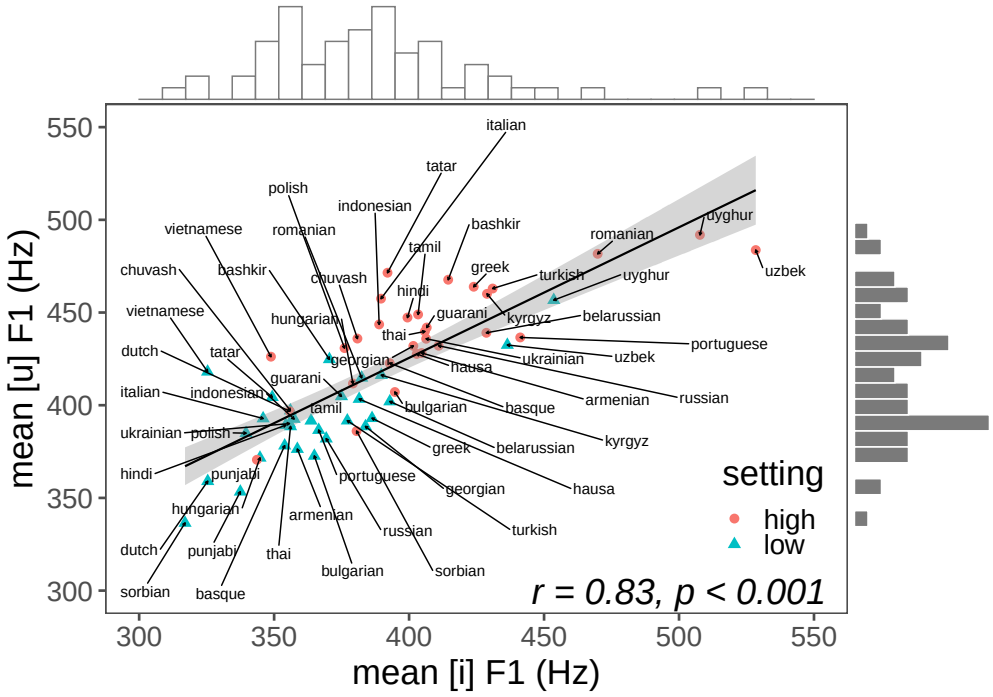

Figure 3.3: Correlation of mean F1 between [i] and [u] across 30 languages with the best-fit regression line. Each point represents a setting- and language-specific pair of F1 means. For example, the blue triangle to the bottom left plots the mean F1 for [u] against the mean F1 for [i] among all low-setting speakers for (Upper) Sorbian. The overall Pearson correlation value is $r = 0.83$ and the significance value is $p < 0.001$. Bar graphics above and to the right of the chart are histograms that demonstrate the frequency at different hertz values for [i] and [u].

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 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.

3.5.1 Methods

For the analysis, we employed the set of formants described in Section 3.3.3 in which either the high or low formant extraction setting was used for each speaker. Focusing on F1 and F2, we used

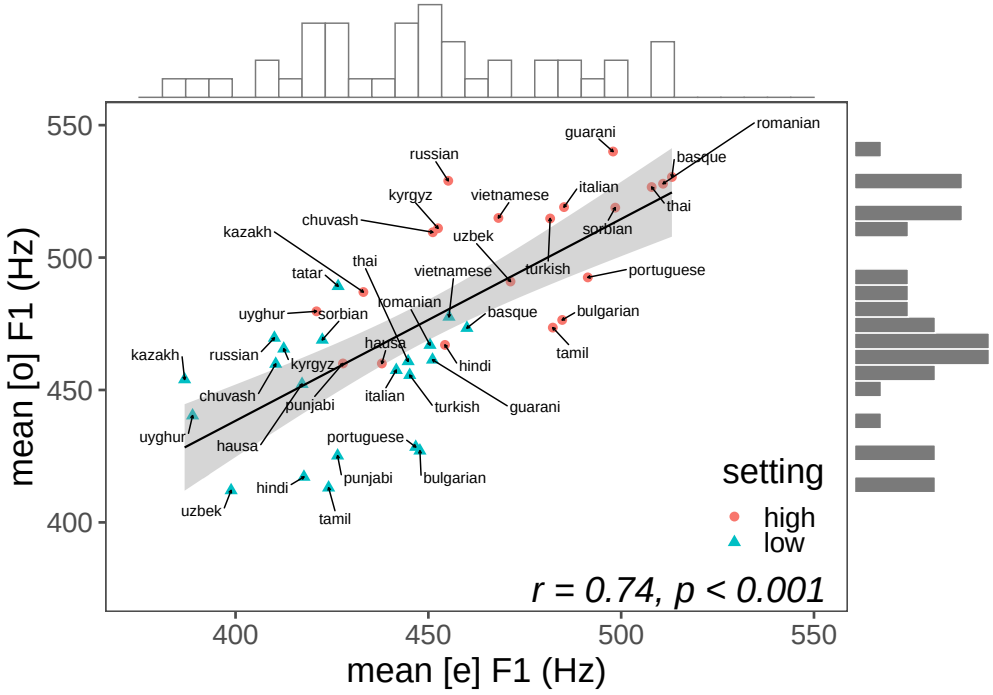

Figure 3.4: Correlation of mean F1 between [e] and [o] across 22 languages with the best-fit regression line. Each point represents a setting- and language-specific pair of F1 means. For example, the red dot to the top right plots the mean F1 for [o] against the mean F1 for [e] among all high-setting speakers for Basque. The overall Pearson correlation value is $r = 0.74$ and the significance value is $p < 0.001$. Bar graphics above and to the right of the chart are histograms that demonstrate the frequency at different hertz values for [e] and [o].

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.

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 3.2: 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 3.3: 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.

3.5.2 Results and discussion

For each of the F1 and F2 analyses, six pairwise correlations reached significance, as shown in Tables 3.2 and 3.3. 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 3.3 and 3.4). 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., 2008; Oushiro,

2019; Schwartz and Ménard, 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 feature. These assumptions warrant additional research.

3.6 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.

Postlude

Since the release of this VoxCommunis dataset, there have been newer releases of the VoxCommunis and Common Voice datasets. What is currently available as of April 2025 is as follows: Common Voice is up to version 21 and covers 134 language varieties.[8] VoxCommunis spans 56 languages with updated models around version 17 of the Common Voice releases.

Several studies have utilized this data for relevant research. Zhu et al. (2024) used a portion of VoxCommunis, combined with other data, to develop a multilingual keyword spotting system and a multilingual forced aligner. Jones and Renwick (2024) trained a speech technology tool using 11 hours of Italian data from VoxCommunis in order to detect marginal contrasts in Italian vowels for sociophonetic research. Pricop (2024) studied consonantal f0 effects in Catalán using over 3,000 hours from the newer release of VoxCommunis that covered Common Voice versions 16 and 17. Moran et al. (2024) sampled ten diverse language datasets from VoxCommunis to study how the variation of vocal tracts impacts phonetic diversity. It is encouraging to see the utility of this corpus on a variety of speech research ranging from linguistics to computer science and biology.

We have answered RQ1 in providing viable methods for processing multilingual data in a semi-automated pipeline. However, after the creation of this initial corpus, one might question the

[8]Not all of these 134 varieties have their transcriptions validated.

quality of extracted alignments and measurements, as none of the output was manually corrected. For instance, a researcher had contacted us and shared that the Armenian dataset from Common Voice was from speakers of two distinct dialects, Eastern and Western Armenian. Dialect metadata was not included in the release. This issue, in addition to another issue where the G2P system did not include schwa epenthesis, made the Armenian VoxCommunis data unusable. The following chapter will provide methods to analyze the quality of VoxCommunis and other similarly derived corpora.

CHAPTER 4
Analyzing Vowel Formant Outliers from a Corpus Phonetics Pipeline

Overview

While the previous chapter addressed RQ1 in how to apply a corpus phonetics pipeline to process multilingual data, the question from RQ2 on the quality of this processed data becomes more relevant. This chapter investigates the quality of the VoxCommunis Corpus for three language datasets. We aim to understand the types of outliers that the corpus phonetics pipeline can produce. We will observe that the method of G2P has a strong impact on how much vowel formant measurements vary. This chapter is largely equivalent to the conference proceedings of *Interspeech* titled *An Outlier Analysis of Vowel Formants from a Corpus Phonetics Pipeline* (Ahn et al., 2023).

4.1 Abstract

Automated systems in a corpus phonetics pipeline may involve grapheme-to-phoneme conversion, forced alignment, and acoustic-phonetic measurement, and each of these stages requires a strong assumption regarding the output quality. We investigated these assumptions by auditing outliers in vowel formants from two multilingual read speech corpora, CMU Wilderness and Mozilla Common Voice, across three languages: Hausa, Kazakh, and Swedish. From this audit, we developed a novel outlier taxonomy that included broad outlier categories of transcript errors, alignment errors, formant tracking errors, linguistic variations, and fine samples. We showed the utility of this outlier analysis in identifying weaknesses in corpus-specific and corpus-general pipeline assumptions, and discovering characteristics of particular languages. From a careful analysis of outliers, we can discern how to better account for the true variation in a language, while reducing the true noise from the system and data.

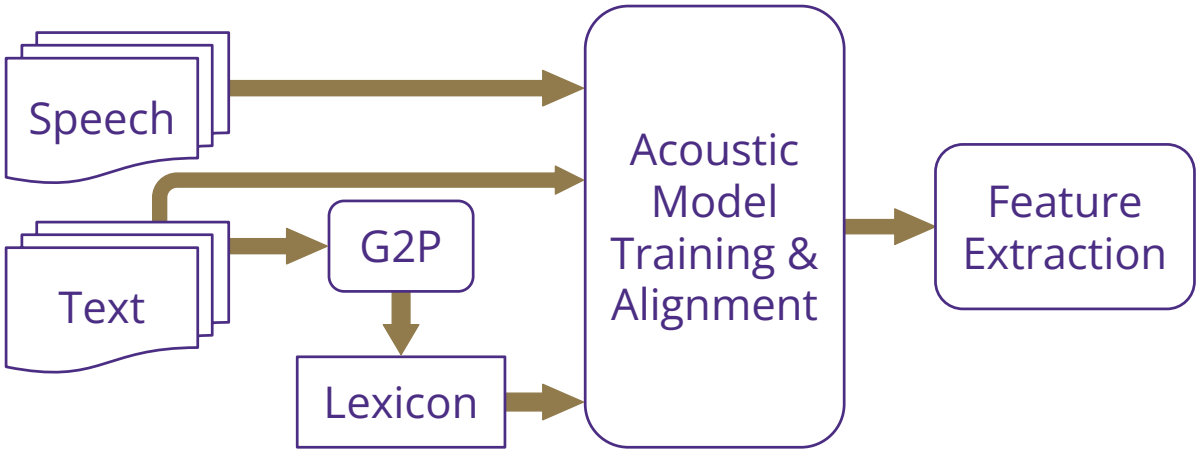

Figure 4.1: An automated pipeline that we analyze in this work. It takes speech and its corresponding text through a grapheme-to-phoneme (G2P) system to produce a lexicon of phone sequences. Then these are used to develop acoustic models and produce time-aligned phone segments. Finally, vowel formants are extracted for analysis.

4.2 Introduction

The growing availability of multilingual speech corpora and speech processing tools has enabled large-scale cross-talker and cross-linguistic investigations of acoustic-phonetic variation. Acoustic-phonetic analysis has typically depended on a data processing pipeline that involves the collection of speech recordings, a time alignment of the relevant units of analysis to the speech signal, and acoustic-phonetic measurements of the relevant units from the speech recording (Figure 4.1). Researchers commonly implement this pipeline manually, with utmost consistency and minimal bias; however, researchers can benefit in terms of time, consistency, replicability, and scalability by automating aspects of this pipeline.

Nevertheless, automation requires that the researcher commit to assumptions that may not always be met. In this chapter, we investigate the degree to which underlying assumptions of automation may be violated in two multilingual read speech corpora through an error analysis of automatically extracted vowel formants. We ultimately suggest that a partial manual audit should always be implemented, but the presented patterns provide some insight to future researchers about likely problematic locations in the overall pipeline.

In processing a large, read speech corpus, we have identified a series of steps that are frequently automated. At each of these steps, the researcher makes certain assumptions about the data and input at hand. If any assumption of a given step is violated, it will have downstream effects on the resulting segmentation and measurement quality.

First, with read speech data, it is frequently assumed that **the script is the transcript.** Though the participant may have intended to read the script faithfully, a script will not contain speech errors or disfluencies that may have occurred.

Second, in converting the words of the script or transcript to a phonetic transcription, it is frequently assumed that **the canonical phonetic transcription is an accurate phonetic transcription**. Grapheme-to-phoneme (G2P) systems and pronunciation dictionaries convert individual words into sequences of sound units that are estimated from the orthography. These can be rule-based, linguist-curated systems such as Epitran (Mortensen et al., 2018) or XPF (Cohen Priva et al., 2021), or ones that involve neural network models (Hammond, 2021; Makarov and Clematide, 2018). These models measure accuracy with Word or Phone Error Rate compared to a gold phone transcription (Ashby et al., 2021; Lee et al., 2020; Mortensen et al., 2018; Novak et al., 2016). In some cases, multiple phonetic transcriptions can be provided for a given word, but the set of transcriptions is nevertheless constrained.

Third in this pipeline is to conduct phonetic forced alignment, which time-aligns the output phone sequence to the audio. Forced aligners rely on acoustic models, which learn distributions of discrete sound units. The forced alignments can be evaluated with segment boundary time displacement (Gonzalez et al., 2020; McAuliffe et al., 2017), a binary accuracy overlap score (Mahr et al., 2021), or even overlap rate (Gonzalez et al., 2020); however, these scores rely on existing gold segmentations, which many researchers may not have for their data. It is therefore frequently assumed that **the segmentation is viable and accurate** given the phonetic transcription and acoustic model.

Last in this pipeline is to extract acoustic-phonetic measurements. Regardless of technique, the measurement frequently relies on an assumption of certain parameters. In this chapter, we focus on formant extraction, which measures the spectral frequency of high energy concentrations, which

typically reflect resonances of the vocal tract. The first two vowel formants are especially representative of vowel quality features such as height and backness (Ladefoged and Johnson, 2014). Errors can occur when the tracker incorrectly identifies a formant as an adjacent formant or harmonic, or the tracker estimates the spectral prominence incorrectly (Chen et al., 2019). Evaluation of formant extraction typically uses mean absolute differences between the gold standard and automatic formant estimates (Barreda, 2021; Evanini et al., 2009). These evaluation methods once again rely on gold data, which many researchers may not have. With the exception of any coarse outlier exclusion protocol (e.g., removing all tokens beyond some threshold), it is otherwise frequently assumed that **the parameters are accurately specified and the acoustic-phonetic measurement is viable and accurate.**

At the end of this pipeline, the researcher ideally wants an acoustic-phonetic measurement that represents the targeted speech. While averages of the data may cut through any noise generated by the assumptions, the present work specifically focuses on the outlying data: these are by definition non-representative tokens of the targeted speech. In a manner similar to an error analysis, this chapter addresses how we can categorize and understand the outliers in vowel formants to gain deeper insight into our assumptions of quality from an automated pipeline that includes grapheme-to-phoneme (G2P) mapping, forced alignment, and vowel formant extraction. In other words, what are the types of "errors" that outlying vowel formants represent, how often do they occur, and why do they occur? This analysis ultimately reveals characteristics of particular languages and data sources through their violations of different assumptions in the pipeline; these are then investigated in a set of case studies.

4.3 Data

The present outlier analysis focuses on subsets of two massively multilingual corpora: the CMU Wilderness Corpus (Black, 2019) and its derivative VoxClamantis Corpus (Salesky et al., 2020), and the Mozilla Common Voice Corpus (Ardila et al., 2020) and its derivative VoxCommunis Corpus (Ahn and Chodroff, 2022; see also Chapter 3). The CMU Wilderness Corpus contains audio recordings of the New Testament in nearly 700 languages; each language has around 20 hours of

data that come from a few speakers, mostly male. The Common Voice data has over 100 languages represented with spoken utterances collected and validated by internet users. These corpora were selected because they are stylistically similar (consisting of read speech), cover a broad range of languages, and provide phonetic alignments and vowel formants.

We chose three languages from both corpora for our analysis. **Hausa (ISO639-3:hau)** is a Chadic language from the Afro-Asiatic language family, spoken around the border between Niger and Nigeria (Wolff, 2023). It used to be transcribed in Ajami, an Arabic alphabet, until the early 1900s when the Latin alphabet became the standardized script. As a tonal language, Hausa has between six and ten vowels (Moran and McCloy, 2019). Vowel length is phonemic, which is not distinguishable from the Latin script. **Kazakh (ISO639-3:kaz)** is a Turkic language spoken primarily in Kazakhstan, as well as parts of China, Uzbekistan, Mongolia, and Afghanistan (Bri, 2023a). Though it uses mainly the Cyrilic script in modern days, it used the Arabic script prior to the 1900s. Linguists do not agree on the number of vowels in its phonemic inventory, as there have been reports ranging from five to eleven (McCollum and Chen, 2021).[1] **Swedish (ISO639-3:swe)** is a Germanic language spoken mainly in Sweden, and it uses the Latin alphabet (Bri, 2023b). There are nine orthographic and phonemic vowels, and each vowel has two corresponding allophones: a long vowel and a short vowel (Riad, 2013).

4.4 Methodology

4.4.1 Data preparation

We downloaded the language-specific data from the Wilderness corpus[2] (sampled at 16kHz and distributed as MP3 files) and from Common Voice[3] 8.0 (as 32kHz MP3 files), and converted these to mono-channel 16kHz waveforms. The conversion to WAV was to satisfy some system assumptions;

[1] As shown in Table 4.1, the two Kazakh datasets used in our work follow two separate inventories: one with six vowels and the other with eleven.

[2] Each reading was individually downloaded from https://www.faithcomesbyhearing.com/audio-bible-resources/mp3-downloads.

[3] Only the validated utterances were downloaded.

the files were lossy from the original MP3 format. The six datasets spanning these two corpora and three languages were all processed by the Epitran G2P toolkit (Mortensen et al., 2018). While the forced alignment for the VoxClamantis (i.e., Wilderness derivative) corpus had used a multilingual ASR model (Wiesner et al., 2019) trained with Kaldi (Povey et al., 2011), we trained new acoustic models and generated alignments on the Common Voice data with the Montreal Forced Aligner (McAuliffe et al., 2017), which also utilizes Kaldi.

The vowel formants were extracted with Praat (Boersma and Weenink, 2019) using the Linear Predictive Coding (Burg method) algorithm. As the Wilderness data was impressionistically dominated by male speakers, the data was processed with a five-formant ceiling of 5,000 Hz, as recommended for the male vocal tract. The Common Voice data was processed with a five-formant ceiling at both 5,000 Hz and 5,500 Hz (recommended for the female vocal tract). As Common Voice had a greater mixture of male and female speakers, a clustering process was applied to classify each speaker as having a high or low formant range (see Section 3.3.3). We used speech from only the low-setting speakers for a better comparison to the Wilderness data. Since the size of the Swedish Common Voice dataset was much larger than the other two Common Voice language datasets (see *Available Corpus* in Table 4.1), we down-sampled it by randomly selecting 1,000 utterances as the starting point for discovering outliers. Formant values were taken at the midpoint from the Wilderness data, and as the mean of values at three timestamps from the Common Voice data: the midpoint, 10 ms prior to the midpoint, and 10 ms after the midpoint. This corresponds to the primary extraction technique from each paper. Table 4.1 gives an overview of the data. Figure 4.2 shows speaker-averaged F1$\times$F2 plots for Common Voice datasets, demonstrating the vowel space across the three languages.

4.4.2 Outlier discovery

To identify outlying formants, we implemented the following procedure. Leys et al. (2018) showed that using the Mahalanobis distance metric based on the Minimum Covariance Determinant is ef-

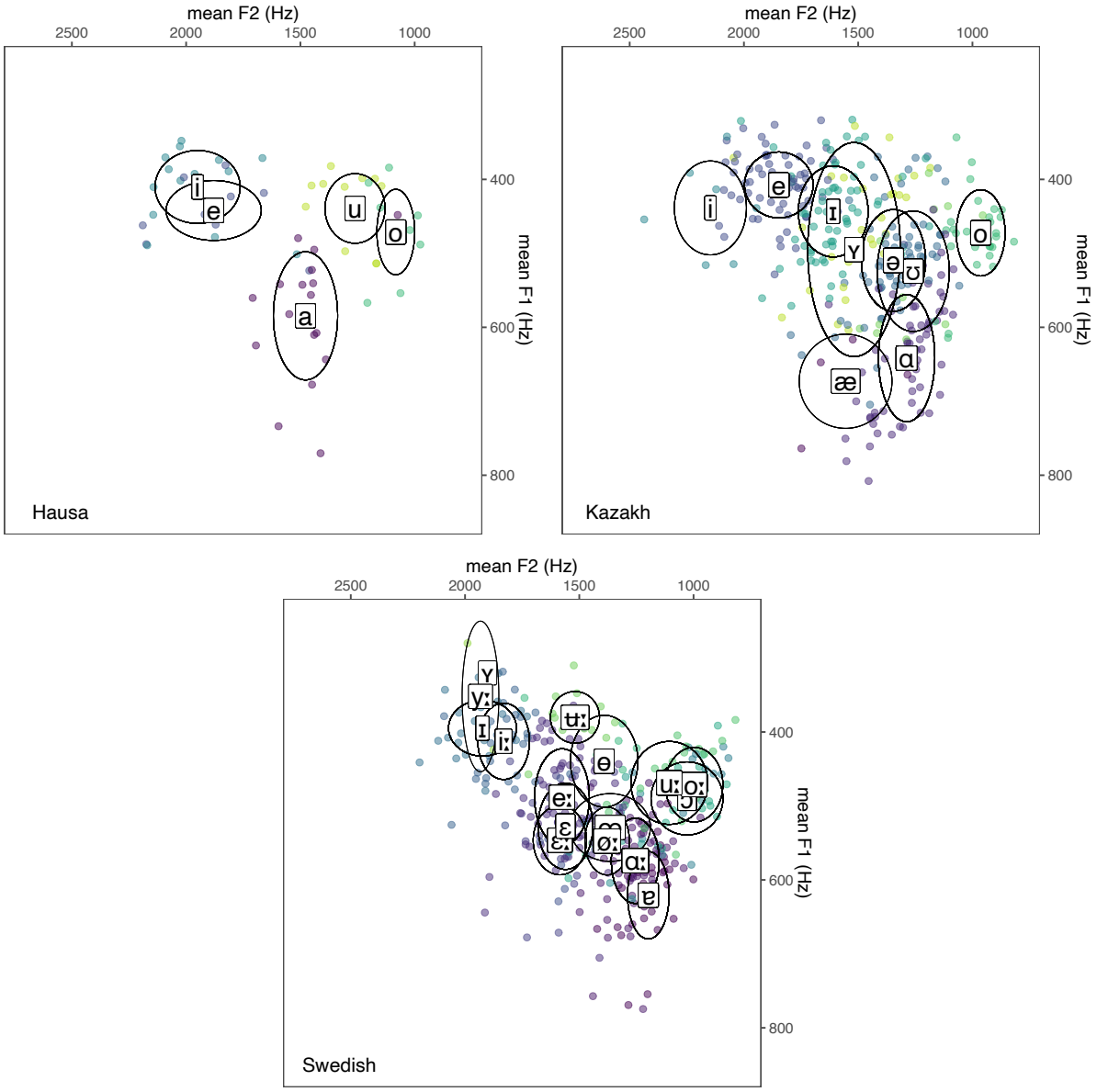

Figure 4.2: Hausa, Kazakh, and Swedish vowels from Common Voice data in F1×F2 space (Hz). Each point represents a speaker-specific pair of means, and vowel labels are placed over the grand means for each vowel. The ellipse covers ± one standard deviation from the mean across speakers.

fective for discovering multivariate outliers.[4] Squizzero and Wassink (2022) reinforced that this metric discovered outliers more accurately than using standard deviations or median absolute deviations for vowel formants. We followed suit and fitted the first two formants into one bivariate Gaussian model[5] per vowel per dataset (i.e., one model for Wilderness Kazakh /i/).[6] Each point's Mahalanobis distance from the mean followed a chi-square distribution, from which we estimated the tail 0.1% of the distribution. This percentage corresponds to an alpha value of 0.001, which is a conservative estimate for outlier exclusion. With two degrees of freedom from measuring two formants at a single timestamp, the outlier threshold corresponded to a Mahalanobis distance of 13.82. From these outliers aggregated across all vowels per dataset, 100 samples were randomly selected for manual annotation. We also randomly selected 40 "near-mean" vowels per dataset that were close to the center of each vowel distribution (Mahalanobis distance less than 1.0) for annotation, as a sanity check and to compare against the outliers. A total of 600 outliers and 240 "near-mean" vowels were analyzed across all datasets. Figure 4.3 displays the outlier discovery process with all /e/ samples in the Kazakh Common Voice data.

4.4.3 Outlier annotation

From our vowel formant audit of the outlying and near-mean vowels, we developed a new taxonomy of errors as a way to evaluate our assumptions from parts of the automated pipeline. First, though we assume the script is the transcript, deviations from the script are most directly reflected in Transcript Errors. Second, though we assume the G2P system provides a faithful phonetic transcription of the speech, if the G2P is not accurate, it could be reflected in the surfacing of Transcript Errors, Alignment Errors, or Linguistic Variations. Some of these violations also arise not necessarily from "accuracy" of the G2P system, but rather the chosen granularity of the G2P system (e.g., broad or narrow transcriptions). The Alignment Error and Formant Error categories respectively

[4]FAVE also used Mahalanobis distance to select the most reliable formants in their formant tracking method (Yuan and Liberman, 2008).

[5]Models were implemented with the MinCovDet (Minimum Covariance Determinant covariance estimator) package from Scikit-Learn.

[6]The minimum number of times the vowel must occur was 100.

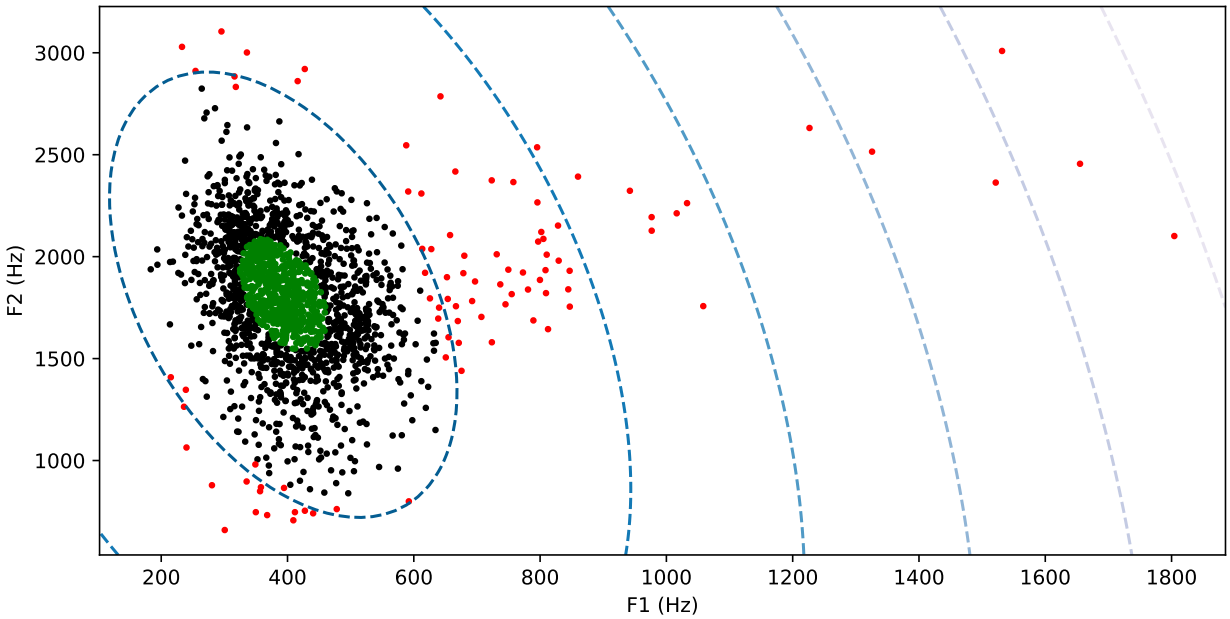

Figure 4.3: A representation of the outlier discovery method for the vowel /e/ in the Kazakh Common Voice data. Data points with a Mahalanobis distance greater than the threshold value 13.82 (i.e., outliers) are marked in red, inliers are in black, and near-mean samples (distance less than 1) are in green. The ellipses in shades of blue represent the shape of the bivariate distribution.

reflect poor performance from the forced aligner and formant tracker. Multiple error types could be applied to a single vowel in a multi-label strategy. The taxonomy includes five broad categories and several fine-grained subcategories:

1. Transcript Error

   Extra Sounds: Extra phones, syllables, or words are spoken but not transcribed.

   Extra Transcript: Extra phones, syllables, or words are transcribed but not spoken. If only the target vowel is not spoken, it is a Linguistic Deletion (see below).

   Broad: The phone sequence does not appear to match the audio at all.

2. Alignment Error

   Target Overlap: The midpoint of the window does not capture the target vowel, and the window either includes extraneous phones or it does not include the full vowel.

	Available Corpus		Analyzed Corpus						
	# Total Hours	# Total Speakers	# Analyzed Hours	# Low Speakers	# Low Utts	# Vowel Types	# Vowel Tokens	# Outliers	% Outliers
Wilderness									
Hausa	20:40	5+*	20:40	5+*	9626	5	303577	9698	3.19%
Kazakh	18:50	5+*	18:50	5+*	8085	6‡	204701	22148	10.82%
Swedish	16:45	1*	16:45	1*	9516	16	182423	15106	8.28%
Common Voice v8									
Hausa	3:23	17	0:57	8	772	5	11490	583	5.07%
Kazakh	1:27	72	1:06	46	796	11‡	10967	642	5.85%
Swedish	39:28	674	1:02	203†	1000†	16	11230	513	4.57%

Table 4.1: The Available Corpus was used for developing the acoustic models, while the Analyzed Corpus was used for the outlier analysis (in the case of Common Voice, only low-formant setting speakers were selected). ∗The number of speakers for the Wilderness data were estimated from an auditory impression of sampled data. †Swedish originally had 19,168 low utterances and 468 low speakers, before subsetting. ‡The number of Kazakh vowel types differed across dataset types due to utilizing different versions of the G2P tool, Epitran.

Broad: There is an alignment issue beyond Target Overlap. However, it can be observed that some of the transcript can be heard in the audio.

3. Formant Error

   The measured formant value does not reflect the frequency of the relevant energy band in the vowel.[7]

4. Linguistic Variation

   Deletion: Only the target vowel is absent, while the surrounding phones are present.

   Change: A different vowel than the target vowel is produced.[8]

5. Fine

   There is no apparent error.

25% of each dataset's samples were annotated by five trained linguists, while the remaining 75% of samples only had one annotator. Inter-annotator agreement across the five annotators was

[7]Formant errors can be due to the tracker getting the wrong formant (e.g. F3 instead of F2) due to coarticulation and other audio issues where the spectrum is too light, or the vowel is devoiced, etc.

[8]The observed vowel may or may not be an item in the language's inventory.

calculated with Krippendorff's Alpha (Krippendorff, 2018). Because the labels could be multiply selected, we followed Martín-Morató and Mesaros (2021) and calculated the agreement for each label. The scores were found to be reliable. Agreement across the five annotators had an average Krippendorff's Alpha of 0.86, aggregated across each of the outlier categories. Agreement tended to be highest for Wilderness outliers (0.9, compared to Common Voice outliers at 0.83) as well as for Transcript Errors (0.91, compared to Linguistic Variations at 0.84).[9]

To produce gold labels for the samples that were annotated by all five annotators, the following heuristic was applied.[10] For each possible label (e.g., Alignment: Target Overlap), if a majority (i.e., three out of five annotators) marked it positive, it was a positive label. If a minority (i.e., only one or two annotators) marked a label positive, then the label from the "most reliable annotator" was assigned. The "most reliable annotator" was designated as the annotator with the highest cosine similarity between their labels and the majority gold labels.[11]

4.5 Results

This section addresses the distribution of outlier vowel category types across languages and datasets. Table 4.1 provides an overview of the data and aggregate quantity of outliers (as determined by the Mahalanobis distance); Figure 4.4 provides raw counts of outlier types across 600 annotated outlying vowels.

As shown in Figure 4.4, the Wilderness Corpus contained more Transcript and Alignment Errors, especially from the Kazakh repository. Even Kazakh's near-mean vowels contained many Transcript and Alignment errors. This was likely an artifact of the Wilderness data processing: while the script

[9] Inter-annotator agreement methodology and calculations were conducted by Sam Briggs.

[10] This heuristic for aggregating gold labels was developed and executed by Anna Batra.

[11] The Formant Error category was added after all the data was annotated, so the "most reliable annotator" re-annotated all samples originally marked as Fine, to differentiate whether or not the vowel experienced formant tracking errors. Occasionally, other categories (e.g., Alignment Error) were found to apply instead of Fine and were re-annotated as such.

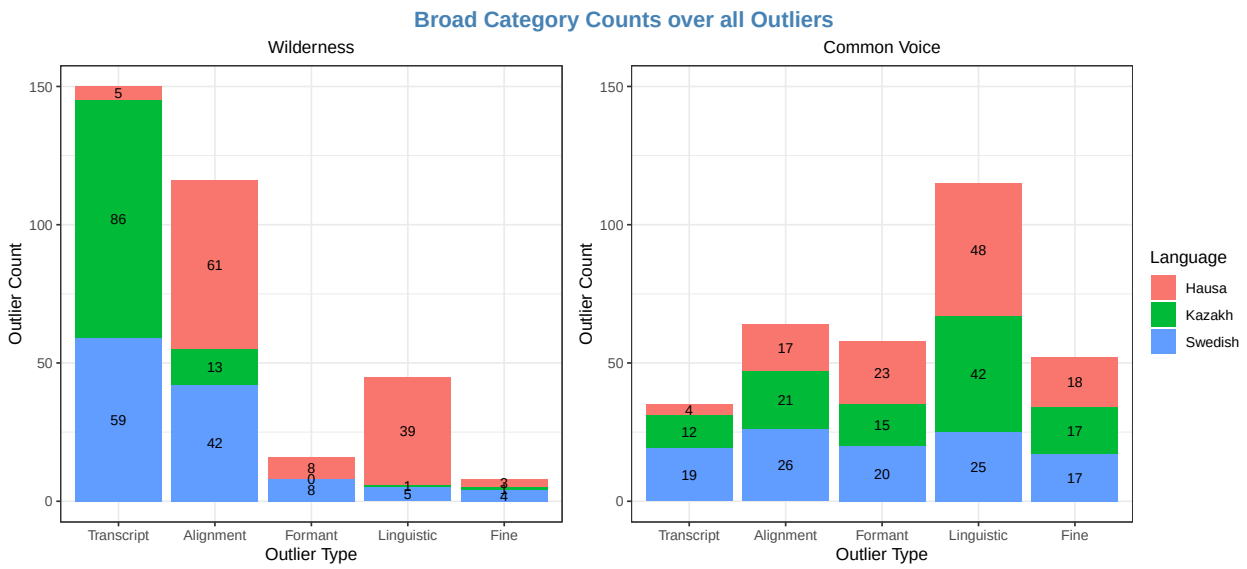

Figure 4.4: Counts of 600 outliers annotated across five broad categories in Wilderness (left) and Common Voice (right).

was manually aligned to the audio at the chapter level, individual chapter sentences were automatically aligned, resulting in some mismatched audio segments (Black, 2019). Meanwhile, the Common Voice Corpus had fewer Transcript Errors as the script-to-audio utterances were considerably shorter and manually validated; in addition, Common Voice had overall more Fine samples. Nevertheless, Common Voice had relatively more Formant Tracking Errors and Linguistic Variations, which reveal a more nuanced violation of our pipeline's assumptions.

Examining the sub-categories more closely in Figure 4.5, we observe that there were different patterns occurring within the Linguistic Variation broad category. Hausa had many more vowel changes than deletions, while the Kazakh data from Common Voice had most of its linguistic outliers due to vowel deletion. These will be analyzed in Section 4.6 as phonological processes that were not captured by the G2P assumptions.

Beyond strict counts, we can further ask how "far off" each type of outlier is by comparing the Mahalanobis distances of each outlier type. Among the set of 600 annotated outliers, Figure 4.6 displays a rough trend that Trancript and Alignment Errors stray furthest from the mean, while Fine outliers are closer to the mean. This follows the intuition in the specification of these categories.

Additionally, we seek to further understand: *What factors are correlated with vowel formants'*

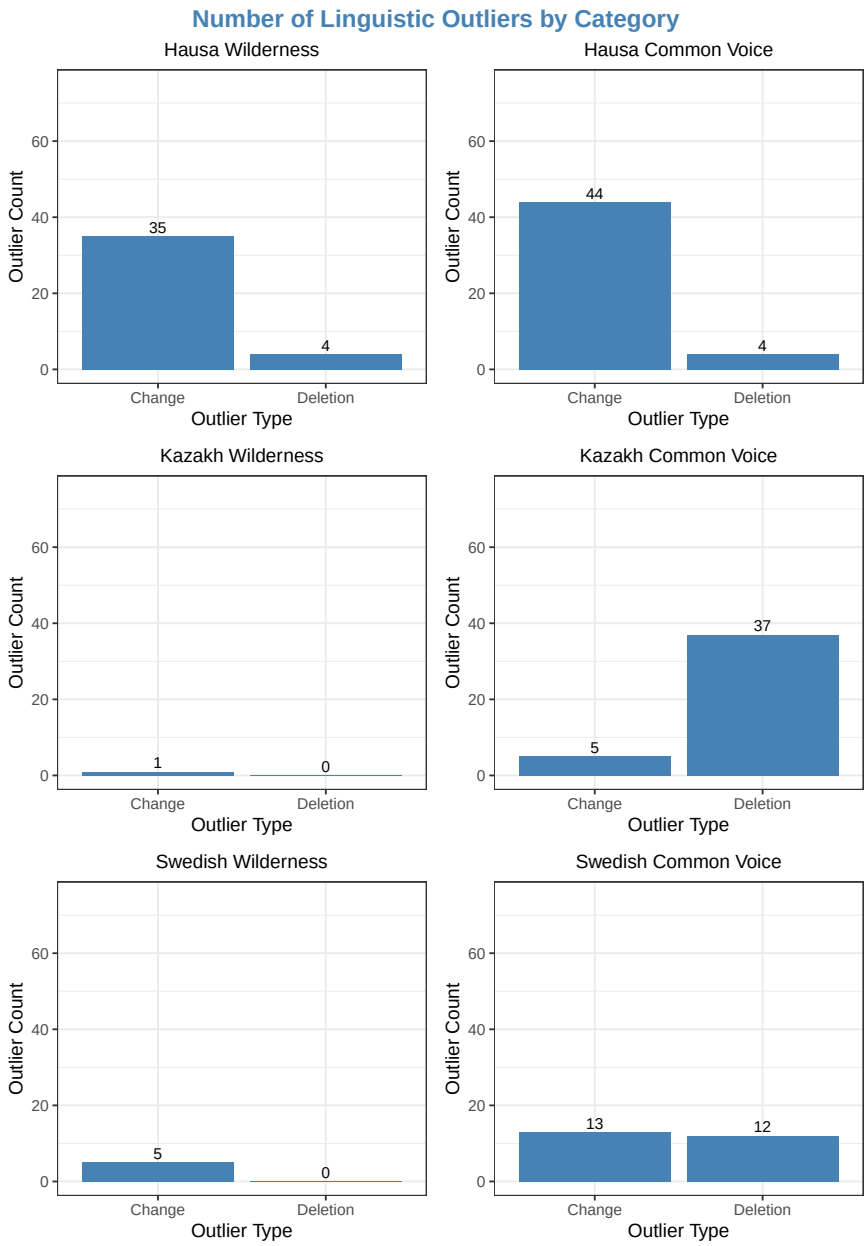

Figure 4.5: A fine-grained comparison of counts across outliers annotated from the Linguistic Variation category from Wilderness (left) and Common Voice (right) datasets, with each language as its own row.

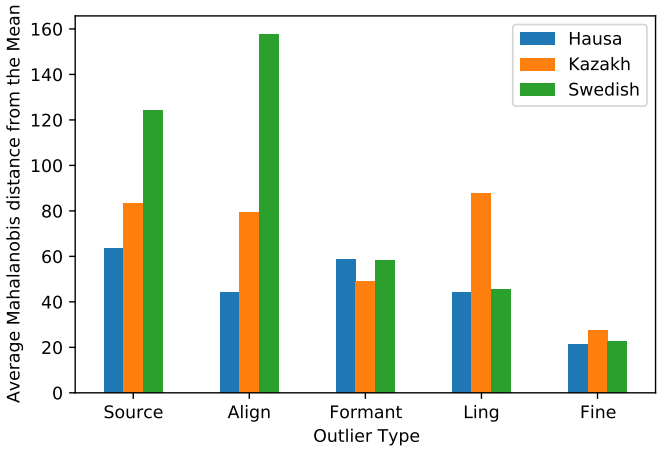

Figure 4.6: The average Mahalanobis distance from the mean across 600 annotated outliers, with each color representing a separate language.

distance from the mean? To answer this question, we ran a linear mixed effects regression analysis (Kuznetsova et al., 2017) in R (R Core Team, 2022) over all the vowels (724,132 samples) in the Wilderness and Common Voice "analyzed corpora." The dependent variable was the Mahalanobis distance, which was a continuous, positive value. Fixed effects were vowel duration, vowel height (high, mid, or low), vowel backness (front, central, or back), vowel roundness, and whether the preceding and following phones were silent. We also accounted for the random effects of language (one of Hausa, Kazakh, or Swedish) and dataset (Wilderness or Common Voice).[12]

Among the fixed effects, it appeared that vowel quality was correlated with higher Mahalanobis distances. High vowels and front vowels were more outlying ($p < 0.001$), back vowels were less outlying ($p < 0.001$), and rounded vowels were more outlying ($p < 0.001$). When either the preceding or following environments contained silence, it also increased the vowel's distance from the mean ($p < 0.001$).[13] From this finding, we can recommend that researchers be more cautious in dealing with vowels that are high, front, and rounded–when working with automatically aligned data. Vowels at the beginning or end of an utterance, as marked by silence in its preceding or

[12]The analysis could not run with the additional random effect of speaker, given that there were not enough annotated samples per unique speaker.

[13]When removing all samples that contained silence in any environment (with 595,081 samples remaining), a vowel's formants were more outlying when it was followed by an obstruent than by a sonorant ($p < 0.001$).

following environment, can also present challenges.

4.6 Case studies

Our manual audit of these outliers revealed linguistic phenomena that were not captured by our G2P assumptions.[14] The choice of granularity in the phonetic transcription can have downstream effects. If the G2P output has too narrow a transcription, it can inhibit cross-linguistic comparisons of the sound inventory. If the G2P output is too broad, it may not accurately represent sounds that undergo phonological processes like allophony, reduction, or assimilation. In both cases, the G2P output may not consistently represent the actual pronunciation. In our data, the broad G2P output resulted in a series of outliers that appeared to reflect systematic phonetic or phonological alternations in Kazakh and Hausa.

4.6.1 High vowel deletion in Kazakh

Results from our annotations indicated that Linguistic Deletions occurred most often in the Kazakh Common Voice data. According to McCollum and Chen (2021), high, short vowels in Kazakh are more susceptible to reduction than other vowels. To test this, we conducted a logistic regression in R (R Core Team, 2022) to determine if high vowels are more correlated with vowel deletion. Across all 840 annotated samples (outliers and near-means), high vowels were 1.7 times more likely to be deleted than non-high vowels ($p < 0.001$). When adding language as an independent variable to the regression, vowels in Kazakh were 2.4 times more likely to be deleted than in other languages ($p < 0.001$).[15] Interestingly, the analyzed cases of vowel deletion frequently occurred in sibilant environments. Our analysis might have picked up on this since sibilants, like vowels, can

[14]The inspiration and foundational work for each of these case studies comes from Anna Batra, Sam Briggs, Emma (Miller) Rhoden, and Qianyue (Ivy) Guo.

[15]Another notable characteristic of the deleted vowels in Kazakh is that upon annotation, it was sometimes difficult to differentiate between whether the vowel was completely deleted or just devoiced (potentially due to voicing assimilation in voiceless environments). Future outlier taxonomies could include an additional category of devoiced vowels as a subcategory of Linguistic Variation outliers.

have measurable formants, but at a considerably higher frequency than would be expected from a vowel.

4.6.2 Vowel length in Hausa

Our second case study examines the implications of vowel length in the G2P transcription of Hausa. Among our annotated Hausa vowels, 44% of the outliers and 64% of the near-means were marked as Linguistic Change. The annotators indicated that they perceived these as reduced and more centralized, e.g., [ə, ʌ, ɪ, ʊ]. Essentially, the centroids of the Hausa vowel formants were not located in the phonetic positions that the G2P inventory might suggest: /a, e, i, o, u/. While linguists do not agree on the exact vowel inventory of Hausa, most Hausa inventories from PHOIBLE include both long and short vowels which could vary in quality (Moran and McCloy, 2019). Vowel length is also not entirely predictable from the orthography. Though there is a predictable pattern that the vowel in a CVC syllable is always short (Newman, 2022), vowel length is not predictable in open syllables (i.e., CV). Therefore, even if the G2P system created separate long and short vowel categories and implemented this one rule, it would miss the open syllable alternation. The lack of vowel length distinction in our G2P system, as well as potential vowel quality differences between long and short vowels, appeared to produce inaccurate distributions of vowel formants in our analysis.

4.6.3 Kazakh 'e' /je/

Lastly, we observed during the annotation process that in Kazakh, the Cyrilic grapheme 'e' (typically pronounced [je]) has alternate phonetic productions in real speech, even though the Epitran G2P system always transcribed 'e' as [je]. We ask the question: *What are systematic environments that affect the phonetic pronunciation of /e/?* To begin, we analyzed all occurrences of 'e' within 30 utterances from the Kazakh Common Voice dataset, which amounted to 46 unique words that contained 'e'. From these, we developed phonological rules to account for these alternations. Table 4.2 displays two of these rules, along with examples.

These rules capture some of the reduction occurring in Kazakh /je/, and mainly that more often than not, the grapheme 'e' is phonetically realized as one of [i], [ɪ], or [e]. Other rules not

Rule	Description	Example
/je/ → [i] / #C[+stop]_C+#	In monosyllabic words, /je/ is reduced to [i] after a stop consonant and before a coda.	/kjeŋ/ → [kiŋ]
/je/ → [i] / C[-son]_C[+son]	In polysyllabic words in a stressed environment, /je/ is reduced to [i] after an obstruent and before a sonorant.	/sɑwdɑgjer/ → [sɑwdɑgir]

Table 4.2: Two hypothesized phonological rules for the reduction of the grapheme 'e' /je/ in Kazakh.

included in this table tell a similar story of /je/ reduction in closed syllable environments and after consonant clusters.[16] We can observe that in the Common Voice Kazakh formants plot in Figure 4.2, the centroid for /e/ has a similar F1 value as /i/, even though we expect /e/ to be a mid vowel, not a high one. This formant distribution of /e/ aligns with our analysis here that 'e' is often reduced from /je/.

Further analysis would need to be done to see if the acoustic models (especially for /e/) are affected by this alternation. This analysis also has implications for the G2P process. Would the acoustic modeling and forced alignment process be improved by a narrower transcription of 'e'? In other words, if we added these rules in Table 4.2 to generate more fine-grained dictionaries, would the alignment be more accurate? Would the distribution of formants be more accurate? Addressing these questions could contribute to the wider discussion of the interface between phonetics and phonology, as well as to the role of symbolic systems in speech technologies.

4.7 Conclusion

When a dataset lacks gold phonetic transcriptions, linguists may utilize a pipeline that takes transcribed speech, passes it through an automated grapheme-to-phoneme system, a forced aligner, and an acoustic-phonetic measurement tool (e.g., a formant tracker). To test the assumptions in this pipeline, we conducted a systematic audit of the outliers in the output of vowel formants, and

[16]These extra rules that we hypothesized do not generalize perfectly to the data; our analysis would benefit from more data points.

developed a taxonomy of errors that may arise. The distribution of these errors sheds light on common issues that arose in the automatic processing of each dataset and language.

Future work may consider discovering outliers via different methods, whether by using an a priori threshold of expected formant values or by extracting features other than formants (e.g., MFCCs). It is also worth applying our outlier audit methodology to test the corpus phonetics pipeline on different types of data (e.g., transcribed conversational speech), in different noise environments, and across more languages. While we recommend always incorporating a partial manual audit to this pipeline, automating the identification of certain outlier categories would be beneficial as well. Being able to distinguish between valid linguistic variation and a technical error is crucial, especially in the context of bias and fairness in language technologies. For example, given the presence of racial bias in ASR (see Wassink et al., 2022), it is important for systems to not treat language features of a minority group as errors. The implications of this work include a call for careful analysis of what seem like errors in the output of automated systems.

Postlude

This chapter addressed both dissertation-wide research questions. In answering RQ1 on viable methods for conducting corpus phonetics on diverse speech data, we developed a taxonomy of outliers to discern how successful each part of the pipeline was. We also addressed RQ2, showing that different types of errors and outliers revealed the inutility for corpus phonetics of datasets such as the Kazakh partition of the Wilderness Corpus.

Reflecting on this work two years after publication, we want to emphasize the reality that conducting this type of thorough outlier analysis is challenging and time-consuming. It is acceptable and may be easier to simply adhere to a cut-off threshold of removing, for example, 5% of the tail end of the distribution, and proceeding with one's research (indeed, the VoxCommunis case study in Section 3.5.1 removed samples greater than two standard deviations from the mean). Yet it is important to investigate the contents of that tail end of the distribution and if they align with the pipeline assumptions. We further discuss the theme of trusting automation in the face of diversity in Section 7.2.1.

CHAPTER 5

Testing Phone Granularity and Cross-language Modeling Strategies for Low-resource Forced Alignment: A study with Panãra field data

Overview

While the previous two chapters dealt with multilingual read speech, we now turn to a different type of diverse speech data: fieldwork speech corpora for which there is less than an hour of transcribed recordings. We address RQ1 in testing methods to make corpus phonetics viable for low-resource field data. Chapter 4 showed that the grapheme-to-phoneme (G2P) step within a forced alignment pipeline has a high impact on the quality of acoustic measures used in downstream analysis. If the input phone sequence to MFA does not match the acoustic signal, measurements can be skewed. One question that surfaced from that work is: what is an appropriate granularity of phones to produce the most accurate alignment? If a phone is broader and more phonemic, it may be more comparable cross-linguistically. This may be important for low-resource settings where there may not be enough data to train a model from scratch; instead, a large pretrained model of another language can be utilized in cross-language forced alignment, and broader phones make the two language inventories more compatible. On the other hand, if phones are modeled narrower and more allophonically, they may more accurately represent the acoustic signal and processes such as reduction, assimilation, and allophony.

In this chapter, we manipulate the precision of the G2P output to try to achieve better alignment (or reduce Alignment errors) in a low-resource language setting. We also explore cross-language modeling techniques. This chapter is largely equivalent to the conference proceedings of *Interspeech* titled *The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panãra* (Ahn et al., 2024).

5.1 Abstract

Automating the time-alignment of phonetic labels in speech facilitates research in language documentation, yet such phonetic forced alignment requires pretrained acoustic models. For low-resource languages, this raises the question as to how and on which data the acoustic model should be trained. To align data from Panãra, an Amazonian indigenous language of Brazil, we investigated three approaches for forced alignment of low-resource languages using the Montreal Forced Aligner. First, we implemented a novel approach of manipulating the acoustic model granularity from phone-specific to increasingly broader natural class categories in training language-specific Panãra models. Second, we trained cross-language English models under two granularity settings. Third, we compared these models to a large, pretrained Global English acoustic model. Results showed that broadening phone categories can improve language-specific modeling, but cross-language modeling performed the best.

5.2 Introduction

The alignment of speech to utterance, word, and phone segments is a useful and often necessary step to studying language phenomena. Language research often faces the "transcription bottleneck", where many audio recordings may exist, but there is limited transcribed data (Shi et al., 2021). Overcoming the transcription bottleneck would also increase the amount of usable data for Natural Language Processing (NLP) applications, such as machine translation and automatic speech recognition, that could serve endangered language communities. Obtaining these transcriptions and alignments is time-consuming to produce manually. Pretrained acoustic models are available for many high-resource languages with large amounts of annotated speech data, yet for lower-resource languages, acoustic models with minimal data need to be trained from scratch.

Alternatively, "cross-language forced alignment" can be implemented in which pretrained acoustic models of high-resource languages are used to align a lower-resource language. While cross-language forced alignment benefits from robust amounts of training data for the acoustic model development, it lacks language-specificity for the target language. The phoneme inventory of the

target language in a cross-language setting generally needs to be remapped to that of the pretrained acoustic model, as phoneme inventories are rarely one-to-one between languages. There are often sounds in target languages of research that do not exist in high-resource languages such as English. Another complication is that phonemic units are commonly debated for a given language, and defining a language's inventory is by no means trivial[1] (Anderson et al., 2023). For example, the Hausa repository in the online PHOIBLE database (Moran and McCloy, 2019) contains five entries with the stated inventory size ranging from 31 to 46 segments (Archive, 2019; see discussion in Section 4.6.2).

Given these concerns, we propose an alternative strategy to retain language-specificity in acoustic model training, which increases the amount of data per phone category by broadening phone categories to larger natural classes. Broad class models may prove to be more universal and would not discriminate against a target language that does not have the exact same phone set as the training data. It also may facilitate language-specific model training by expanding the number of instances per phone. As the task of alignment is to identify the transition point from one segment to the next in the acoustic signal, this may still be achievable with a coarser representation of a segment. This strategy was shown to improve cross-language, sentence-level alignment across five European languages in acoustic model development (Hoffmann and Pfister, 2013). Alternatively, DiCanio et al. (2013) showed that when comparing two English acoustic models' alignments on Yoloxóchitl Mixtec data, the model with a narrower, "context-sensitive" set of phones outperformed the broader, more phonemic model. The narrower phones seemed to better match those of Yoloxóchitl Mixtec; for example, Yoloxóchitl Mixtec only had unaspirated stops, which mapped well to the English model having separate categories for aspirated vs unaspirated stops. A major limitation of creating narrower models is data quantity, and especially in the language-specific setting, we are curious if broader models—that do not require as much data for training—can be beneficial. To our knowledge, our work is the first to utilize broad phone classes in both language-specific and cross-language modeling for phone-level alignment.

[1]For instance, it took Myriam Lapierre approximately 30 weeks of *in-situ* fieldwork in the Panãra Indigenous land, distributed over 5 years, to determine the phonemic inventory of Panãra.

In this work, we defined three settings of phonetic granularity and evaluated the forced alignment performance of Panãra, an Amazonian language of Brazil. We analyzed these settings within both language-specific modeling of Panãra (i.e., training a model on just Panãra), and cross-language modeling (i.e., training on English and adapting to Panãra). We address the following chapter-specific research questions:

1. Does broadening phone categories improve alignment performance in language-specific training?

2. Does broadening phone categories improve alignment performance in cross-language training and modeling?

3. Do any of these strategies perform better than using a large, pretrained English model in cross-language alignment?

5.3 Data

We utilized datasets from two languages: Panãra, an endangered language that is undergoing active documentation by Myriam Lapierre, and English, a high-resource language, which allowed us to investigate generalizability of the methods.

Panãra (ISO-639-3: kre) is a Jê language spoken in the state of Mato Grosso, Brazil, with approximately 700 speakers. Its phoneme inventory, displayed in Tables 5.1 and 5.2, includes typologically less common nasality and length contrasts in both vowels and consonants (Lapierre, 2023b). Our Panãra dataset consisted of four free-speech recordings from four different speakers—two male and two female; the utterances span 35 minutes of speech. There were a total of 771 utterances that averaged 2.76 seconds each. All data was transcribed orthographically by Myriam Lapierre and corrected with a native speaker of Panãra.[2] Phonetic boundaries from two of the four

[2]Though hours of recording were plenty, our available data was limited by the time-consuming process of transcribing audio into Panãra orthography, which requires approximately two hours of work per one minute (100 words) of audio. Orthographic transcription is a task that requires expert knowledge and the presence of both the field linguist and one of the few fully literate speakers of Panãra.

recordings have been manually hand-corrected by a trained phonetician.

	Bilabial	**Alveolar**	**Palatal**	**Velar**
Singleton obstruent	p	t	s	k
Geminate obstruent	pː	tː	sː	kː
Singleton nasal	m	n	ɲ	ŋ
Geminate nasal	mː	nː		
Approximant	w	ɾ	j	

Table 5.1: Consonant phonemes in Panãra, as taken from Table 1 in Lapierre (2023b, p. 187).

Short Oral	**Short Nasal**	**Long Oral**	**Long Nasal**
i	ĩ	iː	ĩː
ɯ	ɯ̃	ɯː	
u	ũ	uː	
e	ẽ	eː	ẽː
ɤ	ɤ̃	ɤː	ɤ̃ː
o	õ	oː	õː
ɛ		ɛː	
a		aː	
ɔ		ɔː	

Table 5.2: Vowel phonemes in Panãra, as modified from Table 5 in Lapierre (2023b, p. 206).

We also selected an English dataset that could be retrained from scratch to incorporate broad phone categories. The motivation for a broad category English model is that it would mimic a more language-independent, multilingual model that would be inclusive of unseen sounds in the target language. The TIMIT English dataset consists of read sentences from 630 speakers across eight dialect regions in the US (Garofolo et al., 1993). This corpus was selected as it has been manually transcribed at the phonetic level, as well as hand-aligned for phonetic boundaries. After removing some problematic data,[3] we utilized just under four hours of TIMIT speech from 519 speakers. This produced 5190 utterances with an average length of 3.08 seconds.

[3]Two of the eight training folders did not run through the Montreal Forced Aligner, so we excluded that data.

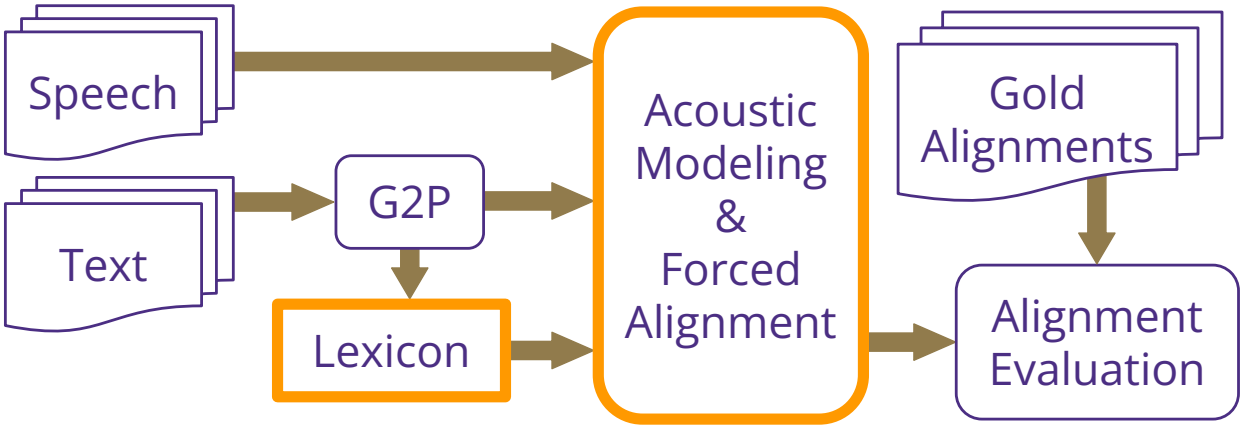

Figure 5.1: Our pipeline takes speech and its corresponding text transcriptions as input. We produce a phone sequence with a grapheme-to-phoneme (G2P) system. We then manipulate the lexicon for each model to allow for broadening phone categories and conducting cross-language alignment. We train acoustic models and produce phone-level alignments, which we then evaluate against human-annotated "gold" alignments.

5.4 Methodology

Our pipeline included the following steps, described in detail below: (i) grapheme-to-phoneme conversion, (ii) lexicon creation, (iii) data management, (iv) acoustic model development and forced alignment, and (v) evaluation of aligned boundaries. As highlighted in Figure 5.1, the lexicon and acoustic modeling are the main components we manipulated for our experiments.[4]

5.4.1 Grapheme-to-phoneme conversion

First, the Panãra data went through an automatic rule-based grapheme-to-phoneme (G2P) system to convert orthography into phones.[5] There were instances of code-switched Portuguese words in the Panãra speech, for which we applied the Epitran G2P in Portuguese (Mortensen et al., 2018). There were only 34 Portuguese words in the entire dataset, and these had been tagged with the language specification directly in the transcription. None of the G2P output was hand-corrected.

	Map from	Map to			
	Panãra-Explicit	**No-Diacritics**	**Broad (SCA)**	**TIMIT-Explicit**	**Global English**
Panãra Phone Types	63	29	17	25	30
Vowels	a	a	A	a	a
	aː	a	A	a	aː
	ã	a	A	a	a
	e	e	E	eɪ	e
	eː	e	E	eɪ	eː
	ẽ	e	E	eɪ	e
	ẽː	e	E	eɪ	eː
	i	i	I	i	i
	iː	i	I	i	iː
	ĩ	i	I	i	i
	ĩː	i	I	i	iː
	o	o	U	oʊ	o
	oː	o	U	oʊ	oː
	õ	o	U	oʊ	o
	õː	o	U	oʊ	oː
	u	u	Y	u	u
	uː	u	Y	u	uː
	ũ	u	Y	u	u
	ũː	u	Y	u	uː
	ɔ	ɔ	U	ɔ	ɔ
	ɔː	ɔ	U	ɔ	ɔ
	ɛ	ɛ	E	ɛ	ɛ
	ɛː	ɛ	E	ɛ	ɛː
	ɤ	ɤ	E	ə	o
	ɤː	ɤ	E	ə	oː
	ɤ̃	ɤ	E	ə	o
	ɤ̃ː	ɤ	E	ə	oː
	ɯ	ɯ	I	ɪ	u
	ɯː	ɯ	I	ɪ	uː
	ɯ̃	ɯ	I	ɪ	u
Flaps	ɾ	ɾ	R	ɾ	ɾ
Approximants	j	j	J	j	j
	w	w	W	w	w
Fricatives	h	h	H	h	h
	s	s	S	s	s
	sː	s	S	s	s
Nasals	m	m	M	m	m
	mː	m	M	m	m
	n	n	N	n	n
	nː	n	N	n	n
	ŋ	ŋ	N	ŋ	ŋ
	ɲ	ɲ	N	n	n
Stops	k	k	K	k	k
	kː	k	K	k	k
	p	p	P	p	p
	pː	p	P	p	p
	t	t	T	t	t
	tː	t	T	t	t
Portuguese	ĩ	i	I	i	i
	ũ	u	U	u	u
	ɐ	ɐ	E	a	ɐ
	ɐ̃	ɐ	E	a	ɐ
	ɛ̃	ɛ	E	ɛ	ɛ
	b	b	P	b	b
	d	d	T	d	d
	g*	g*	K	g	g*
	w̃	w	W	w	w
	ʃ	ʃ	S	ʃ	ʃ
Other	c	c	C	k	k
	g	g	K	g	g*
	jː	j	J	j	j
	l	l	L	l	l
	ɔ̃	ɔ	U	ɔ	ɔ

Table 5.3: (Caption next page.)

Table 5.3: (Previous page.) The lexicon mapping from the explicit Panãra setting to each of the other language-specific and cross-language settings. The number of phone types is the number of distinct phone categories in that model that corresponds to Panãra phones. (For counts of the total phone types in each of the trained models, see Table 5.5 in the Results section.) The Broad category symbols correspond to the classes in List (2012)–for the full list, see Figure 5.2; all other symbols are in IPA. The Other section at the bottom of the table is output from the Panãra grapheme-to-phoneme (G2P) conversion that is not in the phone inventory. It may have been a named entity where the characters from the word transcription were not recognized as valid Panãra orthography. To note, [c] came from the name "Claudio" uttered several times, which for the mapping to explicit English models, we chose the more acoustically accurate [k] sound. The g* refers to the Latin small letter script g. It is a subtly different character than g. The geminate consonant [jː] does not exist in the Panãra inventory, but was created as an artifact of the post-processing of the grapheme-to-phoneme conversion; it appears only once in all the Panãra data.

5.4.2 Lexicon creation

Second, we created three settings of phone categories in the lexicon creation stage that informed the language-specific training for Panãra-only acoustic models: "Explicit", "No Diacritics", and "Broad natural class". The first setting, "Explicit", used the given phone categories, which consisted of 63 phonetic labels from the default output of the G2P system. The second setting, "No Diacritics", had 29 phonetic labels where all length and nasalization markers were ignored. Lastly, we created a "Broad" phone categories setting, which followed the natural classes from the Sound-Class-Based Phonetic Alignment (SCA) tool (List, 2012). Each of our phones was mapped to one of the 28 SCA classes that included six vowels, five fricatives, three plosives, one affricate, two nasals, one laryngeal, two approximants, one trill/tap/flap, and seven tones (see Figure 5.2). The Panãra phone set had 17 of these sound classes.

For our second chapter-specific research question, we trained the TIMIT English acoustic models under two phonetic granularity settings. The "Explicit" setting had 46 fine-grained phonetic labels.[6] The "Broad" setting also used the SCA classes (List, 2012), which included 19 of these sound classes.

For our third chapter-specific research question, we conducted a final comparison of a large pre-trained English model's alignment on Panãra data to address whether large cross-language acoustic

[4]All code is publicly available at https://github.com/emilyahn/force_align.

[5]This G2P system is an internal tool created by collaborator Teela Huff.

[6]While TIMIT originally had 61 phone labels, they are not typically all used (Fernández et al., 2008). We collapsed several categories such as "b-closure" with [b], and [ʌ] with [ə].

Table 4. The SCA sound class model.

No.	Cl.	Description	Examples
1	A	unrounded back vowels	a, ɑ
2	B	labial fricatives	f, β
3	C	dental / alveolar affricates	ʦ, ʣ, ʧ, ʤ
4	D	dental fricatives	θ , ð
5	E	unrounded mid vowels	e, ɛ
6	G	velar and uvual fricatives	ɣ , x
7	H	laryngeals	h, ʔ
8	I	unrounded close vowels	i, ɪ
9	J	palatal approxoimant	j
10	K	velare and uvular plosives	k, g
11	L	lateral approximants	l
12	M	labial nasal	m
13	N	nasals	n, ŋ
14	O	rounded back vowels	œ, ɒ
15	P	labial plosives	p, b
16	R	trills, taps, flaps	r
17	S	sibilant fricatives	s, z, ʃ, ʒ
18	T	dental / alveolar plosives	t, d
19	U	rounded mid vowels	ɔ , o
20	W	labial approx. / fricative	v, w
21	Y	rounded front vowels	u, ʊ, y
22	0	low even tones	11, 22
23	1	rising tones	13, 35
24	2	falling tones	51, 53
25	3	mid even tones	33
26	4	high even tones	44, 55
27	5	short tones	1, 2
28	6	complex tones	214

Figure 5.2: The 28 sound classes from the SCA sound class model, as taken from Table 4 in List (2012). We used these classes as the "Broad natural class" settings for training our Panãra and TIMIT English models.

models may still outperform the smaller acoustic models. The Global English model (McAuliffe and Sonderegger, 2023) was the largest pretrained acoustic model available from MFA, consisting of 3,770 hours of speech from regions including the US, UK, Nigeria, and India. As this model would have been difficult to retrain from scratch, we did not apply our broad phonetic categories methodology.

To align Panãra data with the Explicit versions of these English models, we also created lexicons that mapped explicit Panãra phones to the respective English phone sets. Table 5.3 displays the Panãra phone set and its mappings to the different settings and cross-language phone inventories, while Table 5.4 displays an example Panãra utterance and the phone sequences it corresponded to using each trained model's lexicon. The main differences between the explicit English phone inventories and Panãra were the following: the Global English model did not have nasalized vowels or lengthened consonants, the TIMIT Explicit model did not contain any nasalized or lengthened sounds, and neither of these had the unrounded back vowels [ɯ] and [ɤ].

Panãra Orthography	Haa mämä jynkjân rasu hapôô
Panãra Explicit	h a: m ɤ̃ m ɤ̃ j ɯ ŋ k j ɤ n ɾ a s u h a p o:
Panãra No Diacritics	h a m ɤ m ɤ j ɯ ŋ k j ɤ n ɾ a s u h a p o
Broad (SCA)	H A M E M E J I N K J E N R A S Y H A P U
TIMIT English	h a m ə m ə j ɪ ŋ k j ə n ɾ a s u h a p oʊ
Global English	h a: m o m o j u ŋ k j o n ɾ a s u h a p o:
Translation	*"Well, this time you arrived (where I am)"*

Table 5.4: An example utterance in Panãra including the original orthography, the phone sequence mapping per lexicon and acoustic model, and the English translation.

5.4.3 Data management

For a fair comparison, the Panãra dataset was split into three speakers for training and one speaker for testing. Since we have manually-annotated "gold" alignments for two speakers (one male and one female), we implemented two train/test configurations that each held out one of these two speakers for testing. This cross-validation scheme allowed us to not contaminate the test set with a speaker from the train set, yet utilize all of our manually-annotated data. This resulted in the Panãra train set totaling either 27 or 31 minutes of speech, and the aggregated test set including 12 minutes of speech. For TIMIT, we created two versions of the train set; the "full" version had 224 minutes and 495 speakers, and the "small" version had 26 minutes and 51 speakers. This "small" TIMIT train set was devised to have a comparable duration to the Panãra train set. The resulting acoustic models were then used to align the Panãra test set.

5.4.4 Acoustic model development and forced alignment

After preparing speech audio files, their corresponding phone sequences, and the lexicons, we passed these through the Montreal Forced Aligner (MFA) Version 2.2.17, a tool easily configurable to train and adapt acoustic models from scratch (McAuliffe et al., 2017). For our experiments, we applied the three different granularity settings to a Panãra-only train and test scenario, and then applied two granularity settings to train the TIMIT English models. All English models were adapted using the Panãra train set, which slightly modified the acoustic model parameters. We used the default MFA settings for all model training and adaptation, which employed a triphone GMM-HMM architecture with MFCCs.

Trained Dataset	Trained Settings (# Phone Categories)	Precision (%) ↑
Panãra	Explicit (63)	60.20
	No Diacritics (29)	62.35
	Broad (17)	61.92
TIMIT English	Explicit Full (46)	62.65
	Explicit Small (46)	66.09
	Broad Full (19)	56.07
	Broad Small (19)	61.14
Global English	Explicit (100)	**69.82**

Table 5.5: Alignment performance across various systems on the Panãra test set, as measured by phone boundary onset displacement within 20 ms. English-trained models have been adapted to the Panãra train set.

5.4.5 Alignment evaluation

Our methods for measuring alignment performance used phone onset boundary differences between system and manually-annotated "gold" versions. We define boundary "precision" as the percentage of system onsets that are within 20 milliseconds of the corresponding gold onsets (higher is better) (MacKenzie and Turton, 2020; McAuliffe et al., 2017).[7]

5.5 Results

For our first chapter-specific research question related to the influence of broadening phone categories on alignment performance in language-specific training, we examine the top three rows in Table 5.5. These show that for language-specific training, i.e., the Panãra-trained models, broadening the phone categories beyond the Explicit setting was beneficial. Between removing diacritics and using the broad SCA natural classes, the No Diacritics model performed the highest at a precision of 62.35% for the onset boundary to be within 20 milliseconds of the gold boundary. The Broad SCA Panãra model performed slightly worse at 61.92%, suggesting that broadening the categories too much lost specificity for the acoustic model phone categories. The Explicit Panãra model had

[7] We use the term "precision" in this chapter instead of the term "accuracy", which was used in the proceedings paper (Ahn et al., 2024). This allows for consistent terminology with Chapter 6.

63 phone categories, which was likely too many for its training data size, given that its performance on the test set was only 60.2%.

As for our second chapter-specific research question related to cross-language acoustic model training and alignment, we examine the middle four rows in Table 5.5. For the TIMIT English-trained models that were adapted on Panãra, using broad phone classes degraded performance; Broad SCA models for each data size (Full and Small) performed worse than their respective Explicit models. Even though the broad TIMIT models were aimed at mimicking language-agnostic, multilingual models, this did not help in alignment of Panãra data. The best TIMIT model was the Explicit Small, which yielded a 66.09% precision.

As for our third chapter-specific research question, if any of these strategies above can outperform a large, pretrained model, our answer is no. The Global English model performed the highest across all models. Even with 100 phone categories in its acoustic model, the 30 categories (see Table 5.3) that pertained to the Panãra phone set were language-specific enough to outperform any of the other models.

Beyond these primary questions, we also sought to understand the performance patterns within specific natural sound classes across the different configurations. Figure 5.3 reveals several findings. Among vowels, short vowels were more precise than long vowels, and oral vowels were more precise than nasal vowels. Within Panãra-only systems, the No Diacritics model performed better than the Broad model for long and nasal vowels, implying that combining the vowels into fewer natural class categories lost specificity in the acoustic model training. Some other notable phone-specific patterns involved [h], which performed particularly poorly among fricatives and overall, as well as the tap [ɾ] which performed especially poorly for all Panãra models. The best TIMIT English model outperformed all the other Panãra models except for among long vowels, with which it performed similarly. The Global English model was on par with or outperformed all the other models for each natural class except approximants.

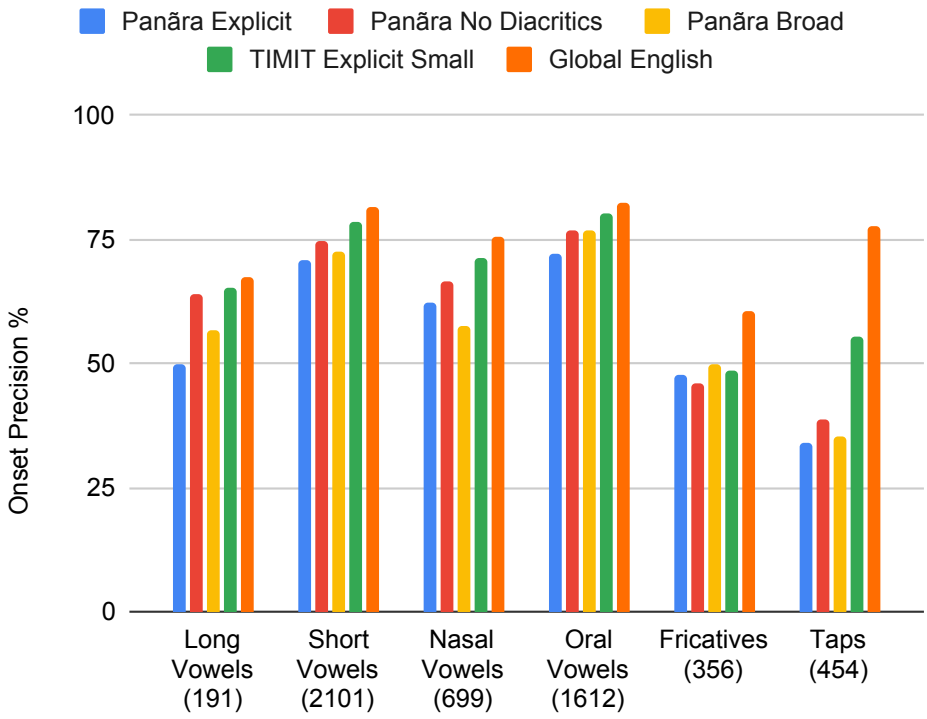

Figure 5.3: Onset boundary precision within 20 ms (y-axis) across a selection of natural classes presented with their token counts (x-axis), on Panãra test data. The colored bars represent five of the systems.

5.6 Discussion

5.6.1 Training strategies

The above results demonstrated that broadening phone categories can be beneficial on limited data, though cross-language forced alignment using the large Global English model had consistently high performance. Considering cross-language configurations, the Global English model outperformed all others, likely due to its nearly 4,000 hours of regionally diverse training data. In addition, the Explicit Small TIMIT model performed better than all remaining cross-language and all language-specific configurations. One potential explanation for this is that the training data for the Explicit Small model was more balanced in its regions of speakers. The higher performance of the Explicit models supports the findings of prior work on Yoloxóchitl Mixtec (DiCanio et al., 2013). Overall, we recommend that language researchers adapt the Global English model, or a similar large pretrained

model, to align their own target dataset.

5.6.2 Natural class analysis

One interesting finding was that long and nasal vowels showed a higher degree of inaccuracy in boundary placement, even for the best-performing Global English model. From a typological angle, long vowels are less common in the world's languages than short vowels, as are nasal vowels compared to oral vowels (Gordon, 2016). The counts of each vowel type in our Panãra data also reflected this imbalance. Despite the diversity of speakers and dialects for the Global English model, typological trends suggest that there were likely fewer instances of vowels being lengthened (which were in the inventory) or nasalized.

In addition, the phonological grammar of Panãra provides some insights into our findings, revealing how language-specific phonological processes make the task of alignment especially challenging. Across all systems, it was more difficult to place the onset boundary for fricatives, nasals, and stops. Fricatives are often distinct in their acoustic and spectral features, yet had low accuracy in our models. As discussed in Lapierre (2023a), the glottal fricative [h] can be variably inserted in onsetless syllables, especially those that are prosodically prominent, such as in word-initial or stressed syllables. Under qualitative review of the data, we observed that word-initial [h], which was consistently present in the transcription, was pronounced at times and silent at others. Because the MFA must assign every phone in the given phone sequence a non-zero duration, deleted or invisible segments can throw off the performance.

Finally, the poor performance of our models on taps may be explained by the fact that our G2P system did not encode a relevant phonological rule. In particular, Lapierre (2023b) described a process of excrescent vowels in complex onsets, such as in the word /kɾɤ/ → [kˠɾɤ] 'thigh'. As a result, the MFA forced an alignment of /kɾɤ/ to an acoustic signal of the form [kVɾɤ], thus encountering an additional, unexpected vowel.

5.7 Conclusion

This chapter investigated the question: how does broadening phone categories of acoustic models affect the temporal degree of accuracy of phone-level forced alignment? We defined three granularity settings of phone categories as input to our acoustic models trained on either Panãra or English, and found mixed results on a Panãra test set: broadening phone categories can be helpful in language-specific training on very limited data, but cross-language alignment with a large Global English model outperformed other configurations. Can our findings for Panãra and English be replicated across other languages?

A main limitation of this work is that because the number of speakers in all of our Panãra data was only four, idiosyncrasies and sociolinguistic factors could have strongly affected our results. The two speakers in the Panãra train set were younger than those in the test set, and age is correlated with higher proficiency in discourse skills such as rhythm and intonation. Also, while most Panãra speakers are monolingual, young males are the only group with conversational proficiency in Portuguese as a second language (Lapierre, 2023b). Although we balanced speaker gender in our data splits, speaker idiosyncrasies could have prevailed. Additionally, the Portuguese and miscellaneous phones present in our Panãra data may have added noise to these results.

One direction for future work is to further investigate the granularity settings of the lexicon. Distinctive features are binary properties in phonology that describe place and manner of articulation, as well as voicing and other properties (Jakobson et al., 1951). As the SCA groupings of natural classes from List (2012) may not optimally reflect the similarity of sounds in the acoustic representation space, using distinctive features or an unsupervised clustering method to group sounds could aid in identifying more optimal groupings for broader sound classes. A multilingual aligner such as one developed by Zhu et al. (2024), trained on data with diverse phone inventories, could also be compared to these strategies.

Postlude

This chapter operationalized the viability of methods from RQ1 into testing two acoustic modeling strategies: cross-language modeling and expanding phone granularity. In the face of limited data

from a fieldwork setting, creative strategies are important for being able to use automated systems to conduct phonetics research.

Continuing the discussion on sound unit representations, we reflect on the impact of similarities and differences between language inventories in cross-language acoustic modeling. In this chapter, we stated that American English does not have lengthened or nasalized contrasts in its phonemic inventory. However, these sounds do exist in the acoustics of American English speech as allophones. As such, Panãra may not be acoustically too dissimilar to English, and the performance of the Global English model could be attributed to a relative overlap between the two languages' sound inventories. It would be interesting in future work to use a pretrained English model to model the acoustics of a language that contains sounds that do not appear at all in English, allophonically or not. In ongoing work, we are collaborating with phoneticians on measuring the acoustics of non-pulmonic sounds such as ejectives and implosives. As these types of sounds rarely occur in English and many other high-resource languages, the methodologies used in this chapter may not be relevant in the acoustic modeling of these sounds. Further discussion on the theme of sound unit representation can be found in Section 7.2.2.

CHAPTER 6

Acoustic Modeling Strategies for Low-resource Forced Alignment of Code-switched Speech: A study with Urum–Russian field data

Overview

This chapter expands the scope of data diversity for which we aim to study corpus phonetics.[1] The presence of Panãra–Portuguese code-switching in the previous chapter, along with the code-switching found in two other field datasets (see Chodroff et al., 2024a), inspired our work in this chapter where we continue to test acoustic modeling and forced alignment strategies for field data that is highly code-switched.

We have discussed methods for using automation in the corpus phonetics pipeline for large, multilingual speech corpora. We also have delved into the realm of field data that is low-resource. Yet one aspect of field data is that there may be more than one language present in the recordings. Another language of broader communication may be used to elicit the target language, or speakers of the target language may switch back-and-forth between the target language and other languages. This chapter investigates ways to conduct corpus phonetics when the input data is multilingual and highly code-switched. How does this affect how we model the acoustics of this language? What are the implications for the quality of phonetic analyses on downstream output? We answer both broad dissertation-wide research questions (RQs) in this chapter's specific research questions.

6.1 Abstract

Code-switching, using multiple languages in a single utterance, is a common means of communication. In the language documentation process, speakers may code-switch between the target

[1] This work has been submitted to the Field Matters workshop of the Association for Computational Linguistics conference, 2025, with co-authors Eleanor Chodroff and Gina-Anne Levow.

language and a language of broader communication; however, how to handle this mixed speech data is not always clearly addressed for speech research and specifically for a corpus phonetics pipeline. This chapter investigates best practices for conducting phone-level forced alignment of code-switched field data using the Urum speech dataset from DoReCo. This dataset comprises 117 minutes of narrative utterances, of which 42% contain code-switched Urum–Russian speech. We demonstrate that the inclusion of Russian speech and Russian pretrained acoustic models can aid the alignment of Urum phones. Beyond using boundary alignment precision and accuracy metrics, we also discovered that the method of acoustic modeling impacted a downstream corpus phonetics investigation of code-switched Urum–Russian.

6.2 Introduction

In the language documentation and analysis pipeline, the fact that the target language is often mixed with another language of broader communication, even within a single utterance, is often overlooked or explicitly ignored. Code-switching is a method for multilingual speakers to communicate in more than one language, where the mixing of languages within a single utterance can be referred to as code-mixing.[2] If the goal of the fieldwork is to document the language of interest, field linguists and language communities may wish to ignore the language of broader communication. On the other hand, it may be useful to include the mixture of languages in the analyzed data for methodological or scientific purposes, for a variety of reasons. This could be that the added data improves methodological precision, or it could be that this mixing more properly reflects how that target language is actually used by the speakers.

Regardless of the analytical use, inclusion of the code-switched language data may benefit processes within the corpus phonetics pipeline. Phonetic forced alignment is a critical part of this pipeline. Generally, when there is more data available to train the acoustic models, the models can perform better alignment. In the case of code-switched speech, there is a question, however, of whether to use only the target language data for training the acoustic models, or to use all of

[2]For this chapter, we use code-mixing and code-switching interchangeably, though recognize that the current research focus is on code-mixing.

the linguistic data for training the acoustic models. If we include code-switched speech instead of ignoring it during training, there would be more data per speaker, which could help build more robust phone-specific acoustic models (as speculated in Chodroff et al., 2024a).

Two prior studies incorporated a language of broader communication when training forced alignment systems on field data, though the effects of this mixed language speech were not explicitly studied. In Chapter 5, we included Portuguese content when developing acoustic models for Panãra, an Amazonian language of Brazil. Chodroff et al. (2024a) retained the Russian speech content in the acoustic modeling of Evenki, a Tungusic language, and Urum, a Turkic language, which is also used in this work (Kazakevich and Klyachko, 2022; Skopeteas et al., 2022).

More relevant to the present study is work by Pandey et al. (2020) who compared methods of training and aligning code-switched Hindi-English speech. Three acoustic models were trained with the Montreal Forced Aligner: Hindi-only, English-only, and Hindi–English mixed, and they discovered that the combined model best aligned English-only speech. It was unclear, however, if the high performance from the Hindi–English mixed acoustic models was due to that model having more tokens in its training data than the other models. Our work extends these findings to a low-resource, field data scenario, and we carefully controlled the variable of training data quantity. We investigated whether including code-switched data could improve the alignment performance of a target low-resource language.

We ask the following research questions:

1. Does the inclusion of Russian code-switched data in acoustic model training help the alignment of target Urum data?

2. Does the method of acoustic modeling impact a downstream corpus phonetics investigation of code-switched Urum–Russian?

In this chapter, we introduce the Urum language (Section 6.3), then discuss the methodology of data preparation, acoustic modeling and forced alignment, evaluation and analysis (Section 6.4). We used the Montreal Forced Aligner (McAuliffe et al., 2017) to train acoustic models from scratch as well as adapt pretrained Russian and English models to our data. With respect to the first chapter-specific research question, we found that the inclusion of code-switched speech and Russian

pretrained models aided alignments of Urum (Section 6.5). To answer the second chapter-specific research question, we tested the impact of acoustic modeling strategies in a bilingual phonetics investigation (Section 6.6): Are vowels in Urum words pronounced differently in monolingual Urum utterances than in code-switched utterances? After discussion, we conclude with methodological recommendations and areas for future work (Section 6.7). All code for replicating this work is publicly available.[3]

6.3 Urum language and dataset

Urum (ISO: uum) is a Turkic language spoken by ethnic Greeks in the Small Caucasus of Georgia and in Ukraine. Also known as Caucasian Urum, it is a variety of Anatolian Turkish that is classified as endangered (Campbell et al., 2022). For the variety documented by Skopeteas et al. (2022) and analyzed in this chapter, the language has been strongly influenced by Russian since the group arrived in Georgia in the early 19th century. Notably, most Urum speakers are bilingual in Russian and often code-switch between the two languages (Skopeteas, 2014). Unlike the examples of code-switching being used in purely meta-linguistic commentary (see Section 2.1.3), Russian portions of speech in this dataset were part of the narrative content by the speakers. The following shows an example of an Urum–Russian utterance with transliterated Russian displayed in brackets:

> äp halhmz egler kissäya **[muzka]** ederh **[malade]** **[tantsuet]** oinamah etmäh
>
> *"All the people get together at the church, we organise **[music]**, and the **[youth]** is **[dancing]**."* (Skopeteas et al., 2022)

We utilized the Urum dataset from the DoReCo corpus, which is a field data repository that contains manual word-level and automatic phone-level alignments of speech (Paschen et al., 2020). Traditional and personal Urum narratives were recorded across 30 speakers (16 female, 14 male) and spanned 117 minutes[4] of speech (Skopeteas et al., 2022). Figure 6.1 visualizes the distribution of Urum-only, Russian-only, and code-switched utterances among speakers. All but one speaker

[3] All code is publicly available at https://github.com/emilyahn/align_cs.

[4] Time was calculated by summing utterance durations, not file or word durations.

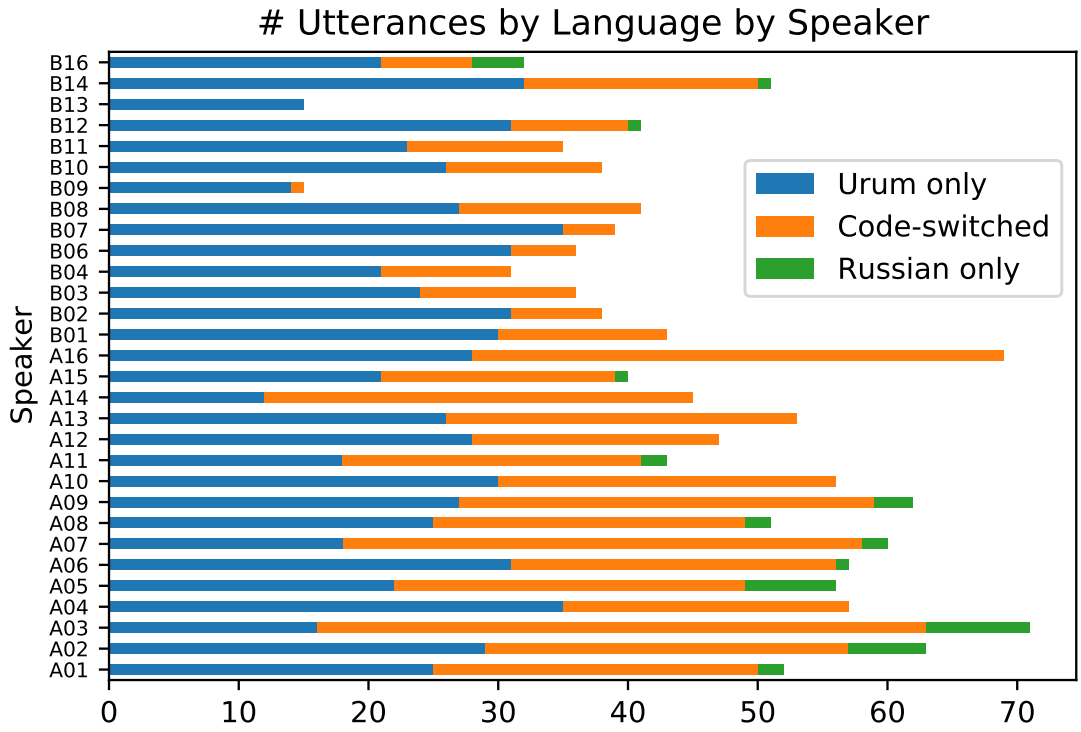

Figure 6.1: Across the 30 speakers in the DoReCo Urum field repository (Skopeteas et al., 2022), almost all produced code-switched utterances (orange, middle) in addition to monolingual Urum (blue, left) and monolingual Russian utterances (green, right).

code-switched. Table 6.1 reveals that while 42% of the utterances were code-switched, they represented 53% of the repository in minutes.[5] Code-switched utterances averaged 6.5 seconds, which was on average longer than non-codeswitched utterances (Urum-only: 4.5 seconds; Russian-only: 2 seconds). Among the code-switched utterances, Urum word tokens were more frequent than Russian word tokens, as seen in Figure 6.2.

[5]If all utterances with "foreign material" were excluded, as was the protocol in Zhu et al. (2024) over the full DoReCo dataset, we would miss out on half the data.

Utt	Count	Time (min)	Avg (sec)
All	1373	117.6	5.1
Urum	752 (55%)	53.5 (45%)	4.3
CS	581 (42%)	62.9 (53%)	6.5
Russ	40 (3%)	1.3 (1%)	2.0

Table 6.1: Distribution of utterances across language usage, by count and time. Notably, code-switched (CS) utterances had longer durations.

6.4 Methodology

6.4.1 Data preparation

Data from the field repository needed to be prepared as input to the acoustic models. The audio files were in WAV format, recorded with a sampling rate of 44.1 kHz, and transcriptions were provided in ELAN. We used Praat (Boersma and Weenink, 2019) to segment the audio files into utterances. Four utterances were filtered out due to encoding issues.

Urum phone sequences were derived automatically by the repository contributors, so our lexicons (two-column text files with words and their corresponding phone sequences) were gathered from these existing phone sequences. Most of the Russian words had been transliterated into Latin script at the word-level, so we used a simple mapping script to build the lexicon (see Table 6.2). The Urum phone set from DoReCo included nine vowels and 30 consonants while the transliterated Russian phone set included six vowels and 24 consonants. Only four Russian phones did not exist in the Urum set, as seen in Table 6.3, and we used the PanPhon tool (Mortensen et al., 2016) to map them to their nearest neighboring Urum phones in the lexicons: ɨ → ɯ, tɕ → tʃ, ʂ → ʃ, ʐ → ʒ. The repository contributors used tags to transcribe content such as filled pauses, prolongations, and false starts.[6] When a tagged word was partially transcribed (such as in this example of a false start, "<fs>ba"), we manually assigned it a phone sequence (in this case, "[b a]") and classified it as an Urum word.[7]

[6]The full list of possible tags is the following: filled pause, false start, prolongation, foreign material, singing, backchannel, ideophone, onomatopoeic, word-internal pause, unidentifiable, and silent pause.

[7]Anecdotally, there were several instances where the partially-tagged word was Russian. For the sake of simplicity, we labeled all tagged content as Urum.

Orthography	IPA
a	a
b	b
d	d
e	e
f	f
g	ɡ
h	x
i	i
j	ʑ
k	k
l	ɬ
m	m
n	n
o	o
p	p
r	r
s	s
t	t
u	u
v	v
y	j
z	z
č	tɕ
ɪ	ɨ
š	ɕ
ž	ʑ
’	ʲ

Table 6.2: We mapped the Russian orthography, which was transcribed in Latin script by the DoReCo repository contributors, to Russian IPA phones.

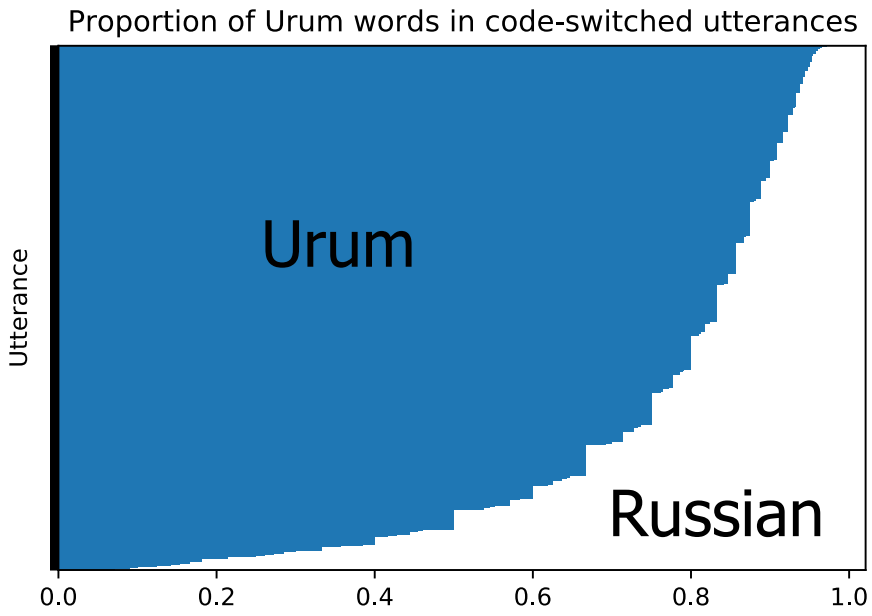

Figure 6.2: Proportion of Urum (blue, shaded) to Russian (white) word tokens in all 581 code-switched utterances, sorted highest to lowest. The majority of these utterances had more Urum than Russian tokens.

6.4.2 Acoustic modeling

We used the Montreal Forced Aligner (MFA version 2.2.17; McAuliffe et al., 2017) to train acoustic models and conduct forced alignment on our data in its default unsupervised manner. We split the DoReCo files into train and test partitions following the same split as Chodroff et al. (2024a): 1,097 utterances (100 minutes) in the train set and 273 utterances (16 minutes) in the test set. For this study, we created further subsets of the training data to answer our first chapter-specific research question. First, we summed the minutes of utterances of each language type and found 47 minutes of monolingual Urum utterances and 52 minutes of code-switched utterances. To keep the quantity of Urum-only and code-switched training data the same, we reduced the number of code-switched utterances to sum to 47 minutes, which would equal the amount of Urum-only speech. Our first results compared the alignment performance of a model trained on 47 minutes of purely

Urum-only	Urum & Russian	Russian-only
ɣ, æ, œ, ɯ	a, e, i, o, u	ɨ
ɟ, c, dː, tː	b, p, d, t, ɡ, k	
sː, ʃ, ʒ, ɣ, dʒ, tʃ	v, f, z, s, x	tɕ, ʂ, ʐ
l, lː, ɾ, mː	j, r, ɬ, m, n	

Table 6.3: The phone sets present in the DoReCo transcriptions across Urum and Russian, with the middle column being their overlap.

Urum speech to a model trained on 47 minutes of Urum–Russian speech.[8] Our third training data partition combined both sets to include 94 minutes of Urum-only and code-switched speech. All Russian-only utterances were excluded from training and evaluation.

Since we showed in Chapter 5 that it was advantageous to use larger, pretrained models for aligning low-resource languages, we chose two relevant MFA models to continue the experiments. The Russian MFA v3.1.0 model was trained on over 400 hours of data from over 3,000 speakers; this model was selected since Russian was frequently spoken in our dataset (McAuliffe and Sonderegger, 2024). The Global English MFA v2.2.1 model was trained on over 3,700 hours of data from over 79,000 speakers across the world (McAuliffe and Sonderegger, 2023). This model has previously proven to be effective in aligning the Urum dataset in Chodroff et al. (2024a). For cross-language modeling and alignment, we developed the lexicons by applying the PanPhon tool (Mortensen et al., 2016) for determining the nearest neighboring phones in cases where the target phone did not exist in the model. Table 6.4 displays these phone mappings. Each pretrained model was adapted to the same three training data partitions as described in the train-from-scratch settings above.

6.4.3 Forced alignment evaluation

Following Chodroff et al. (2024a), we evaluated the quality of the forced alignments using two metrics, precision and accuracy, to compare our system output to the hand-corrected, "gold" phone-alignments from the same study. For precision, we calculated the percent for which the model

[8]While the minutes across the two partitions were the same, the number of utterances was 618 for Urum and 414 for code-switched speech. However, the number of phones in each partition was roughly 27,000.

Russ (CS) to Eng MFA	Urum to Eng MFA	Urum to Russ MFA
tɕ → tʃ	dː → d	ɾ → r
ɨ → ɯ	lː → l	œ→ ɛ
ʂ → ʃ	mː → m	ɯ → ɨ
ʐ → ʒ	r → ɾ	ʃ → ʂ
	sː → s	ʒ → ʐ
	tː → t	d → d̪
	x → ç	dː → d̪ː
	y → ʉ	dʒ → dʐ
	œ→ ɛ	l → ɬ
	ɣ → ç	lː → ɬː
	ɯ → ə	n → n̪
		s → s̪
		sː → s̪ː
		t → t̪
		tː → t̪ː
		tʃ → tʂ
		y → ʉ
		z → z̪

Table 6.4: Urum and Russian (code-switched) phones from DoReCo that did not exist in the pre-trained English or Russian MFA model lexicons were mapped to their nearest neighboring phones, calculated with the PanPhon tool (Mortensen et al., 2016).

onset boundary was within 20 ms (the selected threshold) of the manually aligned onset boundary (MacKenzie and Turton, 2020; McAuliffe et al., 2017). For accuracy, we utilized a measure that calculated the proportion of model-aligned intervals whose midpoints lay within the respective gold intervals (Knowles et al., 2018; Mahr et al., 2021). All evaluation was conducted on the test partition which consisted of 132 Urum utterances and 119 code-switched utterances. Phones from only Urum words were evaluated, ignoring all phones from Russian words.

6.4.4 Analysis

We conducted mixed-effects regressions in R (R Core Team, 2022) using the lme4 package to analyze the variables that contributed to both the precision and the accuracy variables (Kuznetsova et al., 2017). Main effects were the language of the test utterance (Urum or CS), the natural class of the current phone, the natural class of the previous phone, the interaction of these two natural

classes, the proportion of contaminated (tagged) tokens,[9] the utterance duration, the interactions of model configuration with utterance language, and whether or not the speaker of the test utterance was present in the training set. Random effects were the speaker ID and the file ID of the utterances. The current phone class was sum-coded with the held-out level of stop; the previous phone class was sum-coded with the held-out level of silence. The eight classes analyzed were vowels, approximants, taps/trills, nasals, fricatives, affricates, and stops. Models were treatment-coded, each compared to the 47min train-from-scratch Urum-only model. We ran two models: the first was a linear model with the dependent variable as log seconds of onset boundary differences, with 0 sec mapped to 0.001 prior to the log transformation. The second model was a logistic regression with the binary dependent variable being accuracy.

6.5 Results

6.5.1 Alignment precision and accuracy

The following results answer our first chapter-specific research question: Does including Russian code-switched data in acoustic model training help the alignment of target Urum data? The different acoustic model configurations were trained or adapted on subsets of the DoReCo dataset, and they were all tested on the held-out test utterances that included both Urum-only and CS utterances. In the scenario where we trained MFA models from scratch (i.e., no pretrained model was used – note the None column in Table 6.5), we have two findings. When we kept the training data quantity equal at 47 minutes for both Urum-only speech and code-switched speech, the Urum-only model outperformed the purely code-switched one. However, combining these two training sets in the Urum + CS (94m) model substantially improved upon either 47 minute model. Since we only evaluated on phones from Urum words, it was expected that the Urum (47m) model would outperform the CS (47m) model as it was more matched in language setting. It also made sense

[9] Contamination in an utterance was calculated as the number of tagged tokens, such as false starts or prolongations, divided by total number of tokens.

that the combined training set yielded a higher performance, since it included more Urum speech overall.

Train/Adapt Partition	Pretrained model None	Eng	Russ
Precision % ↑			
Urum (47m)	63.2	70.4	71.2
CS (47m)	58.2	70.0	70.4
Urum + CS (94m)	70.9	70.6	**71.3**
Accuracy % ↑			
Urum (47m)	80.6	83.7	84.9
CS (47m)	77.2	83.1	84.4
Urum + CS (94m)	**85.1**	83.6	**85.1**

Table 6.5: Results revealed that the Russian MFA model adapted on all 94 minutes of Urum and code-switched (CS) data performed the best, with maximal training-from-scratch (i.e., Urum + CS (94m)) on par in terms of accuracy. Precision is how often the system phone boundary was within 20ms of the gold boundary. Accuracy is how often a system phone midpoint lay within the gold interval. Highest scores are bolded and shaded.

For the experiments using pretrained models adapted on the various Urum/CS partitions, the Russian MFA model adapted on Urum + CS (94m) produced the best results. Even though the Global English MFA model was trained on nearly 4,000 hours of diverse speech, its alignments did not outperform the smaller Russian MFA model. This is perhaps due to the language similarity of Russian to Urum, or the history of Urum being influenced by Russian contact. All models trained or adapted on the different DoReCo subsets patterned the same where the ranking of best to worst subset was Urum + CS (94m) > Urum (47m) > CS (47m), with the slight exception of accuracy from the Global English MFA with Urum (47m) > Urum + CS (94m).

6.5.2 Regression analysis

The mixed-effects regression analysis revealed several factors that influenced alignment performance. We report all significant findings of $p < 0.05$. Except for the train-from-scratch CS (47m) model which performed significantly worse, all other models performed significantly better than the Urum (47m) model. Longer utterance durations and higher contamination amounts were correlated with worse performance. The speaker appearing in the training data had no significant

	Spkr	VOWELS a	e	i	o	u	y	œ	æ	ɯ
Male	A01									
	A03	n=189			n=57					
Female	A02									
	A07	n=13								
	B08									
	B11									
	B16	n=20								

Table 6.6: Our case study revealed that from the gold data, /a/ for 3 speakers and /o/ for 1 speaker (marked with shaded cells and token counts) in Urum words were pronounced significantly differently in monolingual Urum vs code-switched utterances. For these four instances, Pillai scores indicated that the vowel formants for the two groups in question (Urum vs CS) were significantly non-overlapping.

effect. The language of the test utterance also had no effect, with a slight exception of the CS (47m) model performing slightly worse on Urum-only test utterances.

In terms of precision, boundaries around taps/trills were displaced more significantly, while boundaries around fricatives showed higher precision. Boundaries preceding vowels also performed better. Significantly better precision was found for vowel–tap/trill, fricative–vowel, affricate–vowel, affricate–nasal, stop–vowel, and stop–tap/trill sequences. Significantly worse precision was found for vowel–vowel, vowel–approximant, tap/trill–vowel, nasal–nasal, and fricative–approximant sequences.

As for accuracy, which used a logistic mixed-effects regression model, significantly better performance was found for phone intervals preceded by nasals, fricatives, affricates, and stops, as well as for targeted phone intervals that were fricatives and affricates. Significantly worse accuracy was found for phone intervals preceded by vowels, approximants, and taps/trills, as well as targeted phone intervals of these three classes. These results are largely comparable to the mixed-effects regression results from Chodroff et al. (2024a).

6.6 Case study

Following Babinski et al. (2019), we asked a general phonetics question and observed whether there were significant differences between the outputs of the different model configurations above. In other words, to what degree are we comfortable substituting an automatic alignment for manual alignment in answering a question about code-switching phonetics? We investigated the following: Are vowels in Urum words pronounced differently in monolingual Urum utterances compared to in CS utterances? First, we answered this with the manually-annotated "gold" test data.

6.6.1 Methodology

The Pillai-Bartlett trace, or Pillai score, is a useful metric to measure overlap in vowel category qualities. It takes output from a MANOVA test, which is used for measuring overlap between two distributions across two dependent variables–in our case, the first two formant values. Among four commonly used metrics for vowel overlap, Kelley and Tucker (2020) showed that Pillai scores are among the most reliable. Stanley and Sneller (2023) additionally provided a formula to derive a threshold for determining overlap vs separation, based on the exact sample size of the tokens. We followed these recommendations and calculated Pillai scores for formant values extracted from the gold test set. Formants were first extracted with the Linear Predictive Coding (LPC) tool in Praat (Boersma and Weenink, 2022), searching for 5 formants under 5,000 Hz for reported male speakers and 5,500 Hz for reported female speakers. The formant value analyzed per vowel was an average of values taken at three timestamps: the interval midpoint and 10 milliseconds before and after the midpoint.

6.6.2 Results from manual alignments

The gold test data revealed several instances of within-speaker differences in pronouncing certain Urum vowels. Table 6.6 shows four instances of a particular vowel being marked as significantly non-overlapping across two conditions. For example, the cell for male speaker A03 /a/ was marked with n=189. This meant that A03 uttered 189 /a/ vowels, and his F1×F2 values for /a/ in Urum words from monolingual Urum utterances were significantly different than his F1×F2 values for

Spkr	a	e	i	o	u	y	œ	æ	ɯ
A01	X								
A03	X		X	X					
A02									
A07	X								
B08								X	
B11									
B16									

Table 6.7: The **best-performing** acoustic model (Russian MFA adapted on Urum + CS 94m) yielded 3 true positives (shaded X), 3 false positives (unshaded X), and 1 false negative (shaded empty cell).

Spkr	a	e	i	o	u	y	œ	æ	ɯ
A01	X								
A03			X	X				X	X
A02									
A07	X								
B08									
B11									
B16									

Table 6.8: The **worst-performing** acoustic model (trained on the CS 47m partition) yielded 2 true positives (shaded X), 4 false positives (unshaded X), and 2 false negatives (shaded empty cells).

/a/ in Urum words from code-switched utterances. The same can be said for speaker A03's /o/ (n=57), speaker A07's /a/ (n=13), and speaker B16's /a/ (n=20).

6.6.3 Results from automatic alignments

Second, we calculated Pillai scores from the output of the best-performing and worst-performing models and compared these to the gold scores (Tables 6.7-6.8). From the best-performing model, the Russian MFA model adapted on the Urum + CS (94m) data, it found 6 instances of significant non-overlap. 3 out of the 4 gold instances were correctly identified (i.e., 3 true positives and 1 false negative), while producing 3 spurious significant findings (i.e., 3 false positives). From the worst-performing model, trained on the CS (47m) partition, it produced less congruent findings: only 2 out of the 4 gold instances were correctly identified (i.e., 2 true positives and 2 false negatives), with 4 spurious significant findings (i.e., 4 false positives). We used the phonR package in R (McCloy,

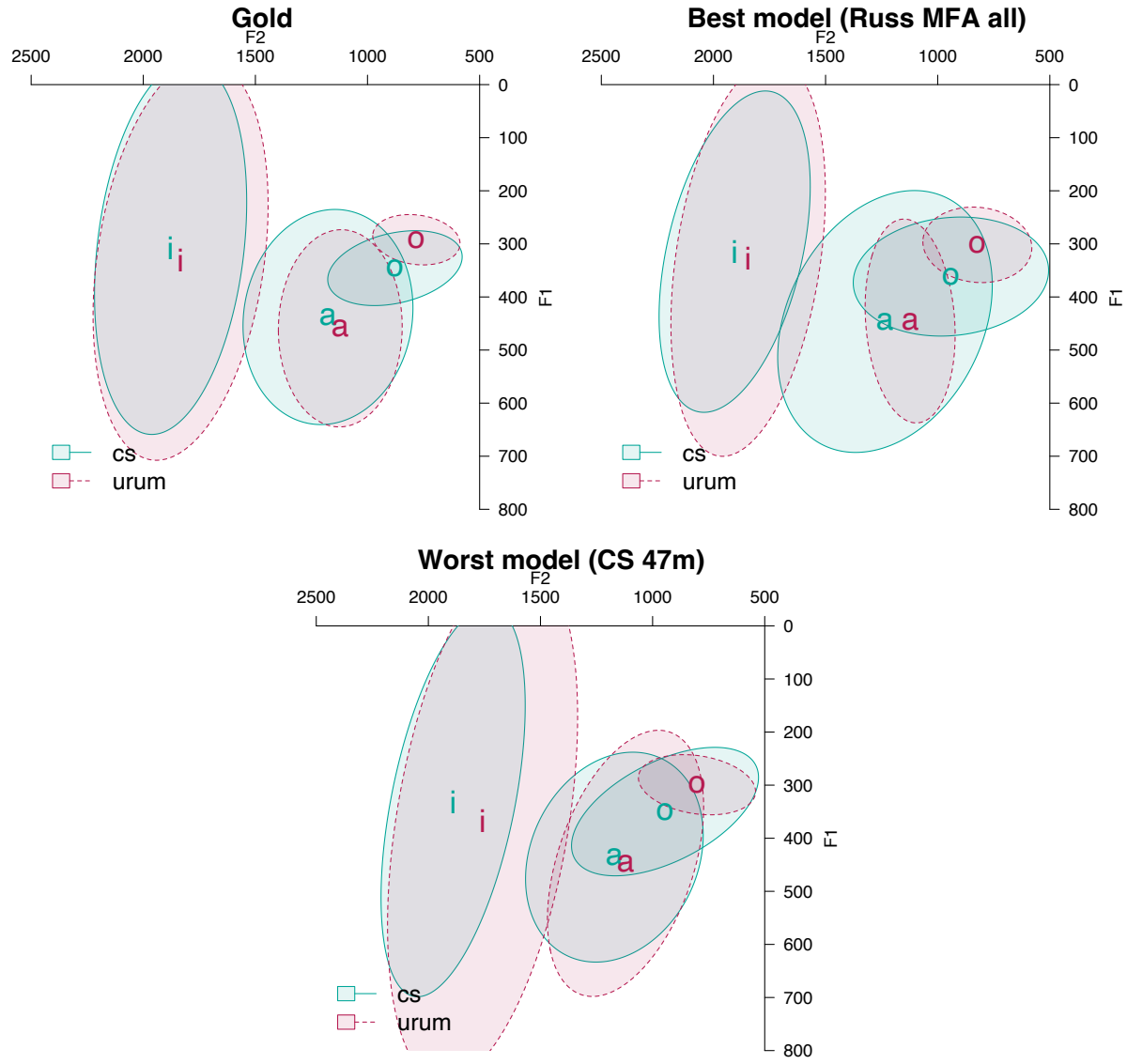

Figure 6.3: These plots reflect the first two formants (in Hz) for three of the nine Urum vowels, /a, i, o/, for male speaker A03. Clockwise from the top-left are formants extracted from the gold alignments, the best model (Russian MFA adapted on Urum + CS 94m) output, and the worst model (CS 47m) output. Vowel labels are positioned at means, and ellipses cover one standard deviation away from the mean.

2012) to plot vowel ellipses for /i, a, o/ for male speaker A03, over the two language conditions, and across the three types of output (Figure 6.3). The gold plot reflects our findings that /a/ and /o/ were significantly different between Urum and CS environments while /i/ was not. The best and worst model, however, show the ellipses increasingly diverging compared to the gold. Both models found spurious differences for /i/, and although /a/ visually appears significantly different for the worst model, it was a false negative.

Essentially, running our vowel overlap analysis on the alignments from the automated systems did not yield the same findings as from the gold alignments. Output from the best- and worst-performing models tended to hallucinate more vowel disparities than were found for the gold output. However, we observed that the best model's vowel disparity predictions more closely aligned to the true findings than the worst model's. The difference between these two models' alignment performances was 11% for precision and 7% for accuracy, but these percentages were uninterpretable. This case study helped interpret the broad metrics to reveal the nuances of alignment performance, as the downstream output yielded different conclusions.

6.7 Conclusion

This work tested methodologies of incorporating code-switched data in acoustic model training and alignment in a low-resource, field data scenario. We tested the inclusion of Urum–Russian code-switched utterances in training acoustic models to align Urum phones, and found that it was helpful to keep the code-switching to produce a larger train set. The maximally trained-from-scratch model performed roughly on-par with a pretrained Russian model adapted to the same field data. If one is fortunate to have 90-some minutes of transcribed data, that is plenty to train models from scratch. Otherwise, utilizing a large, pretrained model performed reasonably when adapted on target data.

In order to functionally assess the quality of the systems, we tested our best and worst systems' alignment outputs against the gold alignments to answer a bilingual phonetics question. Calculating Pillai scores across formant values for individual speakers, we discovered that certain Urum vowels for several speakers were pronounced significantly differently in monolingual Urum utterances than in code-switched utterances. Although the best acoustic model yielded more similar

results to the manually fixed gold alignments than the worst acoustic model, none of the automatic alignments were suitable for reaching conclusions for this particular research question. We recommend manual adjustment of phone boundaries when conducting phonetic analyses whose measurements are highly influenced by segment durations.

As future work, it would be beneficial to this research community to conduct a survey study with qualitative and quantitative statistics on the prevalence of code-switching across field data repositories. How are multiple languages used by the elicitors and by the community members of the language being documented?

Further research could also aim to extend the study of phonetics and phonology for code-switched language more broadly. Our case study only scratched the surface to discover the nature of shifting Urum vowel qualities depending on the languages present in that utterance. It would be interesting to discover if the significantly different Urum vowel formants were becoming more Russian-like when surrounded by Russian context, similar to findings on Korean–English by Seo and Olson (2024). Cross-linguistic interference or transfer could be in effect and is worth investigating.

Limitations

When conducting our regression analyses or case study, we did not take into account code-switching properties at the syntactic or prosodic level. It would be interesting to factor into account whether the code-switched utterance was inter-sentential or intra-sentential (i.e., mixing languages at phrase boundaries or within phrases). When calculating boundary differences, examining how close an Urum word was to a Russian word could have provided useful information. Prosodic factors such as speech rate and pitch would also add insight as, anecdotally, prosody was at times visibly different near the language switch points.

The Urum dataset from the DoReCo repository used in this work was particularly well-annotated for both Urum and Russian. However, the quality and quantity of transcriptions here may not represent other field data repositories, and replication of our findings on other datasets may be challenging.

Ethics statement

The dataset in this study has been made publicly available for download and research use. Speech data that is public carries inherent potential harms for misuse in downstream tasks.

Particularly for our methodological approach of including code-switched speech or the language of broader communication in the analysis of field data, we advise some caution. The speech from the non-target language may have been meant to be ignored and not recorded. If sections of the speech data were not explicitly transcribed, they may not have been intended to be used for analysis.

Postlude

This chapter ties together methods and themes from the previous three contribution chapters. We applied the corpus phonetics pipeline to field data and expanded our use of diverse data to include code-switched speech. With respect to RQ1, it is viable to include code-switched utterances, rather than ignore them, for the acoustic model training and forced alignment of targeted low-resource speech data. It is also beneficial to use a pretrained model from the language of broader communication (Russian in this study) for the acoustic modeling of the target language (Urum in this study). With respect to RQ2, we found that relying on automated processes alone would have degraded the precision of a manually-conducted phonetic analysis. A corpus phonetics analysis using phonetic measures taken from *automatic* alignments did not produce the same analysis results as using manually-annotated alignments. We further discuss the theme of trusting automation in the midst of data diversity in Section 7.2.1.

Part III

Resolution

CHAPTER 7
Conclusion

We first summarize what we learned and our contributions as they relate to the original research questions from Section 1.2. Then we connect the takeaways from this dissertation to broader themes across linguistics and speech technology. This discussion includes avenues for future direction.

7.1 Summary

RQ1 Methodology: What methods are viable for processing diverse speech corpora within the corpus phonetics pipeline?

Each contribution chapter used automated systems to process diverse speech corpora. Chapters 3 and 4 applied grapheme-to-phoneme (G2P) systems, forced alignment, and formant measurement tools to process multilingual read speech datasets. We learned that these tools and pipeline are usable for studying phonetic typology at a broad level. From an audit of outlying vowel formants, we also learned that errors can occur at various stages of the pipeline. Factors including a language's sound system, speaker idiosyncrasies, dataset quality, and choice of automated pipeline systems can all impact the data results. In Chapters 5 and 6, we focused on field data from language documentation settings and specifically manipulated the G2P output and the acoustic modeling to improve forced alignment. Generally successful was the strategy of cross-language acoustic modeling using a large, pretrained model to adapt and align a small, targeted field dataset. Additional strategies were only somewhat useful, including the broadening of phone classes for the Panãra dataset and the inclusion of Russian code-switched speech in the Urum dataset.

RQ2 Impact: How much can one rely on automated processes to obtain interpretable phonetics answers?

Chapter 4 demonstrated, through an audit of outliers in a pipeline of processed read speech, that errors and variation could occur at many levels with varying impact on the downstream output. At the highest level, transcripts that did not match the audio or were retrieved incorrectly (e.g., by sentence-level speech-to-audio alignment) caused the biggest issues. Alignment errors and formant tracking errors were also problematic. When the data quality was lower, these transcript and alignment errors were more abundant; when the data quality was higher, outliers were more attributed to smaller issues of formant tracking errors and linguistic variation. In the case of low data quality, we found it unreliable to use automated processes in the corpus phonetics pipeline.

In our investigation of acoustic modeling of code-switched Urum–Russian in Chapter 6, we answered this impact question directly. We calculated Pillai scores on formants from Urum vowels spoken in monolingual Urum utterances vs in code-switched Urum–Russian utterances, predicting that Urum vowels in code-switched contexts would be realized significantly differently from monolingual contexts. We answered this question with manually-annotated "gold" alignments, and found that the scores and conclusions derived from our automatically aligned segments were different from the gold version. While our best acoustic model yielded closer conclusions to the gold version than the worst acoustic model, none of these outputs were suitable for reaching conclusions on this particular research question.

7.2 Broad themes and future directions

7.2.1 Trusting automation in the face of diversity

With respect to RQ2 above, perhaps we erred more conservatively in saying that output from automated corpus phonetics pipelines is unreliable. Automation always introduces some levels of error, and the question is more so—what is our threshold of acceptability? To arrive at an acceptable state, where and how much might we need to intercede in the pipeline with any manual edits? For the Panãra and Urum field datasets, we obtained gold phone segmentations by first

utilizing G2P and forced alignment, then manually editing the boundaries in Praat. This process is commonly used in phonetics research and is an effective use of automation combined with human intervention. However, there is a tradeoff of resources between doing manual edits and being able to process data more quickly. One of the aspects of diverse data is that there is either not much available data or not many transcriptions or supervised labels for that data. If we are building technology for a low-resource language variety, it may be worth adding more manual edits in the pipeline as there is not much data to begin with.[1]

Another aspect of diverse data is that technology often treats it differently than standard data. We encourage researchers to carefully examine the data and discern what the models treat as accurate and inaccurate. As we discussed in Chapter 4, non-standard speech, or speech that was unaccounted for in the assumptions in the pipeline, might be classified as outlying or erroneous. It is important to distinguish between valid variation outliers—because the data reflects diverse people—and outliers that resulted from technical processing errors. In work that examined the many data partitions to form a large English language model (C4), Dodge et al. (2021) found that pre-processing algorithms frequently removed data partitions from low-resource varieties (e.g., text in African American English) and about minority individuals (e.g., text discussing LGBTQ+ identities). This falls in line with growing research on the discovery that data processing pipelines for large models can exacerbate existing power inequalities in society (Bender et al., 2021).

Therefore, while we embrace automation, it is still important to be cautious with it. We encourage future research to focus on developing technology to work with non-standard speech data, especially types we have not investigated in this dissertation such as tonal language varieties, child speech, and speech disorders. In doing so, we invite users of this automation, whether corpus phoneticians or speech technology developers, to consider the benefits of manual audits. With regards to the idea of a "robot phonetician" or a phonetician's "robot assistant" (Liberman, 2019), we encourage the latter concept in order to prevent technology (without human intervention) from doing harm such as removing diverse data partitions in Dodge et al. (2021). A key to trusting automation in these scenarios is knowing that a human has approved the work of the "robot assistant."

[1]Conversely, it could be harder to obtain manual edits if there are not as many speakers of that variety to verify the work. It can be costly for field linguists to consult experts of the language for checking transcriptions.

Lastly, if we aim to celebrate the diversity of people and cultures that these languages represent, it is extremely important for linguistics research (and therefore in our case, computational linguistics research) to align with the goals and needs of the communities and people that the research concerns. Work by Levow et al. (2017) shows how dialogue between language community members and researchers can inform directions for future technologies and language applications to support language documentation (for a detailed overview of possible directions, see also van Esch et al., 2019).

7.2.2 Representing sound units

Another theme that this dissertation touches on is: What is a good representation of a speech unit that is faithful to the acoustics yet generalizable (across that language variety, across speakers, or across languages)? This question is important for both speech technology and linguistics. In line with what we classified as a Linguistic Variation in Chapter 4, we discovered an unanswered question of what we should do when the acoustic-phonetic realization of a sound does not match the transcription because of valid linguistic variation. This could occur for a number of reasons including: (1) the transcript was at the phonemic level but the spoken sound was a different allophone of that phoneme; (2) the speech was reduced (e.g., the speaker said "wanna" instead of "want to"); (3) there are dialectal or idiosyncratic variations in the speakers; etc. What would happen if the phone transcription was more specific, allophonic, and faithful to the acoustics? Perhaps it would perform better on the data, but that phone set may lose generalization to newer or different data, as there is a risk of over-fitting to that trained dataset or variety.

For linguistics research, the phonetics-phonology interface has been a delicate balance of acknowledging that phonology is not strictly discrete and that phonetics is not strictly continuous (Cohn, 2007). In Chapter 5, we broadened phone categories, collapsing several sounds that were similar in place and manner of articulation, and the results were mixed. We believe that broadening sound classes achieved small performance gains in cases where there was very little training data per phone. Once more data was present, it was better to retain specificity (e.g., that /o/ and /u/ are distinct phones). This approach had the appeal of being generalizable across languages,

which is applicable for cross-language or multilingual acoustic modeling. However, it was clear that modeling specificity was important as well. Perhaps data quantity could help discern where in the spectrum of phonetics we can afford to be discrete. In other words, we can broaden categories if the data is limited, but then make it more specific when the data is more plentiful.

An avenue for future work related to sound representations is to pursue phoneme recognition. Phoneme recognition tools have a potential for aiding the transcription process in language documentation, especially for oral languages that do not have writing systems (Bird, 2021). It would be interesting to analyze if phoneme recognition tools could also help field linguists in determining the phone inventory of their language varieties of interest. As we observed in Section 4.6.2, linguists on PHOIBLE did not agree on the vowel inventory for Hausa (Moran and McCloy, 2019), and the formant data indicated that vowels had differing realizations from the given labels. In general, phoneme inventories vary widely for many languages due to both systematic and unpredictable reasons (Anderson et al., 2023). Perhaps an automated system could provide alternative inventory suggestions during the language documentation process, as well as offer systematic choices when defining phoneme inventories cross-linguistically.

As computer scientists continue to innovate on speech technologies such as acoustic modeling, forced alignment, and phoneme recognition, it will be important to consider the differences between modeling acoustic contrasts vs. linguistic contrasts. Phoneme inventories are defined based on the mental representation of sounds and how they contrast, instead of on the pure acoustics. Because machine learning is by nature empirical, there may be an over-emphasis on modeling speech output functionally as opposed to the speech processes mentally. In the possible future scenario that phoneme recognition technology is "perfected," linguists will still need to make decisions on the accuracy of those phonemes based on speakers' perceptions and the appropriateness of the granularity of the recognized sounds.

7.2.3 Utilizing linguistics in language technologies

A goal of this dissertation has been to use language technologies to further scientific research on spoken languages. We might now ask the reverse question if there is a benefit that phonetics

research has on technology. With the rise of speech and language technologies and large systems that can perform at-par with humans on certain tasks, we believe there is still a need for linguistics to inform these systems.

Opitz et al. (2025) argued that linguistics contributes to language technologies by providing curated resources, evaluating and interpreting models, adding critical insights in low-resource settings, and allowing the technology to contribute to meaningful language science discovery rather than to purely commercial applications. In a perspective paper, Yang et al. (2025) emphasized the need for language technologies to have social awareness—that is, for systems to take social contexts into account. Their goal is for technology to interact responsibly with human society. Towards this goal, computer scientists need to collaborate with experts in other domains such as linguistics, psychology, and sociology.

Following these recommendations, there are several pathways for phonetics research and linguistics to improve speech technology. First, linguists can aid in curating speech corpora that it is high quality and covers diverse language varieties, speakers, speech styles, and recording environments. Second, linguists can make sense of the output that large models are producing and discover patterns from their output. If such patterns reflect human linguistic processes, this could contribute to scientific discovery, as we found in the outliers of vowel formants in Chapter 4. Third, linguists can provide creative methodologies to design speech technologies that follow linguistic theory, such as our phonetic granularity exploration in Chapter 5. We encourage research to continue along these lines and for linguistics and language technology to not only coexist but complement one another in a symbiotic relationship.

BIBLIOGRAPHY

(2023a). Kazakh Langauge. *Encyclopaedia Britannica*.

(2023b). Swedish Langauge. *Encyclopaedia Britannica*.

Ahn, E. P. and Chodroff, E. (2022). VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 5286--5294.

Ahn, E. P., Chodroff, E., Lapierre, M., and Levow, G.-A. (2024). The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panãra. In *Interspeech 2024*, pages 1505--1509.

Ahn, E. P., Levow, G.-A., Wright, R. A., and Chodroff, E. (2023). An Outlier Analysis of Vowel Formants from a Corpus Phonetics Pipeline. In *Interspeech 2023*, pages 2573--2577.

Anderson, C., Tresoldi, T., Greenhill, S. J., Forkel, R., Gray, R., and List, J.-M. (2023). Variation in Phoneme Inventories: Quantifying the Problem and Improving Comparability. *Journal of Language Evolution*.

Archive, S. P. (2019). *Hausa Sound Inventory (SPA)*. Max Planck Institute for the Science of Human History, Jena.

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 12th Language Resources and Evaluation Conference*, pages 11--16.

Ashby, L. F., Bartley, T. M., Clematide, S., Del Signore, L., Gibson, C., Gorman, K., Lee-Sikka, Y., Makarov, P., Malanoski, A., Miller, S., Ortiz, O., Raff, R., Sengupta, A., Seo, B., Spektor, Y., and Yan, W. (2021). Results of the Second SIGMORPHON Shared Task on Multilingual Grapheme-to-Phoneme Conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 115--125.

Babinski, S., Dockum, R., Craft, J. H., Fergus, A., Goldenberg, D., and Bowern, C. (2019). A Robin Hood Approach to Forced Alignment: English-trained Algorithms and their Use on Australian Languages. *Proceedings of the Linguistic Society of America*, 4:3:1--12.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449--12460. Curran Associates, Inc.

Bailey, G. (2016). Automatic Detection of Sociolinguistic Variation using Forced Alignment. In *University of Pennsylvania Working Papers in Linguistics: Selected Papers from New Ways of Analyzing Variation (NWAV 44)*, pages 10--20.

Barreda, S. (2021). Fast Track: Fast (Nearly) Automatic Formant-Tracking using Praat. *Linguistics Vanguard*, 7(1).

Becker-Kristal, R. (2010). *Acoustic Typology of Vowel Inventories and Dispersion Theory: Insights from a Large Cross-linguistic Corpus*. PhD thesis, University of California, Los Angeles.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610--623, New York, NY, USA. Association for Computing Machinery.

Bird, S. (2021). Sparse Transcription. *Computational Linguistics*, 46(4):713--744.

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.

Boersma, P. and Weenink, D. (2019). Praat: Doing Phonetics by Computer [Computer Program]. (Version 6.1.08).

Boersma, P. and Weenink, D. (2022). Praat: Doing Phonetics by Computer [Computer Program]. (Version 6.0.3).

Bullock, B. E. and Toribio, A. J. (2009). *The Cambridge Handbook of Linguistic Code-switching*. Cambridge University Press.

Campbell, L., Lee, N. H., Okura, E., Simpson, S., and Ueki, K. (2022). The Catalogue of Endangered Languages (ElCat). Database available at http://endangeredlanguages.com/userquery/download/, accessed 2022-08-28.

Chen, W.-R., Whalen, D., and Shadle, C. (2019). F0-induced Formant Measurement Errors Result in Biased Variabilities. *The Journal of the Acoustical Society of America*, 145:EL360–EL366.

Chodroff, E. (2018). Corpus Phonetics Tutorial. *arXiv preprint arXiv:1811.05553*.

Chodroff, E., Ahn, E. P., and Dolatian, H. (2024a). Comparing Language-specific and Cross-language Acoustic Models for Low-resource Phonetic Forced Alignment. *Language Documentation & Conservation*.

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.

Chodroff, E., Pažon, B., Baker, A., and Moran, S. (2024b). Phonetic Segmentation of the UCLA Phonetics Lab Archive. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12724--12733, Torino, Italia. ELRA and ICCL.

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*, 98(2):250--289.

Cohen Priva, U., Strand, E., Yang, S., Mizgerd, W., Creighton, A., Bai, J., Mathew, R., Shao, A., Schuster, J., and Wiepert, D. (2021). The Cross-linguistic Phonological Frequencies (XPF) Corpus.

Cohn, A. C. (2007). Phonetics in Phonology and Phonology in Phonetics. *Working Papers of the Cornell Phonetics Laboratory*, 16:1--31.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., and Bapna, A. (2023). FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798--805. IEEE.

Coto-Solano, R., Nicholas, S. A., and Wray, S. (2018). Development of Natural Language Processing Tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26--33, Dunedin, New Zealand.

Coto-Solano, R. and Solórzano, S. F. (2017). Comparison of Two Forced Alignment Systems for Aligning Bribri Speech. *CLEI ELECTRONIC JOURNAL*, 20(1).

Coto-Solano, R., Stanford, J. N., and Reddy, S. K. (2021). Advances in Completely Automated Vowel Analysis for Sociophonetics: Using End-to-End Speech Recognition Systems with DARLA. *Frontiers in Artificial Intelligence*, 4.

Deuchar, M., Davies, P., Herring, J., Couto, M. C. P., and Carter, D. (2014). Building Bilingual Corpora. *Advances in the Study of Bilingualism*, pages 93--111.

DiCanio, C., Nam, H., Whalen, D. H., Timothy Bunnell, H., Amith, J. D., and García, R. C. (2013). Using Automatic Alignment to Analyze Endangered Language Data: Testing the Viability of Untrained Alignment. *The Journal of the Acoustical Society of America*, 134(3):2235--2246.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (2021). Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-T., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286--1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eek, A. and Meister, E. (1994). Acoustics and Perception of Estonian Vowel Types. *Phonetic Experimental Research*, XVIII:146--158.

Evanini, K., Isard, S., and Liberman, M. (2009). Automatic Formant Extraction for Sociolinguistic Analysis of Large Corpora. In *Tenth Annual Conference of the International Speech Communication Association*.

Fernández, S., Graves, A., and Schmidhuber, J. (2008). Phoneme Recognition in TIMIT with BLSTM-CTC. *arXiv preprint arXiv:0804.3269*.

Fricke, M., Kroll, J. F., and Dussias, P. E. (2016). Phonetic Variation in Bilingual Speech: A Lens for Studying the Productioncomprehension Link. *Journal of Memory and Language*, 89:110--137.

Fromont, R. and Hay, J. (2012). LaBB-CAT: an Annotation Store. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 113--117.

Fruehwald, J. (2017). The Role of Phonology in Phonetic Change. *Annual Review of Linguistics*, 3:25--42.

Gao, H., Hasegawa-Johnson, M., and Yoo, C. D. (2024). G2PU: Grapheme-To-Phoneme Transducer with Speech Units. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10061--10065.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1. Technical Report 93, NASA/STI Recon.

Goldman, J.-P. (2011). EasyAlign: an Automatic Phonetic Alignment Tool under Praat. In *Interspeech 2011*, pages 3233--3236.

Gonzalez, S., Grama, J., and Travis, C. (2020). Comparing the Performance of Forced Aligners Used in Sociophonetic Research. *Linguistics Vanguard*, 5.

Gordon, M. K. (2016). *Phonological Typology*, volume 1. Oxford University Press.

Gorman, K., Howell, J., and Wagner, M. (2011). Prosodylab-Aligner: A Tool for Forced Alignment of Laboratory Speech. *Canadian Acoustics*, 39(3):192--193.

Gorman, K., McCarthy, A. D., Cotterell, R., Vylomova, E., Silfverberg, M., and Markowska, M. (2019). Weird Inflects but OK: Making Sense of Morphological Generation Errors. In *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, pages 140--151.

Guy, G. R. and Hinskens, F. (2016). Linguistic Coherence: Systems, Repertoires and Speech Communities. *Lingua*, 172(173):1--9.

Hammond, M. (2021). Data Augmentation for Low-Resource Grapheme-to-Phoneme Mapping. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 126--130.

Harper, M. (2011). The IARPA Babel Multilingual Speech Database. Accessed: 2020-05-01.

Hoffmann, S. and Pfister, B. (2013). Text-to-Speech Alignment of Long Recordings using Universal Phone Models. In *Interspeech 2013*, pages 1520--1524.

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.

Hu, K., Chen, Z., Yang, C.-H. H., Żelasko, P., Hrinchuk, O., Lavrukhin, V., Balam, J., and Ginsburg, B. (2025). Chain-of-thought Prompting for Speech Translation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1--5.

Jakobson, R., Fant, C. G., and Halle, M. (1951). Preliminaries to Speech Analysis: The Distinctive Features and their Correlates.

Johnson, K. A., Babel, M., Fong, I., and Yiu, N. (2020). SpiCE: A New Open-Access Corpus of Conversational Bilingual Speech in Cantonese and English. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4089--4095, Marseille, France. European Language Resources Association.

Jones, A. and Renwick, M. E. L. (2024). Evaluating Italian Vowel Variation with the Recurrent Neural Network Phonet. In *Interspeech 2024*, pages 3679--3683.

Jones, C., Li, W., Almeida, A., and German, A. (2019). Evaluating Cross-linguistic Forced Alignment of Conversational Data in North Australian Kriol, an Under-resourced Language. *Language Documentation & Conservation*, 13:281--299.

Kazakevich, O. and Klyachko, E. (2022). Evenki DoReCo Dataset. In Seifart, F., Paschen, L., and Stave, M., editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin and Lyon. Accessed on 17 Nov 2022.

Keating, P. A. (2003). Phonetic and Other Influences on Voicing Contrasts. In Solé, M.-J., Recasens, D., and Romero, J., editors, *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 20--23, Barcelona, Spain.

Kelley, M. C., Perry, S. J., and Tucker, B. V. (2024). The Mason-Alberta Phonetic Segmenter: A Forced Alignment System Based on Deep Neural Networks and Interpolation. *Phonetica*, 81(5):451--508.

Kelley, M. C. and Tucker, B. V. (2020). A Comparison of Four Vowel Overlap Measures. *The Journal of the Acoustical Society of America*, 147(1):137--145.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual Processing of Speech via Web Services. *Computer Speech & Language*, 45:326--347.

Knowles, T., Clayards, M., and Sonderegger, M. (2018). Examining Factors Influencing the Viability of Automatic Acoustic Analysis of Child Speech. *Journal of Speech, Language, and Hearing Research*, 61(10):2487--2501.

Krippendorff, K. (2018). *Content Analysis: An Introduction to its Methodology*. SAGE Publications.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82:1--26.

Labov, W., Rosenfelder, I., and Fruehwald, J. (2013). One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. *Language*, 89(1):30--65.

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.

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

Ladefoged, P. and Maddieson, I. (1996). Recording the Phonetic Structures of Endangered Languages. *UCLA Working Papers in Phonetics*, 93:1--7.

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

Lane, W. and Bird, S. (2021). Local Word Discovery for Interactive Transcription. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2058--2067.

Lapierre, M. (2023a). The Phonology of Panãra: A Prosodic Analysis. *International Journal of American Linguistics*, 89(3):333--356.

Lapierre, M. (2023b). The Phonology of Panãra: A Segmental Analysis. *International Journal of American Linguistics*, 89(2):183--218.

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.

Levow, G.-A., Ahn, E. P., and Bender, E. M. (2021). Developing a Shared Task for Speech Processing on Endangered Languages. In *Proceedings of the Workshop on Computational Methods for Endangered Languages (ComputEL)*, volume 1, pages 96--106.

Levow, G.-A., Bender, E. M., Littell, P., Howell, K., Chelliah, S., Crowgey, J., Garrette, D., Good, J., Hargus, S., Inman, D., Maxwell, M., Tjalve, M., and Xia, F. (2017). STREAMLInED Challenges: Aligning Research Interests with Shared Tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39--47.

Leys, C., Klein, O., Dominicy, Y., and Ley, C. (2018). Detecting Multivariate Outliers: Use a Robust Variant of the Mahalanobis Distance. *Journal of Experimental Social Psychology*, 74:150--156.

Li, X., Mortensen, D. R., Metze, F., and Black, A. W. (2021). Multilingual Phonetic Dataset for Low Resource Speech Recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6958--6962.

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

Liberman, M. Y. (2019). Corpus Phonetics. *Annual Review of Linguistics*, 5:91--107.

Liljencrants, J. and Lindblom, B. (1972). Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast. *Language*, 48(4):839--862.

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

Lindblom, B. and Maddieson, I. (1988). Phonetic Universals in Consonant Systems. In Hyman, L. M. and Li, C., 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.

List, J.-M. (2012). SCA: Phonetic Alignment Based on Sound Classes. In Slavkovik, M. and Lassiter, D., editors, *New Directions in Logic, Language, and Computation*, pages 32--51. Springer.

Littell, P., Joanis, E., Pine, A., Tessier, M., Huggins Daines, D., and Torkornoo, D. (2022). ReadAlong Studio: Practical Zero-Shot Text-Speech Alignment for Indigenous Language Audiobooks. In Melero, M., Sakti, S., and Soria, C., editors, *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 23--32, Marseille, France. European Language Resources Association.

Liu, Y., Yang, X., and Qu, D. (2024). Exploration of Whisper Fine-tuning Strategies for Low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Macaire, C., Schwab, D., Lecouteux, B., and Schang, E. (2022). Automatic Speech Recognition and Query By Example for Creole Languages Documentation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2512--2520, Dublin, Ireland. Association for Computational Linguistics.

MacKenzie, L. and Turton, D. (2020). Assessing the Accuracy of Existing Forced Alignment Software on Varieties of British English. *Linguistics Vanguard*, 6(s1):20180061.

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

Mahr, T. J., Berisha, V., Kawabata, K., Liss, J., and Hustad, K. C. (2021). Performance of Forced-Alignment Algorithms on Children's Speech. *Journal of Speech, Language, and Hearing Research*, 64(6S):2213--2222.

Makarov, P. and Clematide, S. (2018). Imitation Learning for Neural Morphological String Transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2877--2882.

Martín-Morató, I. and Mesaros, A. (2021). What is the Ground Truth? Reliability of Multi-Annotator Data for Audio Tagging. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 76--80. IEEE.

Mathad, V. C., Mahr, T. J., Scherer, N., Chapman, K., Hustad, K. C., Liss, J., and Berisha, V. (2021). The Impact of Forced-Alignment Errors on Automatic Pronunciation Evaluation. In *Interspeech 2021*, pages 1922--1926. ISCA.

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.

McAuliffe, M. and Sonderegger, M. (2023). English MFA Acoustic Model v2.2.1. Technical report, https://mfa-models.readthedocs.io/en/latest/acoustic/English/English%20MFA%20acoustic%20model%20v2_2_1.html.

McAuliffe, M. and Sonderegger, M. (2024). Russian MFA Acoustic Model v3.1.0. Technical report, https://mfa-models.readthedocs.io/en/latest/acoustic/Russian/Russian%20MFA%20acoustic%20model%20v3_1_0.html.

McCloy, D. R. (2012). Vowel Normalization and Plotting with the phonR Package. *Technical Reports of the UW Linguistic Phonetics Laboratory*, 1:1--8.

McCollum, A. G. and Chen, S. (2021). Kazakh. *Journal of the International Phonetic Association*, 51(2):276--298.

Meer, P. (2020). Automatic Alignment for New Englishes: Applying State-of-the-art Aligners to Trinidadian English. *The Journal of the Acoustical Society of America*, 147(4):2283--2294.

Ménard, L., Schwartz, J.-L., and Aubin, J. (2008). Invariance and Variability in the Production of the Height Feature in French Vowels. *Speech Communication*, 50:14--28.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and Guillaume, S. (2018). Integrating Automatic Transcription into the Language Documentation Workflow: Experiments with Na data and the Persephone Toolkit. *Language Documentation & Conservation*, 12:393--429.

Moran, S., Danner, T., de León, M. P., and Zollikofer, C. (2024). Interindividual Vocal Tract Diversity Influences the Phonetic Diversification of Spoken Languages. *bioRxiv*.

Moran, S. and McCloy, D., editors (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Mortensen, D. R., Dalmia, S., and Littell, P. (2018). Epitran: Precision G2P for Many Languages. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Paris, France. European Language Resources Association (ELRA).

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. (2016). PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475--3484.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press.

Newman, P. (2022). *A History of the Hausa Language: Reconstruction and Pathways to the Present*. Cambridge University Press, 1 edition.

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.

Oganyan, M., Levow, G.-A., Squizzero, R., Ahn, E. P., Ng, S., Deaton, E., and Wright, R. A. (2024). Investigating the Acoustic Fidelity of Vowels across Remote Recording Methods. *Linguistics Vanguard*, 10(1):63--79.

Opitz, J., Wein, S., and Schneider, N. (2025). Natural Language Processing Relies on Linguistics. *Computational Linguistics*, pages 1--23.

Oushiro, L. (2019). Linguistic Uniformity in the Speech of Brazilian Internal Migrants in a Dialect Contact Situation. In Calhoun, S., Escudero, P., Tabain, M., and Warren, P., editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 686--690, Melbourne, Australia. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Pandey, A., Gogoi, P., and Tang, K. (2020). Understanding Forced Alignment Errors in Hindi-English Code-Mixed Speech—a Feature Analysis. In *Proceedings of the First Workshop on Speech Technologies for Code-Switching in Multilingual Communities*, pages 13--17.

Paschen, L., Delafontaine, F., Draxler, C., Fuchs, S., Stave, M., and Seifart, F. (2020). Building a Time-aligned Cross-linguistic Reference Corpus from Language Documentation Data (DoReCo). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2657--2666, Marseille, France. European Language Resources Association.

Peri, R., Sadjadi, S. O., Garcia-Romero, D., Vishnubhotla, S., and Han, K. J. (2025). Knowledge Distillation from Ensemble for Spoken Language Identification. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1--5.

Pine, A., William Littell, P., Joanis, E., Huggins-Daines, D., Cox, C., Davis, F., Antonio Santos, E., Srikanth, S., Torkornoo, D., and Yu, S. (2022). G_i2P_i Rule-based, Index-preserving Grapheme-to-phoneme Transformations. In Moeller, S., Anastasopoulos, A., Arppe, A., Chaudhary, A., Harrigan, A., Holden, J., Lachler, J., Palmer, A., Rijhwani, S., and Schwartz, L., editors, *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52--60, Dublin, Ireland. Association for Computational Linguistics.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. *Speech Communication*, 45(1):89--95.

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.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. (2024). Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1--52.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757--2761.

Pricop, B. (2024). Massively Multilingual Speech Corpora as a Resource for Investigating Phonetic Universals: Consonant F0 Effects in Catalan. Master's thesis, University of Zurich.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust Speech Recognition via Large-scale Weak Supervision. In *International Conference on Machine Learning*, pages 28492--28518. PMLR.

Reddy, S. and Stanford, J. N. (2015). Toward Completely Automated Vowel Extraction: Introducing DARLA. *Linguistics Vanguard*, 1(1).

Riad, T. (2013). *The Phonology of Swedish*. Oxford University PressOxford, 1 edition.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite.

Rousso, R., Cohen, E., Keshet, J., and Chodroff, E. (2024). Tradition or Innovation: A Comparison of Modern ASR Methods for Forced Alignment. In *Interspeech 2024*, pages 1525--1529.

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. 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 *Interspeech 2021*, pages 3655--3659.

Samir, F., Ahn, E. P., Prakash, S., Shwartz, V., Sóskuthy, M., and Zhu, J. (2025). A Comparative Approach for Auditing Multilingual Phonetic Transcript Archives. *Transactions of the Association for Computational Linguistics*. Forthcoming.

San, N., Bartelds, M., Ògúnrèmí, T., Mount, A., Thompson, R., Higgins, M., Barker, R., Simpson, J., and Jurafsky, D. (2022). Automated Speech Tools for Helping Communities Process Restricted-access Corpora for Language Revival Efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 41--51.

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.

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.

Schwartz, J.-L. and Ménard, L. (2019). Structured Idiosyncrasies in Vowel Systems. *OSF Preprints*.

Seo, Y. and Olson, D. J. (2024). Phonetic Shifts in Bilingual Vowels: Evidence from Intersentential and Intrasentential Code-switching. *International Journal of Bilingualism*, 0(0):1--16.

Shi, J., Amith, J. D., Castillo García, R., Guadalupe Sierra, E., Duh, K., and Watanabe, S. (2021). Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yolóxochitl Mixtec. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134--1145, Online. Association for Computational Linguistics.

Skopeteas, S. (2014). Caucasian Urums and Urum Language. *Journal of Endangered Turkish Languages*, 3(1):333--364.

Skopeteas, S., Moisidi, V., Tsetereli, N., Lorenz, J., and Schröter, S. (2022). Urum DoReCo Dataset. In Seifart, F., Paschen, L., and Stave, M., editors, *Language Documentation Reference Corpus*

(DoReCo) 1.2. Leibniz-Zentrum Allgemeine Sprachwissenschaft & Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin and Lyon. Accessed on 17 Nov 2022.

Squizzero, R. and Wassink, A. B. (2022). A Comparison of Three Methods for Identifying Dynamic Formant Tracking Errors via Outlier Detection.

Stanley, J. A. and Sneller, B. (2023). Sample Size Matters in Calculating Pillai Scores. *The Journal of the Acoustical Society of America*, 153(1):54--67.

van Esch, D., Foley, B., and San, N. (2019). Future Directions in Technological Support for Language Documentation. In Arppe, A., Good, J., Hulden, M., Lachler, J., Palmer, A., Schwartz, L., and Silfverberg, M., editors, *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 14--22, Honolulu. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Wassink, A. B., Gansen, C., and Bartholomew, I. (2022). Uneven Success: Automatic Speech Recognition and Ethnicity-Related Dialects. *Speech Communication*, 140:50--70.

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.

Wiesner, M., Adams, O., Yarowsky, D., Trmal, J., and Khudanpur, S. (2019). Zero-Shot Pronunciation Lexicons for Cross-Language Acoustic Model Transfer. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1048--1054.

Williams, S., Foulkes, P., and Hughes, V. (2024). Analysis of Forced Aligner Performance on L2 English Speech. *Speech Communication*, 158:103042.

Winata, G., Aji, A. F., Yong, Z. X., and Solorio, T. (2023). The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936--2978, Toronto, Canada. Association for Computational Linguistics.

Wolff, H. (2023). Hausa Langauge. *Encyclopaedia Britannica*.

Xu, Q., Baevski, A., and Auli, M. (2022). Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. In *Interspeech 2022*, pages 2113--2117.

Yang, D., Hovy, D., Jurgens, D., and Plank, B. (2025). Socially Aware Language Technologies: Perspectives and Practices. *Computational Linguistics*, pages 1--15.

Yuan, J. and Liberman, M. (2008). Speaker Identification on the SCOTUS Corpus. *Journal of the Acoustical Society of America*, 123(5):3878.

Zhao, J., Pratap, V., and Auli, M. (2025). Scaling a Simple Approach to Zero-shot Speech Recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1--5.

Zhu, J., Yang, C., Samir, F., and Islam, J. (2024). The Taste of IPA: Towards Open-Vocabulary Keyword Spotting and Forced Alignment in Any Language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 750--772. Association for Computational Linguistics.

Zhu, J., Zhang, C., and Jurgens, D. (2022). ByT5 Model for Massively Multilingual Grapheme-to-phoneme Conversion. In *Interspeech 2022*, pages 446--450.