INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

An Outlier Analysis of Vowel Formants from a Corpus Phonetics Pipeline

Emily P. Ahn[1], *Gina-Anne Levow*[1], *Richard A. Wright*[1], *Eleanor Chodroff*[2]

[1]University of Washington, USA
[2]University of Zurich, Switzerland

eahn@uw.edu, levow@uw.edu, rawright@uw.edu, eleanor.chodroff@uzh.ch

Abstract

With the growing availability of large-scale spoken databases, linguists are increasingly relying on automated tools to obtain time alignments of sound units to the speech signal. A typical automated 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 investigate 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 develop a novel outlier taxonomy that includes the broad outlier categories of transcript errors, alignment errors, formant tracking errors, linguistic variations, and fine samples. We show the utility of this outlier analysis in identifying weaknesses in corpus-specific and corpus-general pipeline assumptions, and discovering characteristics of particular languages.

Index Terms: corpus phonetics, acoustic phonetics, forced alignment, error analysis, G2P conversion, vowel formants

1. 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. 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. 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 [1]; a phonetician can compare the effectiveness of remote recording methods on retaining accurate acoustic measures [2]; a sociolinguist can discover measurable vowel quality differences between groups of speakers in a large corpus [3, 4]; and, a typologist can test the degree to which analytic constraints may account for crosslinguistic patterns in phonetic realization [5, 6].

Nevertheless, automation requires that the researcher commit to assumptions that may not always be met. In the following paper, 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 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 [7] or XPF [8], or ones that involve neural network models [9, 10]. These models measure accuracy with Word or Phone Error Rate compared to a gold phone transcription [7, 11, 12, 13]. 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, in which the output phone sequence is time-aligned to the audio. Forced aligners rely on acoustic models, which learn statistical distributions of acoustic properties from discrete sound units. The quality of the alignment can be evaluated with segment boundary time displacement [14, 15], a binary accuracy overlap score [16], or overlap rate [15]; 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 paper, 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 [17]. A popular method for extracting formants is the Burg linear predictive coding (LPC) algorithm that relies on several pre-specified parameters [18, 19, 3, 20]. Evaluation of formant extraction typically uses mean absolute differences between the gold standard and automatic formant estimates [21, 20]. These evaluation methods once again rely on gold data, which many researchers may not have. With the ex-

10.21437/Interspeech.2023-1052

ception 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.**

The end of this pipeline would ideally produce acoustic-phonetic measurements that represent the targeted speech. While averages of the data may cut through any noise generated by the assumptions, the present paper 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 paper 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.[1]

2. Data

The present outlier analysis focuses on subsets of two massively multilingual corpora: the CMU Wilderness Corpus [22] and its derivative VoxClamantis corpus [5], and the Mozilla Common Voice corpus [23] and its derivative VoxCommunis corpus [6]. 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), a Chadic language spoken in Niger and Nigeria [24]; Kazakh (ISO639-3:kaz), a Turkic language spoken primarily in Kazakhstan [25]; and Swedish (ISO639-3:swe), a Germanic Indo-European language spoken mainly in Sweden [26].

3. Methodology

3.1. Data Processing

We downloaded the language-specific data from the Wilderness corpus[2] (sampled at 16kHz and distributed as MP3 files) and 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; the files were lossy from the original MP3 format. The six datasets spanning these two corpora and 3 languages were all processed by the Epitran G2P toolkit [7]. While the forced alignment for the VoxClamantis (i.e., Wilderness derivative) corpus used a multilingual ASR model [27] trained with Kaldi [28], we trained acoustic models and generated alignments on the Common Voice data with the Montreal Forced Aligner [14], which also utilizes Kaldi.

The vowel formants were extracted with Praat [18] 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 5000 Hz, as recommended for the male vocal tract. The Common Voice data was processed with a five-formant ceiling at both 5000 Hz and 5500 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 [6]. 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 1), we down-sampled it by randomly selecting 1000 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 3 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 1 gives an overview of the data.

3.2. Outlier Discovery

To identify outlying formants, we implemented the following procedure. First, [29] showed that using the Mahalanobis distance metric based on the Minimum Covariance Determinant is effective for discovering multivariate outliers. We followed suit and fitted the first two formants into one bivariate Gaussian model[4] per vowel per dataset (e.g., one model for Wilderness Kazakh /i/).[5] 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. (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.

3.3. 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 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 tax-

[1] Code for this work is available at https://github.com/emilyahn/outliers.

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

[3] Only the validated utterances were downloaded.

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

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

	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 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 Kazakh vowel types differed across dataset types due to utilizing different versions of the G2P tool, Epitran.*

onomy 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.

   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.

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.

5. Fine

   There is no apparent error.

25% of the samples from each dataset were annotated by five trained linguists, while the remaining 75% had one annotator. Inter-annotator agreement across the five annotators was calculated with Krippendorff's Alpha [30]. Because the labels could be multiply selected, we followed [31] 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) and for Transcript Errors (0.91, compared to Linguistic Variations at 0.84).

To produce gold labels for the samples that were annotated by all five annotators, the following heuristic was applied. 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.[6]

4. Results

This section addresses the distribution of outlier vowel category types across languages and datasets. Table 1 provides an overview of the data and aggregate quantity of outliers (as determined by the Mahalanobis distance); Figure 1 provides raw counts of outlier types across 600 annotated outlying vowels. As shown in Figure 1, the Wilderness corpus contained more "upstream" 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 was manually aligned to the audio at the chapter level, individual chapter sentences were automatically aligned, resulting in some mismatched audio segments [22]. 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 "downstream" Formant Tracking Errors and Linguistic Variations. While MP3 compression, as found in Common Voice, has been shown to have minimal influence on formant tracking [32], formant tracking can be further improved through using informed thresholds and removing f0 biases [33, 34]. Altogether, these findings reveal more nuanced violations of our pipeline's assumptions.

5. Case Studies

Our manual audit revealed several linguistic phenomena that were not captured by our G2P assumptions, indicating that the

[6] 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.

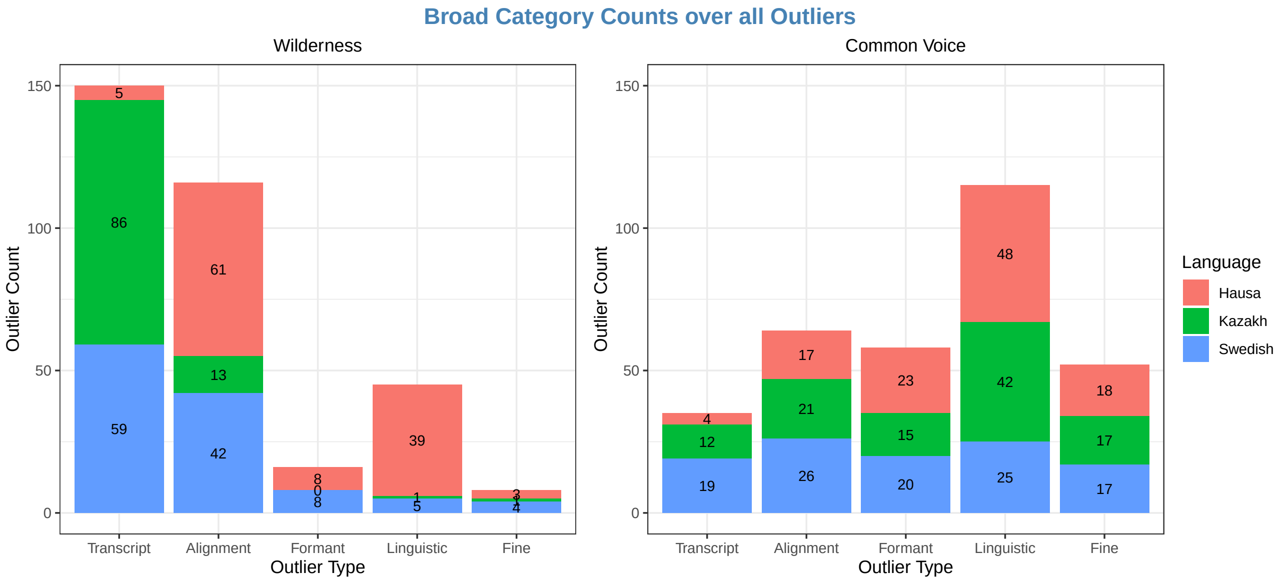

Figure 1: *Counts of 600 outliers annotated across 5 broad categories in Wilderness (left) and Common Voice (right).*

choice of granularity in the phonetic transcription can have downstream effects. Narrow transcriptions can inhibit cross-linguistic comparisons of the sound inventory, whereas broad transcriptions 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.

5.1. High Vowel Deletion in Kazakh

Results from our annotations indicated that Linguistic Deletions occurred most often in the Kazakh Common Voice data. According to [35], high, short vowels in Kazakh are more susceptible to reduction than other vowels. To test this, we conducted a logistic regression in R [36] 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 ($p < 0.001$). Interestingly, the analyzed cases of vowel deletion frequently occurred in sibilant environments. Our analysis may have identified several of these tokens since sibilants can have measurable formants, but at higher frequencies than would be expected from a vowel.

5.2. 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 are 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 [37]. (Vowel length is not entirely predictable from the orthography.) The lack of vowel length distinction in our G2P system, as well as potential vowel quality differences between long and short vowels, appear to produce inaccurate distributions of vowel formants in our analysis.

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

Future work may consider discovering outliers via alternative methods, whether by using an a priori threshold of expected formant values (e.g., [33]) 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 speech registers, noise environments, and across more languages. While we recommend always incorporating a partial manual audit in 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 today. The implications of this work include a call for careful analysis of what seem like errors in the output of automated systems.

7. Acknowledgements

We thank Anna Batra, Sam Briggs, Ivy Guo, and Emma Miller for their work on the annotations, processing of results, and exploratory data analyses. Author EPA was supported by NSF GRFP grant DGE-2140004, and Author EC by SNSF grant 208460.

8. References

[1] R. Coto-Solano and S. F. Solórzano, "Comparison of Two Forced Alignment Systems for Aligning Bribri Speech," *CLEI ELECTRONIC JOURNAL*, vol. 20, no. 1, 2017.

[2] J. Calder, R. Wheeler, S. Adams, D. Amarelo, K. Arnold-Murray, J. Bai, M. Church, J. Daniels, S. Gomez, J. Henry, Y. Jia, B. Johnson-Morris, K. Lee, K. Miller, D. Powell, C. Ramsey-Smith, S. Rayl, S. Rosenau, and N. Salvador, "Is Zoom Viable for Sociophonetic Research? A Comparison of In-Person and Online Recordings for Vocalic Analysis," *Linguistics Vanguard*, 2022.

[3] S. Reddy and J. N. Stanford, "Toward Completely Automated Vowel Extraction: Introducing DARLA," *Linguistics Vanguard*, vol. 1, no. 1, 2015.

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

[5] E. Salesky, E. Chodroff, T. Pimentel, M. Wiesner, R. Cotterell, A. W. Black, and J. Eisner, "A Corpus for Large-Scale Phonetic Typology," in *Association for Computational Linguistics*, 2020, pp. 4526–4546.

[6] E. P. Ahn and E. Chodroff, "VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis," in *Proceedings of the 13th Language Resources and Evaluation Conference*, 2022, pp. 5286–5294.

[7] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for Many Languages," in *Proceedings of the 11th Language Resources and Evaluation Conference*, 2018, pp. 2710–2714.

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

[9] P. Makarov and S. Clematide, "Imitation Learning for Neural Morphological String Transduction," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2877–2882.

[10] M. Hammond, "Data Augmentation for Low-Resource Grapheme-to-Phoneme Mapping," in *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2021, pp. 126–130.

[11] L. F. Ashby, T. M. Bartley, S. Clematide, L. Del Signore, C. Gibson, K. Gorman, Y. Lee-Sikka, P. Makarov, A. Malanoski, S. Miller, O. Ortiz, R. Raff, A. Sengupta, B. Seo, Y. Spektor, and W. Yan, "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*, 2021, pp. 115–125.

[12] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring Grapheme-to-Phoneme Conversion with Joint N-gram Models in the WFST Framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2016, publisher: Cambridge University Press.

[13] J. L. Lee, L. F. Ashby, M. E. Garza, Y. Lee-Sikka, S. Miller, A. Wong, A. D. McCarthy, and K. Gorman, "Massively Multilingual Pronunciation Modeling with WikiPron," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4223–4228.

[14] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech*. ISCA, 2017, pp. 498–502.

[15] S. Gonzalez, J. Grama, and C. Travis, "Comparing the Performance of Forced Aligners Used in Sociophonetic Research," *Linguistics Vanguard*, vol. 5, 2020.

[16] T. J. Mahr, V. Berisha, K. Kawabata, J. Liss, and K. C. Hustad, "Performance of Forced-Alignment Algorithms on Children's Speech," *Journal of Speech, Language, and Hearing Research*, pp. 2213–2222, 2021.

[17] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Cengage Learning, 2014.

[18] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer (Version 6.0.16)," 2019. [Online]. Available: http://www.praat.org/

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

[20] S. Barreda, "Fast Track: Fast (Nearly) Automatic Formant-Tracking Using Praat," *Linguistics Vanguard*, vol. 7, no. 1, 2021.

[21] K. Evanini, S. Isard, and M. Liberman, "Automatic Formant Extraction for Sociolinguistic Analysis of Large Corpora," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[22] A. W. Black, "CMU Wilderness Multilingual Speech Dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5971–5975.

[23] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 11–16.

[24] H. Wolff, "Hausa Langauge," Jan 2023. [Online]. Available: https://www.britannica.com/topic/Hausa-language

[25] "Kazakh Langauge," Feb 2023. [Online]. Available: https://www.britannica.com/topic/Kazakh-language

[26] "Swedish Langauge," Jan 2023. [Online]. Available: https://www.britannica.com/topic/Swedish-language

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

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi Speech Recognition Toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[29] C. Leys, O. Klein, Y. Dominicy, and C. Ley, "Detecting Multivariate Outliers: Use a Robust Variant of the Mahalanobis Distance," *Journal of Experimental Social Psychology*, vol. 74, pp. 150–156, 2018.

[30] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*. SAGE Publications, 2018.

[31] I. Martín-Morató and A. Mesaros, "What is the Ground Truth? Reliability of Multi-Annotator Data for Audio Tagging." IEEE, 2021, pp. 76–80.

[32] R. J. J. H. van Son, "A Study of Pitch, Formant, and Spectral Estimation Errors Introduced by Three Lossy Speech Compression Algorithms," *Acta Acustica United with Acustica*, vol. 91, no. 4, pp. 771–778, 2005.

[33] M. Lee, J. van Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 741–750, 2005.

[34] C. H. Shadle, H. Nam, and D. Whalen, "Comparing Measurement Errors for Formants in Synthetic and Natural Vowels," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 713–727, 2016.

[35] A. G. McCollum and S. Chen, "Kazakh," *Journal of the International Phonetic Association*, vol. 51, no. 2, pp. 276–298, 2021.

[36] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/

[37] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/