An Outlier Analysis of Vowel Formants from a Corpus Phonetics Pipeline

Emily P. Ahn, Gina-Anne Levow, Richard A. Wright, & Eleanor Chodroff

University of Washington, USA — *University of Zurich, CH*

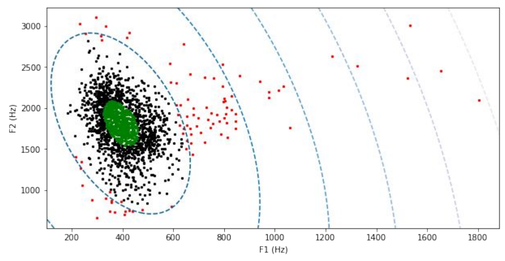
1. Motivation

> **Automatic forced alignment and phonetic measurement aids field linguists, phoneticians, sociolinguists**

> **Understand outliers from a fully automated corpus phonetics pipeline**
- Distinguish between technical errors & true linguistic variation
- Develop taxonomy of error types

2. Methods

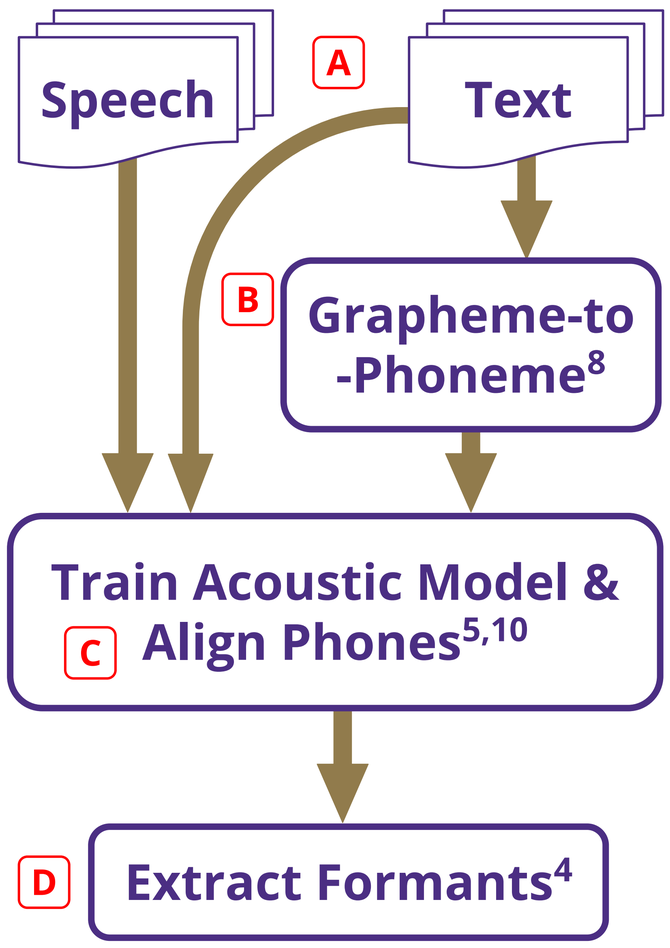
1. **Download 2 read speech corpora**
   - Wilderness[3] Bible & VoxClamantis[9] alignments/formants
   - Mozilla Common Voice[2] sentences & VoxCommunis[1] alignments/formants
2. **Discover vowel formant outliers**
   - **Mahalanobis distance**



3. **Annotate with new taxonomy**
   - 840 vowel samples (600 outliers, 240 near-means)
   - 5 trained linguists (Krippendorff's alpha = 0.86, strong agreement)

	Available Corpus		Analyzed Corpus						
Wilderness	**Total Hours**	**Total Spkrs**	**Hours**	**Spkrs**	**Utts**	**Vowel Types**	**Vowel Tokens**	**# Outliers**	**% Outliers**
Hausa	20:40	5+	20:40	5+	9626	5	303577	9698	3.19%
Kazakh	18:51	5+	18:51	5+	8085	6	204701	22148	10.82%
Swedish	16:46	1	16:46	1	9516	16	204701	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%

3a. Pipeline



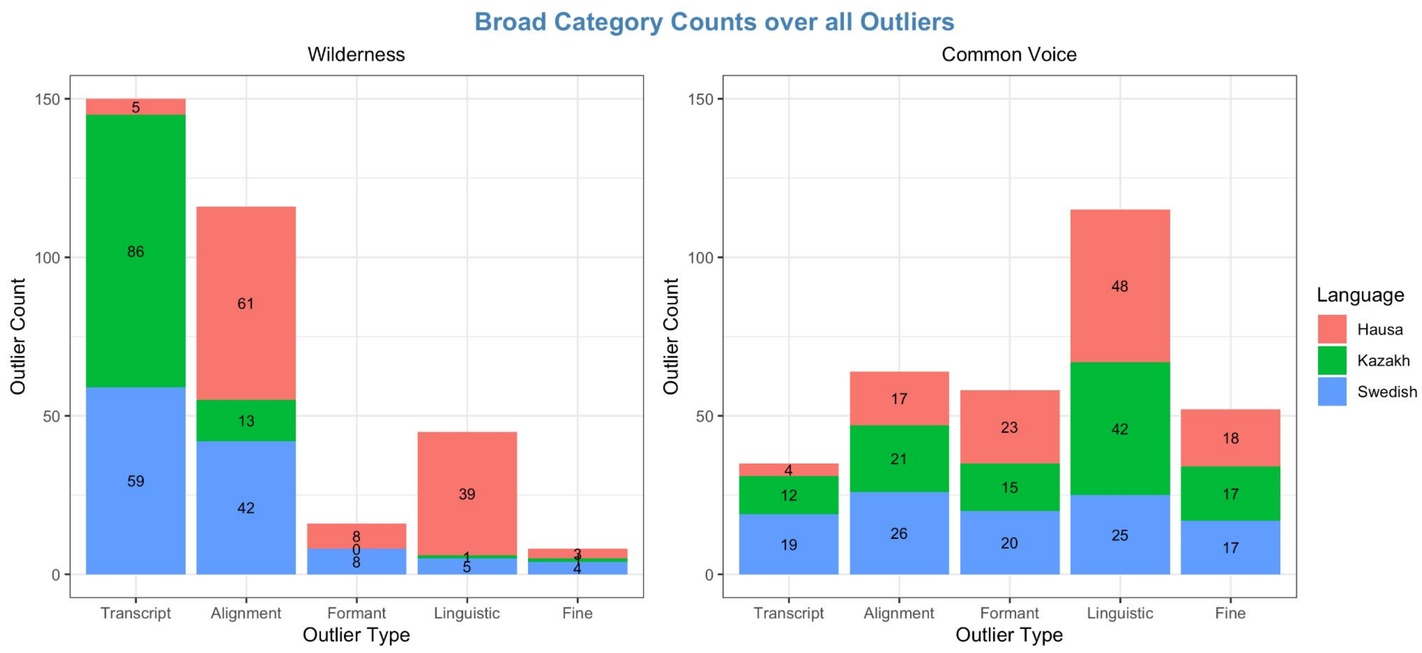
3b. Assumptions

A. Script = Transcript
B. Phonetic transcription is accurate
C. Segmentation is accurate
D. Acoustic-phonetic measurement is accurate

3c. Taxonomy

1. **Transcript Error**
   - Extra sounds (phones, syllables)
   - Extra transcript
   - Broad mismatch
2. **Alignment Error**
   - Target overlap
   - Broad alignment issue
3. **Formant Error**
   - Tracker and formant Hz mismatch
4. **Linguistic Variation**
   - Deletion of target vowel
   - Change (different vowel produced)
5. **Fine**

4. Results



5. Case Studies

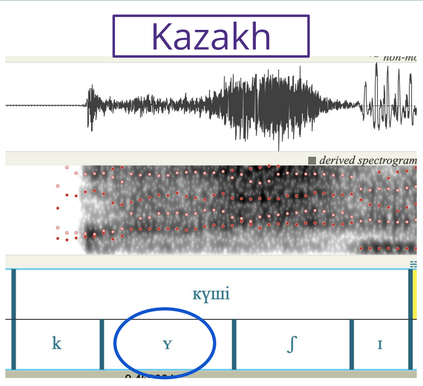
Why are these vowel formants outlying?

→ G2P specification & output may not explicitly represent pronunciation

1) High Vowel Deletion in Kazakh

High (short) vowels in Kazakh are more susceptible to reduction (McCollum & Chen, 2021).

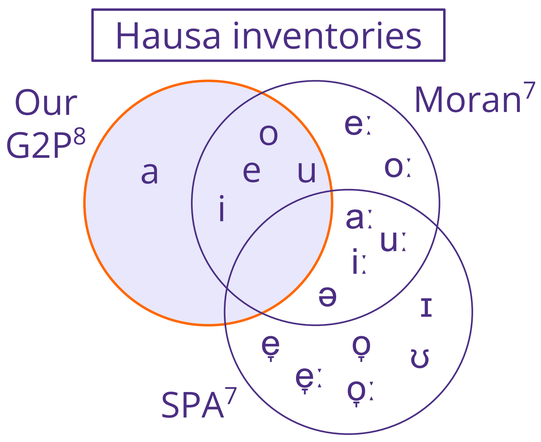
→ High vowels are 1.7 times more likely to be deleted than non-high vowels (*p<0.001*), especially in Kazakh (*p<0.001*)



2) Vowel Length in Hausa

> 44% of Hausa outliers & 64% of Hausa near-means marked as Linguistic Change

> Issue: disagreement of phoneme inventory
- In PHOIBLE[7], some linguists include long vowels, while our G2P does not
  - Vowel length semi-predictable, phonemic?
  - Vowel quality more centralized?



6. Conclusion

Summary

1. We develop a novel outlier taxonomy for a corpus phonetics pipeline
2. The distribution of outliers reveals dataset quality & quirks, language-specific phenomena

Future Work

> Apply taxonomy to new data, measures

> Identify efficient solutions to avoid & correct errors

References

1. Ahn, E., & Chodroff, E. (2022). VoxCommunis: A corpus for cross-linguistic phonetic analysis. In *LREC*.
2. Ardila et al. (2020). Common Voice: A massively-multilingual speech corpus. In *LREC*.
3. Black, A. W. (2019). CMU wilderness multilingual speech dataset. In *ICASSP*.
4. Boersma, B. & Weenink, D. (2019). Praat: Doing Phonetics by Computer (Version 6.0.16).
5. McAuliffe et al. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*.
6. McCollum, A. G. and Chen, S. (2021). Kazakh. In *JIPA*.
7. Moran, S. & McCloy, D., Eds. (2019). PHOIBLE 2.0.
8. Mortensen et al. (2018). Epitran: Precision G2P for many languages. In *LREC*.
9. Salesky et al. (2020). A corpus for large-scale phonetic typology. In *ACL*.
10. Wiesner et al. (2019). Zero-shot pronunciation lexicons for cross-language acoustic model transfer. In *IEEE ASRU*.

Acknowledgments

We thank Anna Batra, Sam Briggs, Ivy Guo, and Emma Miller. This work was supported in part by NSF GRFP grant DGE-2140004 and SNSF grant 208460.