Testing the Effect of Speaker Diarization and Speech Separation on Vowel Formant Estimates

Joseph A. Stanley
Brigham Young University

Lisa Morgan Johnson
Brigham Young University

Earl Kjar Brown
Brigham Young University

Annual Meeting of the Linguistics Society of America
January 12, 2025
Philadelphia, Pennsylvania

Technological Advancements

- Recent sociophonetic software has made analyzing large datasets possible.
  - Manual transcription: Transcriber (Boudahmane et al. 1998) and ELAN (Brugman & Russel 2004)
  - Automated transcription: DARLA (Reddy & Stanford 2015), Bed Word (Ma, Glass & Stanford 2024), and recently developed AI tools (Radford et al. 2023)
  - Forced-aligners: MAUS (Schiel 1999), ProsodyLab (Gorman, Howell & Wagner 2011), and MFA (McAuliffe et al. 2017)
  - Automated extraction of acoustic data: FAVE (Rosenfelder et al. 2014) and Fast Track (Barreda 2021)
- This facilitates analyzing data originally gathered for linguistic analysis
  - public speeches (Harrington, Palethorpe & Watson 2000; Bowie 2003; Wolfram et al. 2016; Holliday 2024)
  - personal vlogs (Mendoza-Denton 2011; Lee 2017; Cheng 2018, 2023).
  - oral histories collected by folklorists and historians (Hickey 2017; Strelluf & Gordon 2024 among many others)

We are all indebted to these developers!

Overlapping Speech

- Recording any social interaction likely involves overlapped speech.
- What can we do?
  - Go through and code it and then exclude it? (Olsen et al. 2017)
  - Consider it an acceptable loss?
- Excluding overlapped speech is especially hard for smaller datasets
  - Archival recordings, infrequent variables, etc.
- Potential solution
  - AI speech diarization and source separation

Speaker Diarization

- Applies speaker labels to segments in a single audio track
  - Answers "Who spoke when?"
- How is it done?
  - Like word embedding vectors, it extracts speaker embedding vectors based on voice characteristics.
  - Clusters those vectors and assigns a label.

Source Separation

- Produces separate audio files based on diarization.
- In theory, good models recover audio that would otherwise be excluded in sociophonetic analysis.

Is the output source separation good enough for sociophonetic analysis?

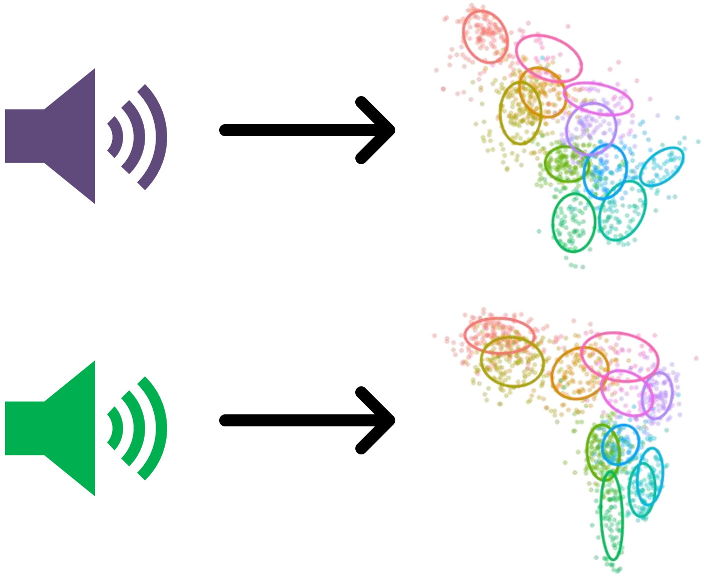
Can we recover some data that was otherwise lost?

Can we save on resources needed to manually tag overlap?

Methods

Baseline Measurements

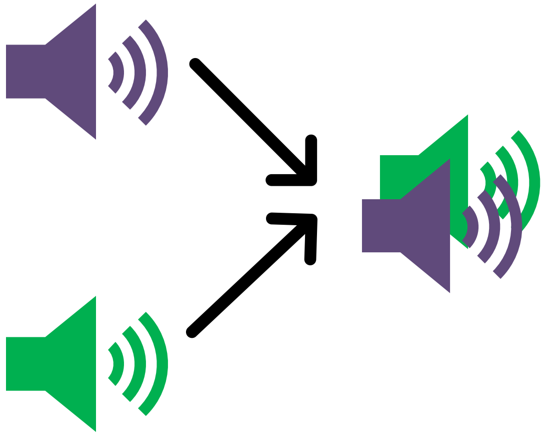
- Two speakers read 300 sentences in a sound booth.
  - "Olivia": female, 20, Asian-American, Atlanta; high-pitched, standard-sounding
  - "Tyler": male, 22, White, Atlanta; lower-pitched, slightly southern-sounding
- Processing
  - Manual utterance-level transcriptions
  - downsampled from 44.1kHz to 16kHz for direct comparison
  - MFA (McAuliffe et al. 2017) and FAVE (Rosenfelder et al. 2014) via DARLA (Reddy & Stanford 2015)



Artificial Overlap

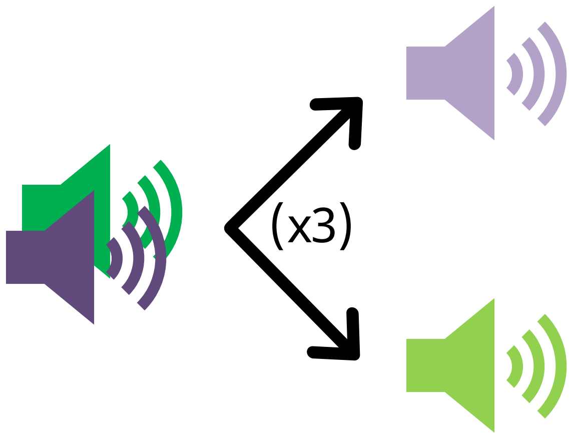
- Merged Tyler’s and Olivia’s audio into a single mono audio file
  - Swapped first and second halves of Tyler’s audio
  - Trimmed Tyler’s audio from 36 min to 33 min to match Olivia’s

- 53.6% of audio was overlapped speech.



Source Separation

- We used three (freely available) SepFormer models trained using SpeechBrain AI
  - Libri2mix
  - Whamr16K
  - WSJ02mix
  - They're only different in the data used to train them.
  - Concatenated 30-second chunks by identified speaker.



- Evaluation
  - Manually spot-checked 6 concatenated files.
  - Sent files through DARLA for alignment and extraction.

Evaluations

- Data comparison
  - Analyzed the new files using the same transcriptions and processing steps.
  - MFA and FAVE via DARLA
  - No manual interventions.

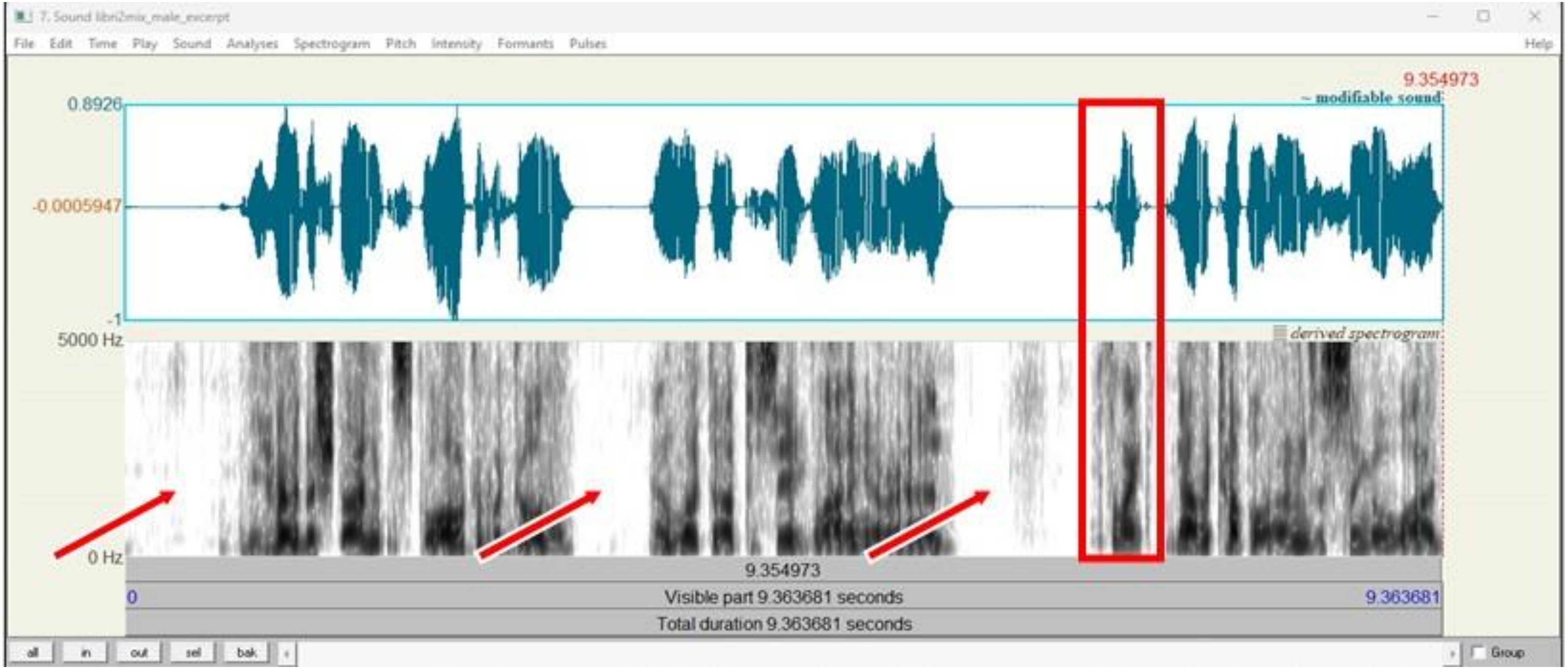
- Today's focus: midpoints of stressed, preobstruent monophthongs.
  - Mean: 1039 tokens per file.

Results

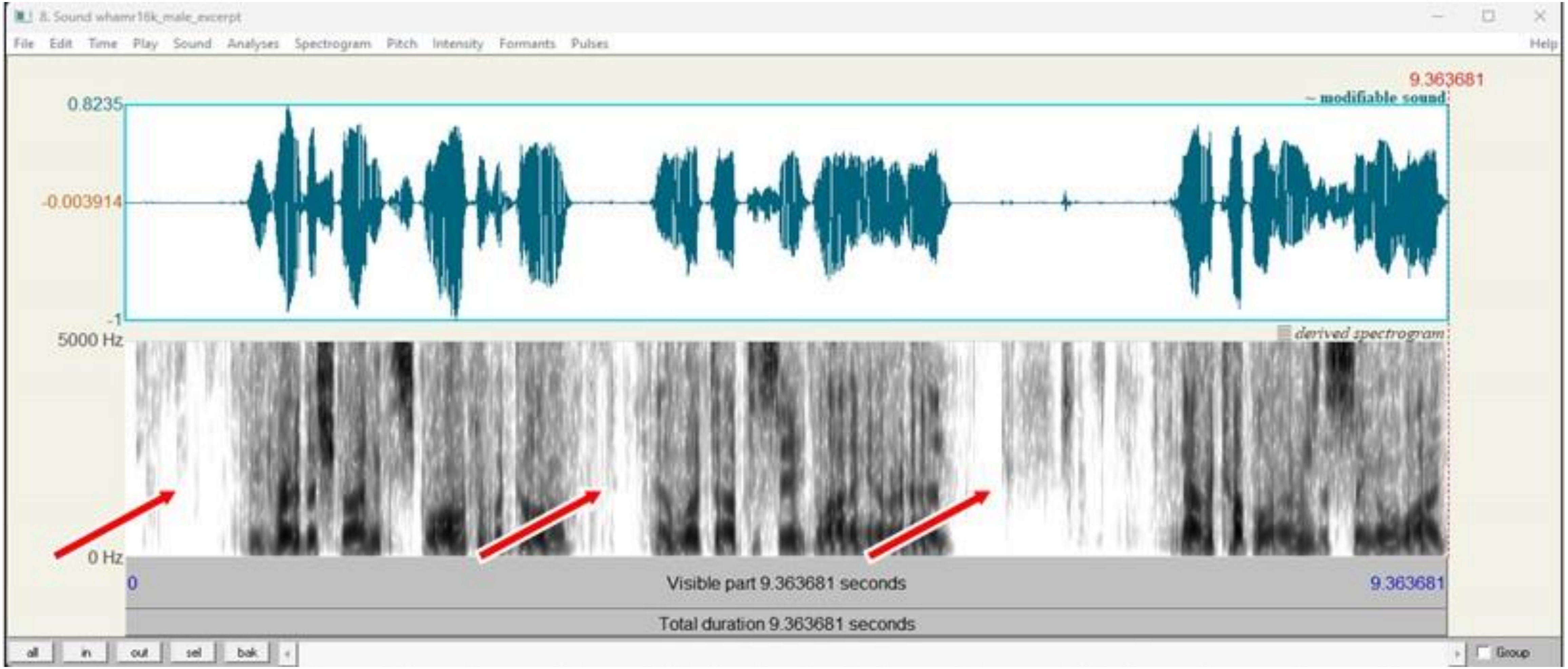
Auditory Checks for Performance

- Libri2
- WhamR
- WSJ02

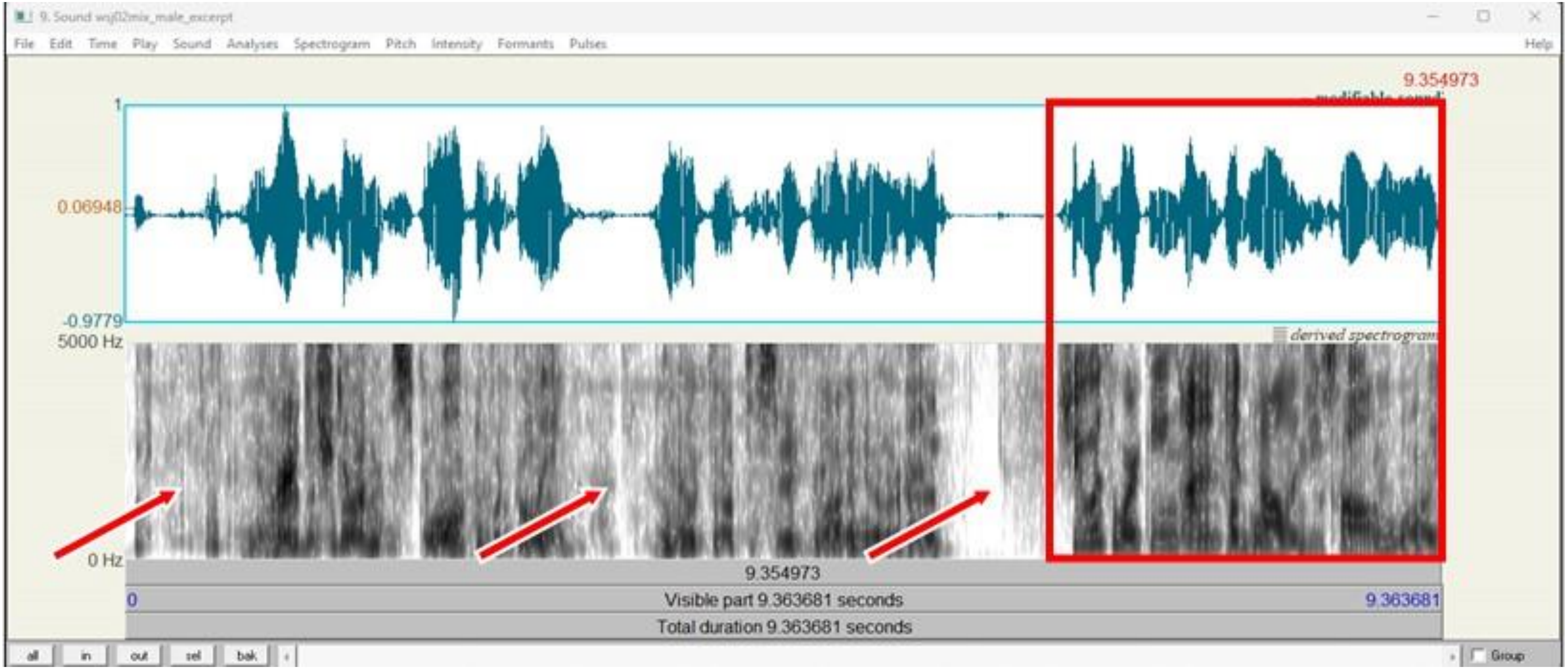
Example: Libri2



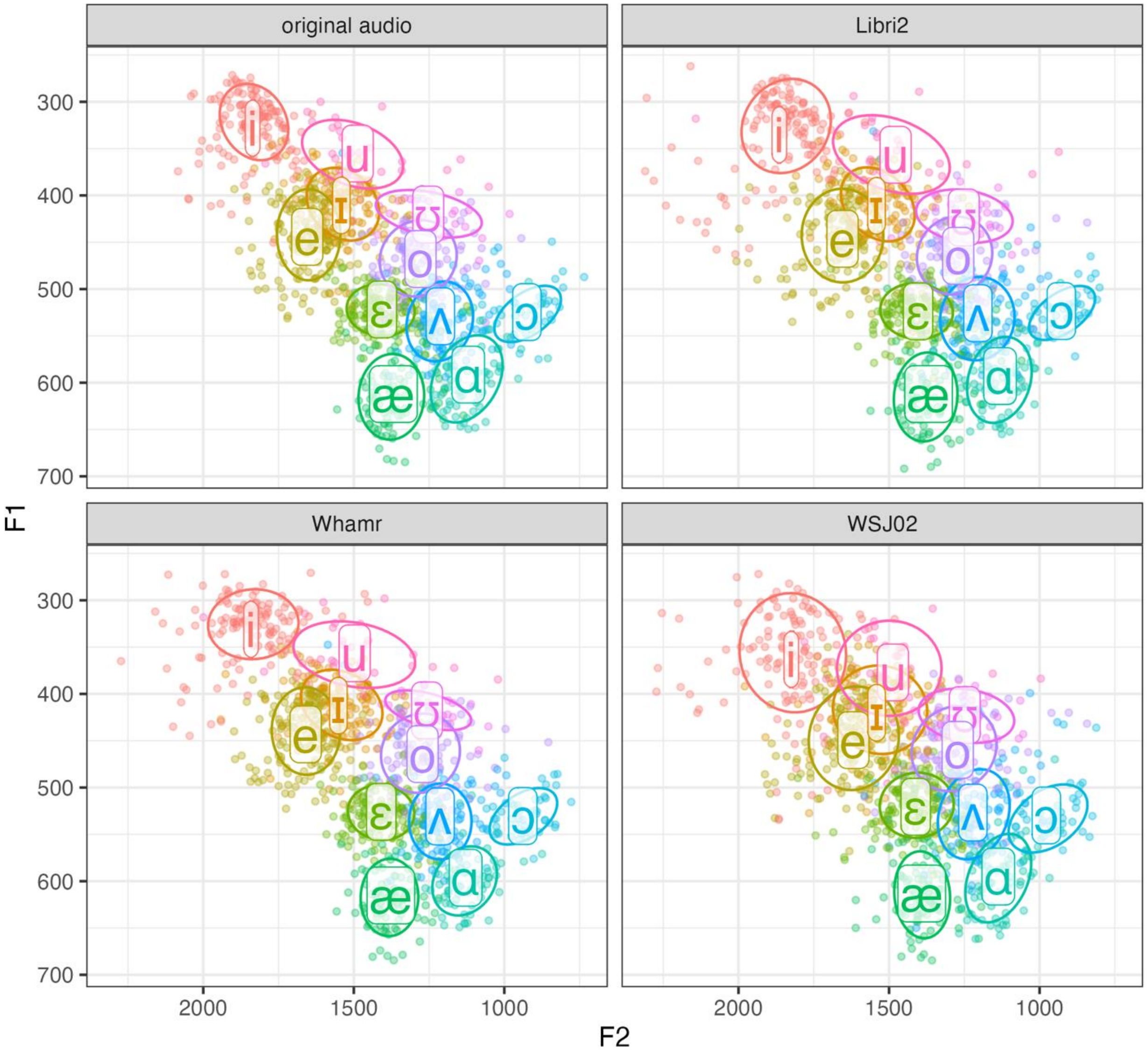
Example: Whamr



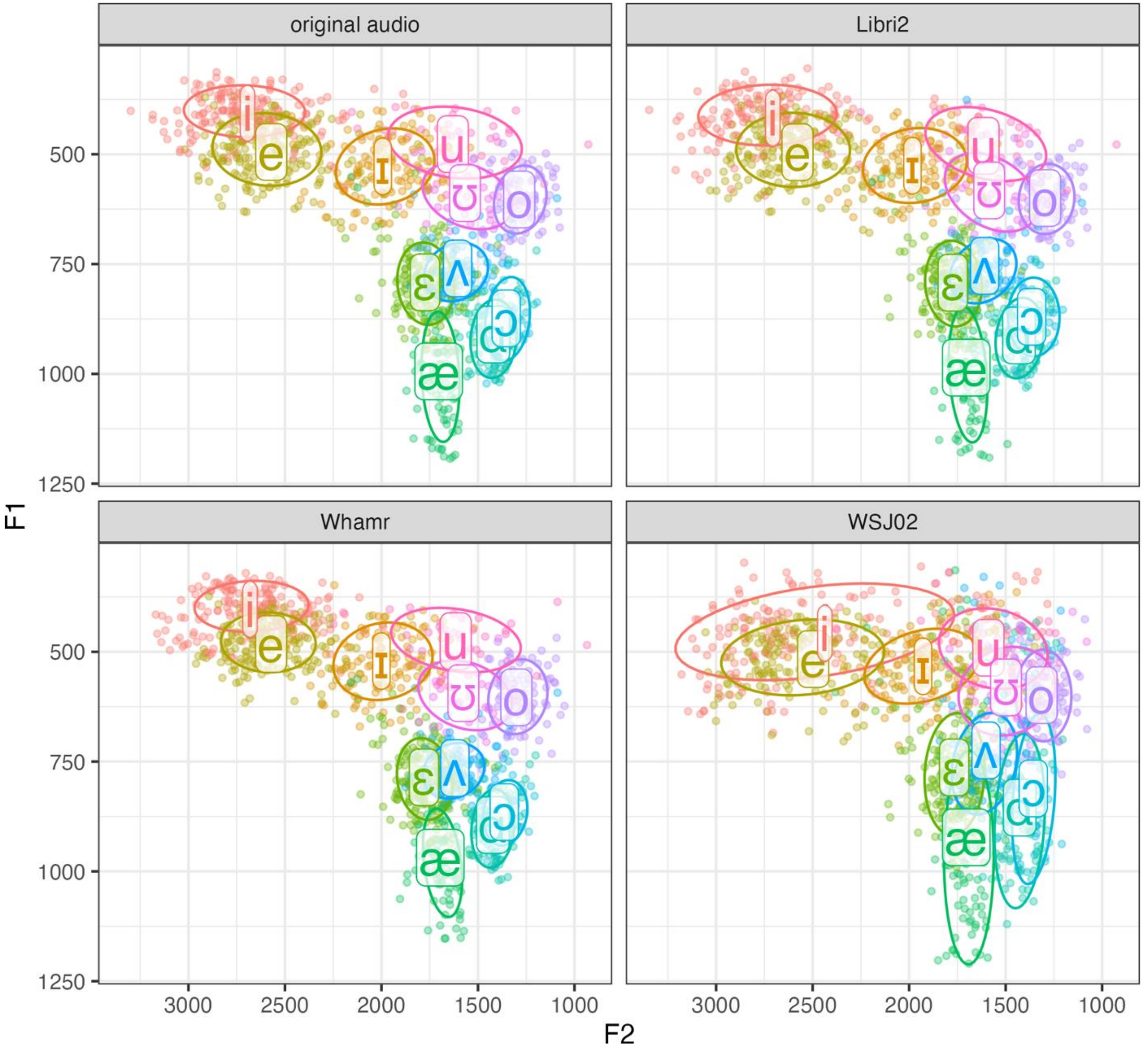
Example: WSJ02

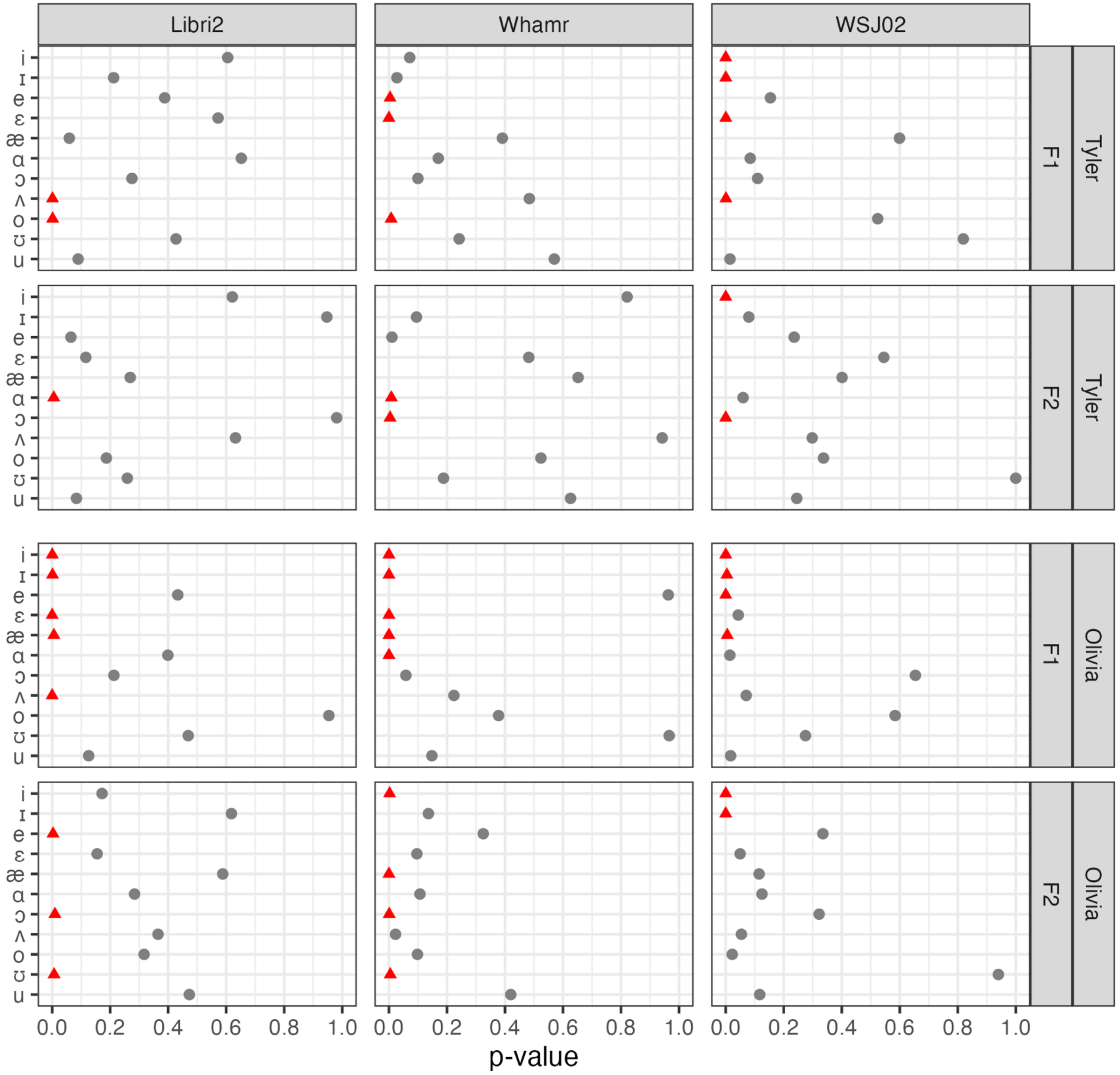


"Tyler" (male, 22, White, Atlanta; lower-pitched, slightly southern-sounding)

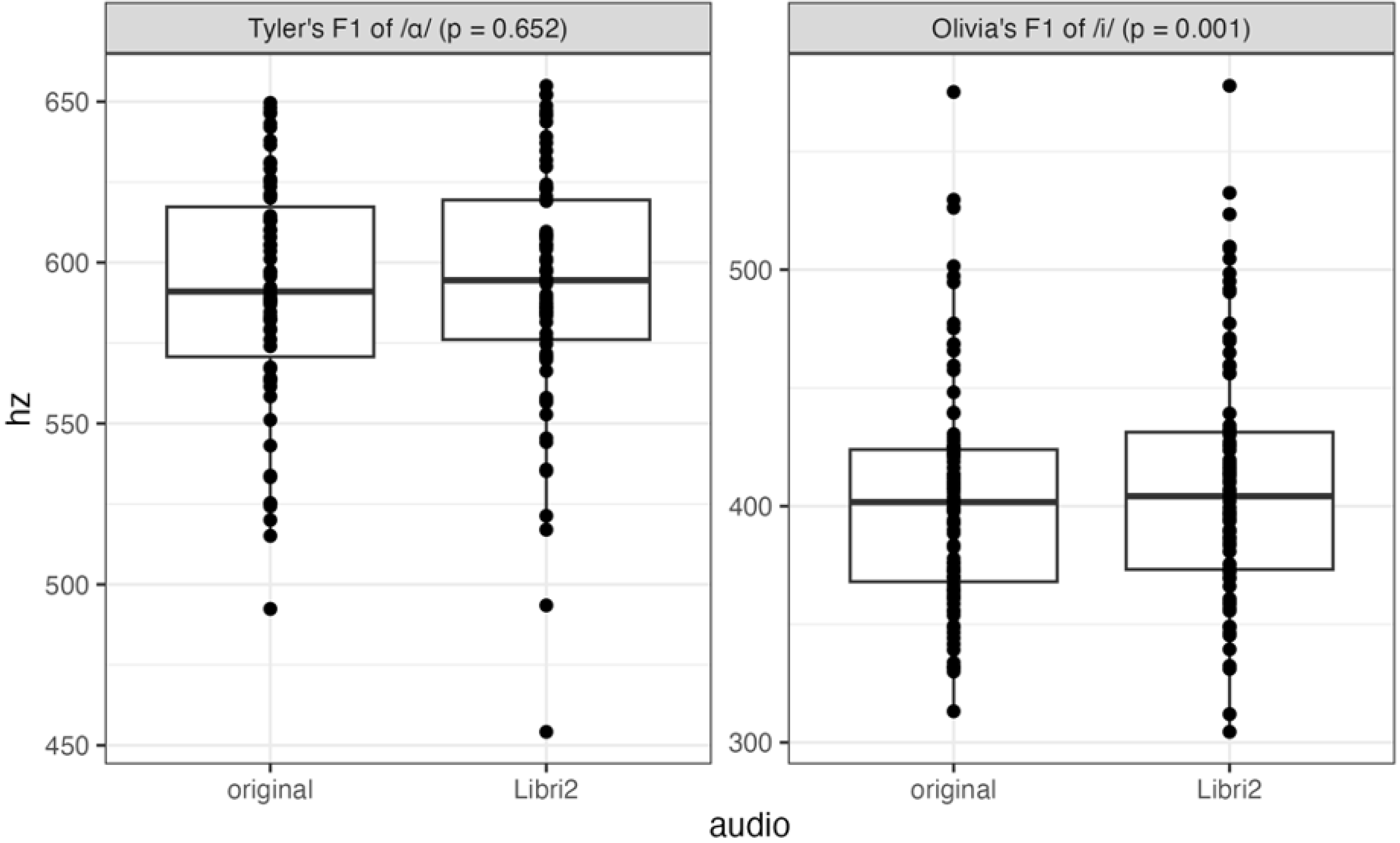


"Olivia" (female, 20, Asian-American, Atlanta; high-pitched, standard-sounding)

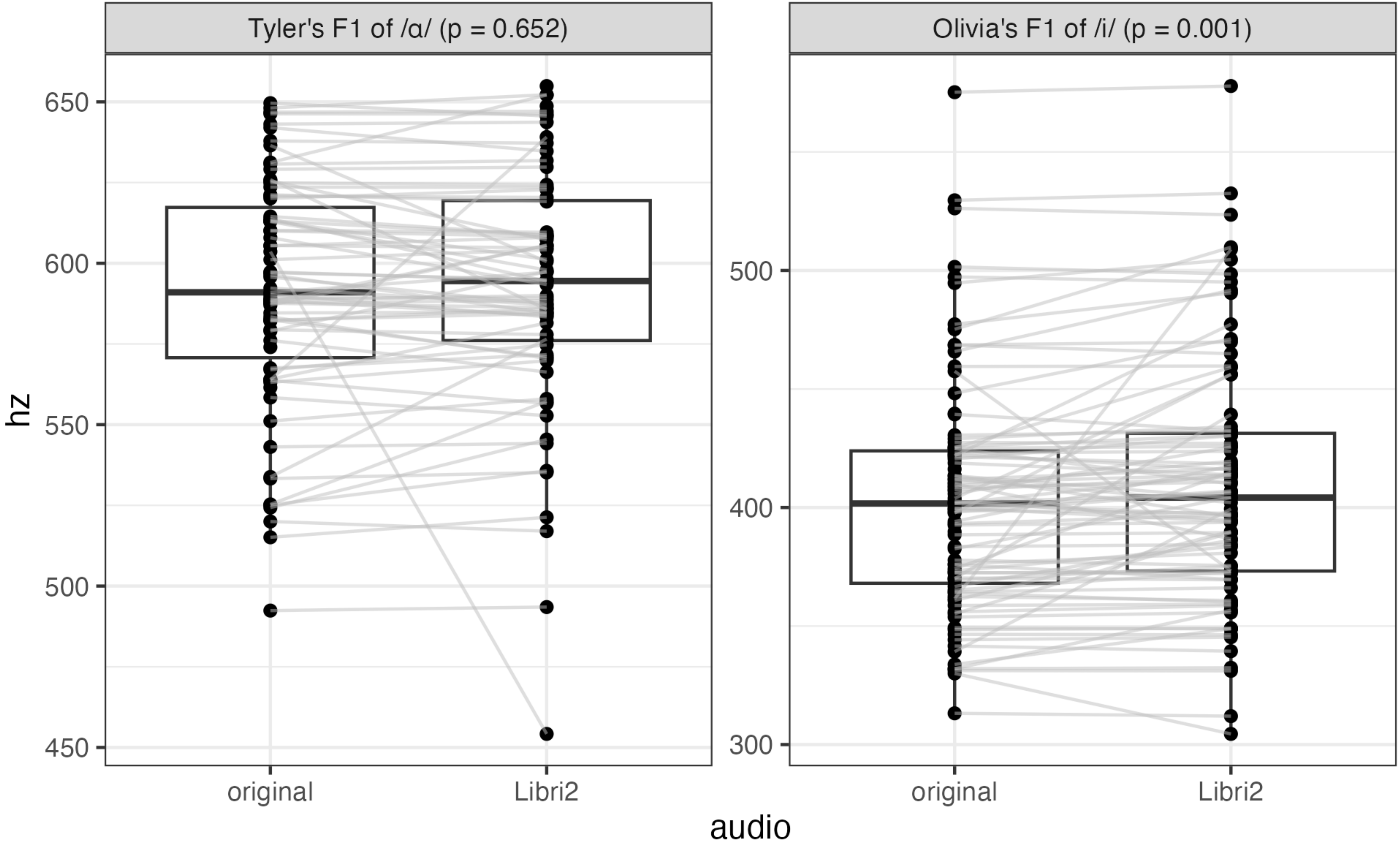




Libri2
Whamr
WSJ02
i
ɪ
e
ɛ
æ
ɑ
ɔ
ʌ
o
ʊ
u
F1
F2
Tyler
Olivia
0.0
0.2
0.4
0.6
0.8
1.0
p-value



Tyler's F1 of /ɑ/ (p = 0.652)
Olivia's F1 of /i/ (p = 0.001)
hz
650
600
550
500
450
500
400
300
original
Libri2
original
Libri2
audio



Tyler's F1 of /ɑ/ (p = 0.652)
Olivia's F1 of /i/ (p = 0.001)
hz
650
600
550
500
450
500
400
300
original
Libri2
original
Libri2
audio

Discussion

Overview

- For Libri2 and Whamr, the audio was remarkably pretty clean.

- While the mean formant measurements per vowel were usually small, differences for each observation were larger in unpredictable ways.
  - Differences were usually within the range of formant estimation variability (Kendall and Vaughn 2020)
  - Remarkably similar to Strelluf & Gordon (2024 chapter 3), who compared various interventions and automatic methods to hand-extracted formants.
  - Our differences were smaller than theirs, suggesting that using source separation has less of an effect than other types of cleaning.

- We're cautiously optimistic about these results.

Applications to Real-World Overlap

- Natural conversation has less overlap so it'll probably work better.

- Our many open-ended questions
  - Multiple speakers?
  - Non-pristine audio?
  - Equal volume?
  - Speaker dyads of more similar voices?
  - Non-standard varieties of English?

Recommendations

- Experiment with different models and continuously explore potentially better tools.
- Split audio at natural breaks rather than equal intervals.
- Listen to the output to ensure clean separation.
- Ensure that transcriptions match the new audio before conducting acoustic analysis.
- Treat formant estimates at the token level with caution. To be safe, only do analyses on vowel summaries like averages.
- Carefully document and report all methodological choices and human interventions.
- Do additional research on source separation for linguistic studies!

References

Barreda, Santiago. 2021. Fast Track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard* 7(1). https://doi.org/10.1515/lingvan-2020-0051.

Boudahmane, Karim, Mathieu Manta, Fabien Antoine, Sylvian Galliano & Claude Barras. 1998. Transcriber. http://trans.sourceforge.net/.

Bowie, David. 2003. Early development of the card-cord merger in Utah. *American Speech* 78(1). 31–51. https://doi.org/10.1215/00031283-78-1-31.

Brugman, Hennie & Albert Russel. 2004. Annotating multimedia/multi-modal resources with ELAN. Proceedings of the Fourth International Conference on Language Resources and Evaluation. Lisbon, 26–28 May. http://lrec-conf.org/proceedings/lrec2004/ (accessed 6 January 2025).

Cheng, Andrew. 2018. A longitudinal acoustic study of two transgender women on YouTube. *UC Berkeley Phonology Lab Annual Reports* 14. https://doi.org/10.5070/P7141042480.

Cheng, Andrew. 2023. Second dialect acquisition “in real time”: Two longitudinal case studies from YouTube. *American Speech* 98(2). 194–224. https://doi.org/10.1215/00031283-9766922.

Cosentino, Joris, Manuel Pariente, Samuele Cornell, Antoine Deleforge & Emmanuel Vincent. 2020. LibriMix: An open-source dataset for generalizable speech separation. arXiv. http://arxiv.org/abs/2005.11262.

Gorman, Kyle, Jonathan Howell & Michael Wagner. 2011. Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3). 192–193.

Harrington, Jonathan, Sallyanne Palethorpe & Catherine Watson. 2000. Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen’s Christmas broadcasts. *Journal of the International Phonetic Association* 30(1–2). 63–78. https://doi.org/10.1017/S0025100300006666.

Hickey, Raymond (ed.). 2017. *Listening to the past: Audio records of accents of English* (Studies in English Language). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781107279865.

Holliday, Nicole. 2024. Complex variation in the construction of a sociolinguistic persona: The case of Vice President Kamala Harris. *American Speech* 99(2). 135–166. https://doi.org/10.1215/00031283-10867240.

Kendall, Tyler & Charlotte Vaughn. 2015. Measurement variability in vowel formant estimation: A simulation experiment. In Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: University of Glasgow. https://www.internationalphoneticassociation.org/icphs/icphs2015.

Kendall, Tyler & Charlotte Vaughn. 2020. Exploring vowel formant estimation through simulation-based techniques. *Linguistics Vanguard* 6(s1). https://doi.org/10.1515/lingvan-2018-0060.

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347. https://doi.org/10.1016/j.csl.2017.01.005.

Lee, Sarah. 2017. Style-shifting in vlogging: An acoustic analysis of “YouTube Voice”. *Lifespans and Styles* 3(1). 28–39. https://doi.org/10.2218/ls.v3i1.2017.1826.

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

References

Ma, Marcus, Lelia Glass & James Stanford. 2024. Introducing Bed Word: A new automated speech recognition tool for sociolinguistic interview transcription. *Linguistics Vanguard.* https://doi.org/10.1515/lingvan-2023-0073.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. In *Proceedings of the 18th Conference of the International Speech Communication Association [Interspeech].* Stockholm, Sweden.

Mendoza-Denton, Norma. 2011. The semiotic hitchhiker's guide to creaky voice: Circulation and gendered hardcore in a Chicana/o gang persona. *Journal of Linguistic Anthropology* 21(2). 261–280. https://doi.org/10.1111/j.1548-1395.2011.01110.x.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv. http://arxiv.org/abs/1301.3781.

Olsen, Rachel M., Michael L. Olsen, Joseph A. Stanley, Margaret E. L. Renwick & William A. Kretzschmar, Jr. 2017. Methods for transcription and forced alignment of a legacy speech corpus. *Proceedings of Meetings on Acoustics* 30(1). 060001. https://doi.org/10.1121/2.0000559.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey & Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. Honolulu, HI.

Reddy, Sravana & James N. Stanford. 2015. Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1(1). https://doi.org/10.1515/lingvan-2015-0002.

Renwick, Margaret E. L. & D. Robert Ladd. 2016. Phonetic distinctiveness vs. lexical contrastiveness in non-robust phonemic contrasts. *Laboratory Phonology* 7(1). https://doi.org/10.5334/labphon.17.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard & Jiahong Yuan. 2014. FAVE (Forced alignment and vowel extraction) program suite, version 1.2.2.

Schiel, Florian. 1999. Automatic phonetic transcription of non-prompted speech. In *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco, CA.

Stanley, Joseph A. 2022. Order of operations in sociophonetic analysis. *University of Pennsylvania Working Papers in Linguistics* 28(1). Available at: https://repository.upenn.edu/pwpl/vol28/iss2/17.

Strelluf, Christopher & Matthew J. Gordon. 2024. *The origins of Missouri English: A historical sociophonetic analysis.* Lanham: Lexington Books.

Subakan, Cem, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi & Jianyuan Zhong. 2020. Attention is all you need in speech separation. arXiv. https://doi.org/10.48550/arXiv.2010.13154.

Wolfram, Walt, Caroline Myrick, Jon Forrest & Michael J. Fox. 2016. The significance of linguistic variation in the speeches of Rev. Dr. Martin Luther King Jr. *American Speech* 91(3). 269–300. https://doi.org/10.1215/00031283-3701015.

THANK YOU

Joseph A. Stanley
joey_stanley@byu.edu

Lisa Morgan Johnson
lisamorganjohnson@byu.edu

Earl Kjar Brown
Earl_brown@byu.edu