The Speech Articulation Toolkit (SATKit): Ultrasound image analysis in Python

Matt Faytak[1], Scott R. Moisik[2], Pertti Palo[3]

[1] UCLA [2] Nanyang Technological Univ. [3] Queen Margaret Univ.

faytak@ucla.edu; scott.moisik@ntu.edu.sg; pertti.palo@taurlin.org

UCLA NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE Queen Margaret University CLINICAL AUDIOLOGY, SPEECH AND LANGUAGE RESEARCH CENTRE

ISSP 12, December 2020

Purpose

A free, open-source collection of Python 3.x methods for high-throughput quantitative analysis of ultrasound imaging data

- We focus on lingual and laryngeal ultrsound here, but our methods are adaptable to any 2D grayscale image data (video, MRI), in theory
- Designed to work with AAA raw scanline data: large user base; already a locus for development [1]

Our initial focus is on **non-contour methods** for ultrasound analysis

- Automatic tongue surface contour extraction (e.g. [13, 6]) is increasingly fast and accurate
- But not the only approach, or even a suitable approach, for all data types or research questions

Have a look

We are still developing SATKit, which is hosted on GitHub at **giuthas/satkit**

Scan to visit the repo:

Or, use this URL:

git.io/JIPVA

Feedback, requests, etc. are appreciated!

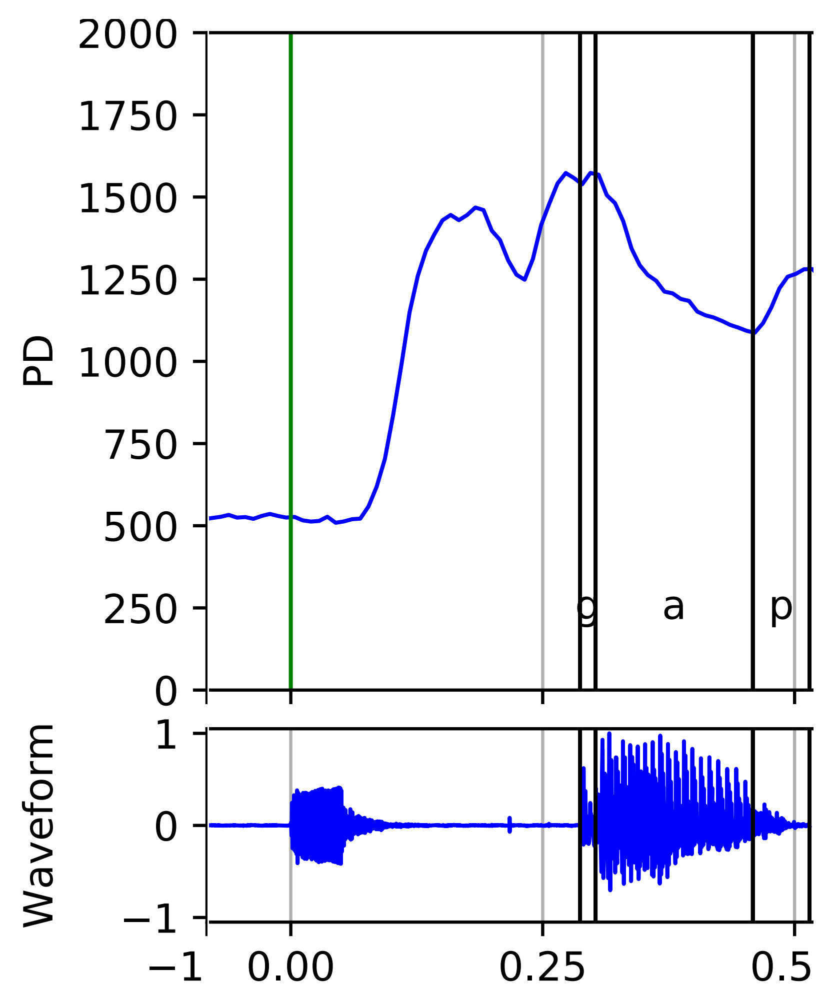
Pixel difference

Euclidean distance in terms of pixel intensity between pairs of images

- Captures **change** over entire image: surface contours, but also internal musculature
- SATKit implements two pixel difference methods from Palo [10]
  - Whole-image method: calculates PD over all matched pixels in pair of images
  - Scanline-based method: calculates PD for each column of pixels (more localised measure)

Among other things, well-suited to locating onset of articulation

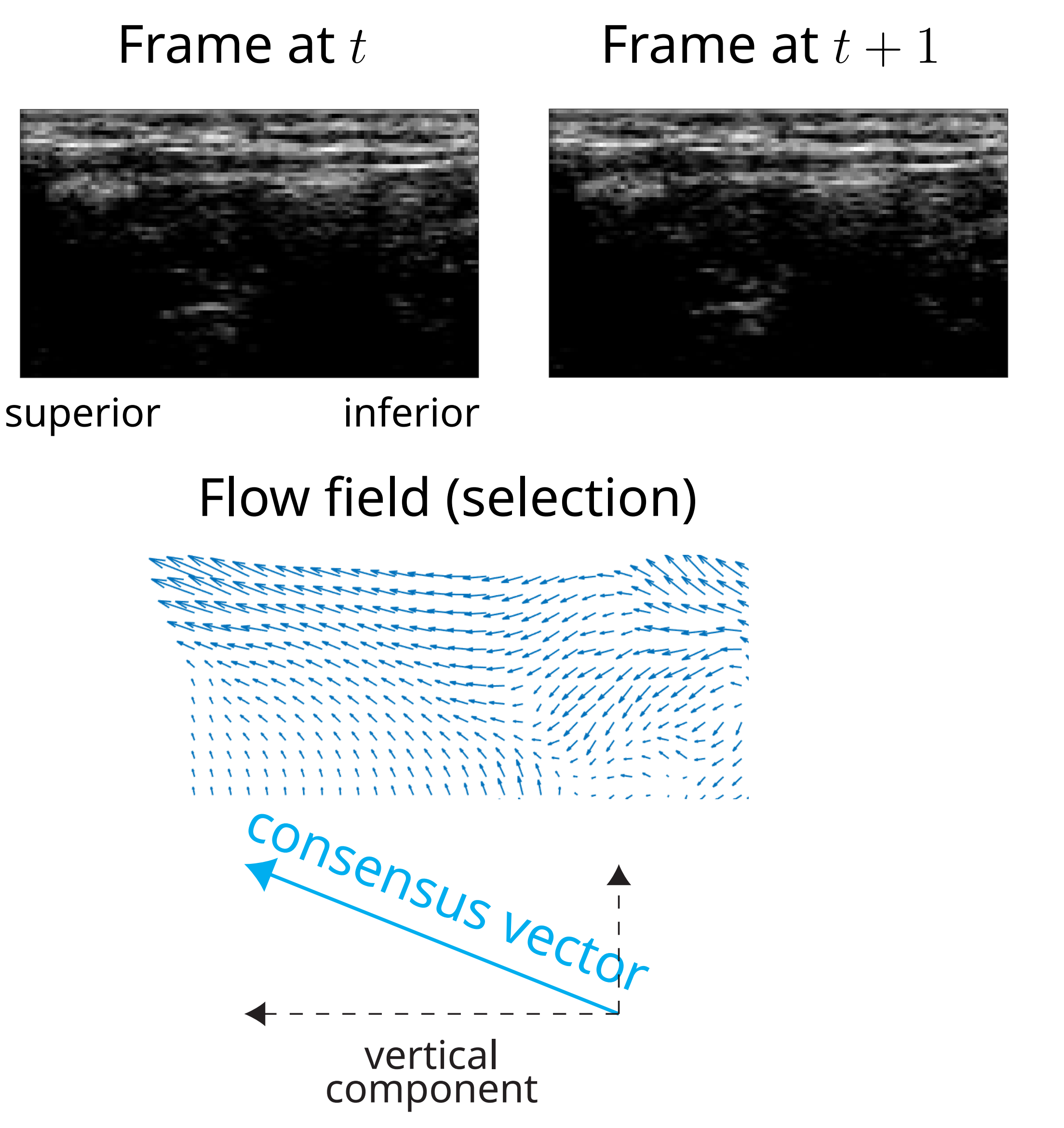
PD changes after go-signal (0s), but before release of /g/, in 'gap'



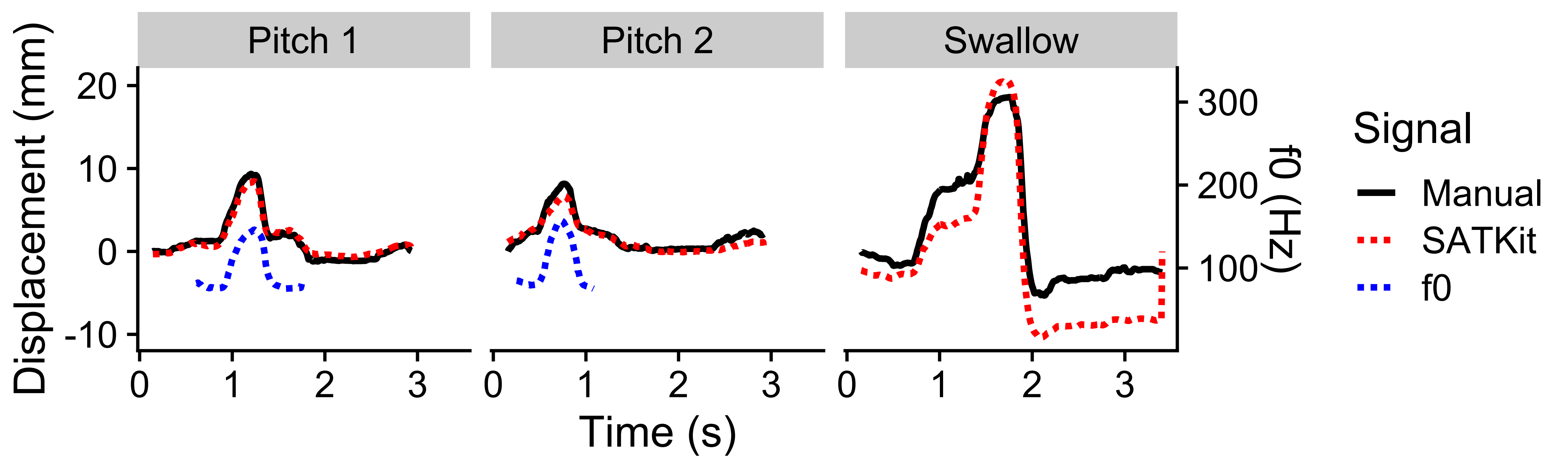
Optical flow

Characterizes direction and magnitude of **apparent motion** between pairs of frames [4]

- Especially well suited to laryngeal data (no single surface to track)
- SATKit implements method similar to Moisik et al. [9], but using dense optical flow, resulting in a **flow field** (one flow vector per pixel)
- **Consensus vectors** obtained by averaging entire fields or regions of interest; can be decomposed into horizontal/vertical velocity components
- **Displacement** can be estimated from cumulative integration of velocity signal



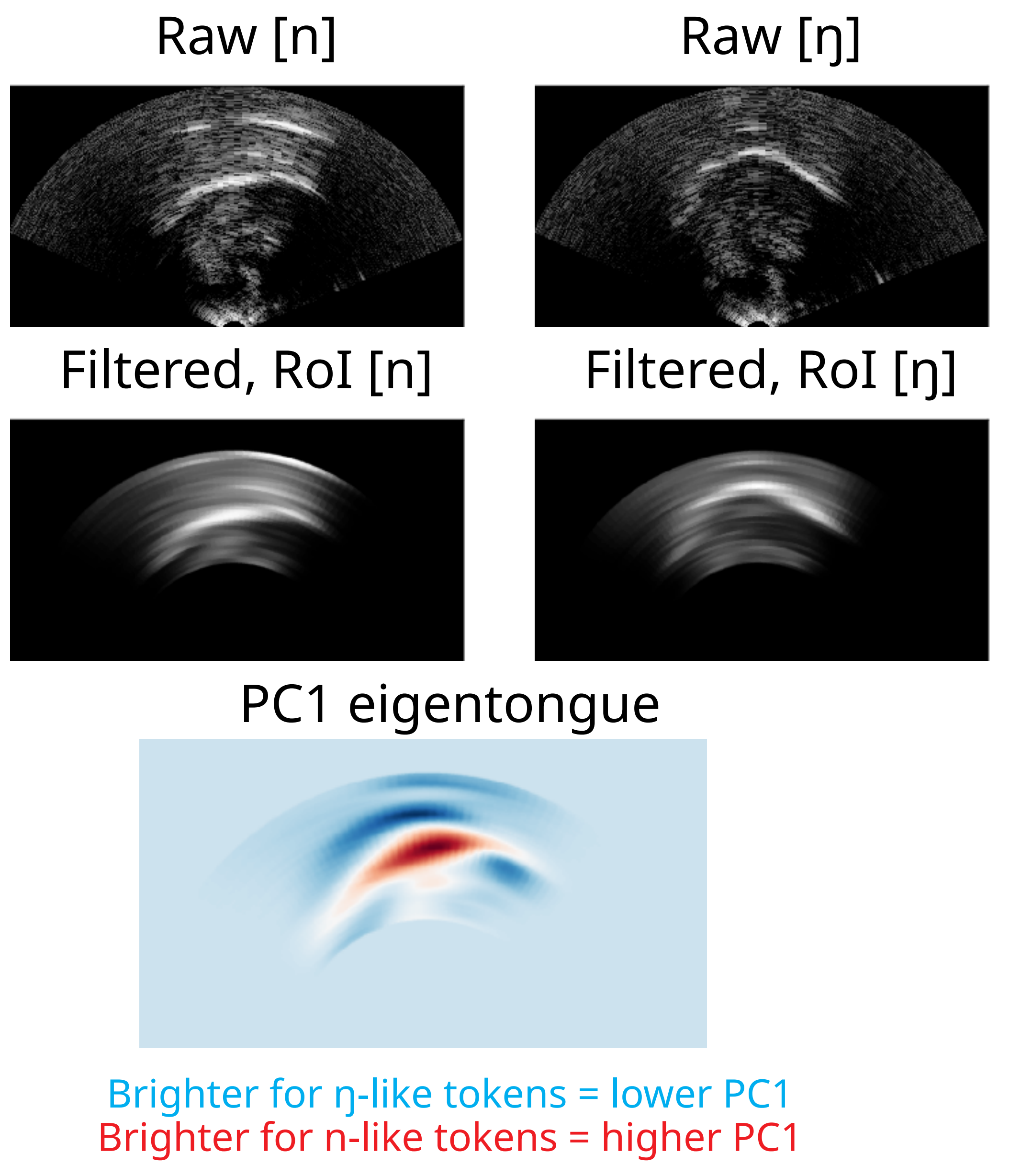
Estimated vertical displacement of larynx closely tracks manually validated displacement; covaries tightly with f0



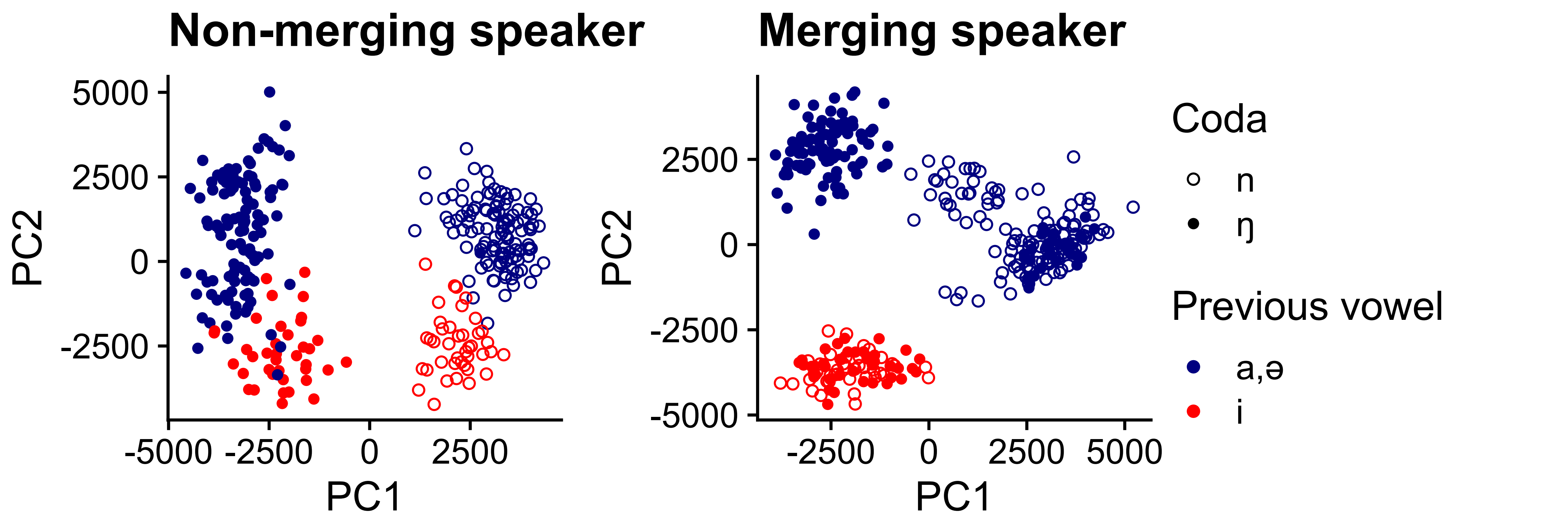
Dimensionality reduction

Discovers important, orthogonal **dimensions of variation** in a data set: here, patterns of covariation in pixel brightness [5, 8, 3, 7]

- SATKit uses principal component analysis (PCA) from scikit-learn [11]
- Utilities to support:
  - Filtering and applying region of interest masks
  - Reshaping and rescaling to eigentongues [5] or eigenlarynges, which help with interpretation of PCs
  - Linear discriminant analysis (LDA) can be used to generate time-varying articulatory signals from PCs, à la [8, 12]



Mandarin /n/–/ŋ/ contrast (see above); data from Faytak et al. [2]



Coming soon

- Separable GUIs and analysis functions
- Improved features (i.e. region of interest selection for pixel difference and optical flow methods)
- Additional documentation and sample data; unit testing

Acknowledgements

Thanks to Alan Wrench for AAA advice.

Poster PDF with **references**:

References

[1] A. Eshky, M. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. Scobbie, and A. Wrench. UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions. In *Proc Interspeech 2018*, pages 1888–1892, 2018.

[2] M. Faytak, S. Liu, and M. Sundara. Nasal coda neutralization in Shanghai Mandarin: Articulatory and perceptual evidence. *Laboratory Phonology*, 11(1), 2020.

[3] P. Hoole and M. Pouplier. Öhman returns: New horizons in the collection and analysis of imaging data in speech production research. *Computer Speech & Language*, 45:253–277, 2017.

[4] B. Horn and B. Schunck. Determining optical flow. In *Techniques and Applications of Image Understanding*, volume 281, pages 319–331. International Society for Optics and Photonics, 1981.

[5] T. Hueber, G. Aversano, G. Cholle, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone. Eigentongue feature extraction for an ultrasound-based silent speech interface. In *ICASSP 2007*, volume 1, pages I–1245. IEEE, 2007.

[6] C. Laporte and L. Ménard. Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. *Medical Image Analysis*, 44:98–114, 2018.

[7] J. Lin and S. Moisik. The lingual voice quality settings of Standard Singapore English and Singapore Colloquial English. In *Proc ICPhS 19*, 2019.

[8] J. Mielke, C. Carignan, and E. Thomas. The articulatory dynamics of pre-velar and pre-nasal/æ/-raising in English: An ultrasound study. *JASA*, 142(1):332–349, 2017.

[9] S. Moisik, H. Lin, and J. Esling. A study of laryngeal gestures in mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (sllus). *JIPA*, 44(1):21–58, 2014.

[10] P. Palo. *Measuring Pre-Speech Articulation*. PhD thesis, Queen Margaret University, 2019.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] J. Shaw, C. Carignan, T. Agostini, R. Mailhammer, M. Harvey, and D. Derrick. Phonological contrast and phonetic variation: The case of velars in Iwaidja. *Language*, 96(3):578–617, 2020.

[13] K. Xu, T. Csapó, P. Roussel, and B. Denby. A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization. *JASA*, 139(5):EL154–EL160, 2016.