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

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

[1]*University of California, Los Angeles, United States of America*
[2]*Nanyang Technological University, Singapore*
[3]*CASL Research Center, Queen Margaret University, United Kingdom*
faytak@ucla.edu, scott.moisik@ntu.edu.sg, pertti.palo@taurlin.org

Abstract

This paper introduces the Speech Articulation Toolkit, an open-source collection of Python 3 utilities for quantitative analysis of ultrasound image data. SATKit presently emphasizes pixel-based measures as a complement to more commonly used contour tracking methods. We provide an overview of the core utilities of the present version of SATKit: pixel difference, which characterizes the amount of change from frame to frame; optical flow, which gauges the magnitude and direction of apparent motion between frames; and dimensionality reduction, here focusing on capturing patterns of covariation in pixel brightness.

Keywords: ultrasound tongue imaging, laryngeal ultrasound, optical flow, pixel difference, dimensionality reduction

1. Introduction

The analytical landscape in ultrasound research is mainly based on contour tracking methods (Stone 2005; Kochetov 2019). Segmentation of contours from ultrasound images often involves time-consuming manual intervention or hand-correction, which may introduce replicability concerns (cf. Hoole & Pouplier 2017; Roettger 2019). Inter-speaker comparison is also often hindered by the need to normalize for differences in morphology, overall tongue size, and probe orientation and stabilization method (Slud *et al.* 2002; Heyne *et al.* 2019).

Even as automated tongue contour tracking has improved in accuracy with gradual refinements to the method (e.g. Li *et al.* 2005; Xu *et al.* 2016; Laporte & Ménard 2018), certain fundamental limitations remain. For instance, contour tracking methods cannot detect potentially informative changes to the tongue musculature below the contour surface (Koppenhaver *et al.* 2009; Vasseljen *et al.* 2009) and are not suitable for tracking articulators that cannot be treated as a single deformable edge, in particular the larynx.

In this article, we introduce the Speech Articulation Toolkit (SATKit; Palo *et al.* 2021), a freely available collection of Python 3 methods with an initial focus on direct quantitative analysis of the pixels in articulatory imaging data. We view these whole-image methods as complementary to contour tracking methods and potentially useful to a wider range of researchers using two-dimensional ultrasound imaging. All methods are designed to use raw scanline data stored by the Articulate Assistant Advanced software suite, an emerging standard for data collection, compact storage, and corpus development (Eshky *et al.* 2018; Ribeiro *et al.* 2021) which is accessible to a growing number of theoretical and clinical researchers.

2. Overview of features

SATKit is under development as of the writing of this manuscript. At the present moment, methods included in SATKit include pixel difference and optical flow, which characterize differences between pairs of images, and dimensionality reduction utilities for extracting dimensions of variation in pixel brightness. We review these features below.

2.1. Pixel difference

The *pixel difference* of a given pair of images is calculated as the Euclidean distance between them in terms of pixel intensity. This unitless measure captures the presence of change in an ultrasound signal, including tongue contour movement and changes in activation of the tongue's intrinsic musculature. Pixel difference methods are particularly well-suited to gauging the onset of articulation as a complement to reaction times measured from acoustics (e.g. McMillan & Corley 2010; Drake *et al.* 2013).

SATKit provides implementations of both of the pixel difference methods described in Palo (2019): a whole-image method calculates pixel difference over all pixels in the pair; and a scanline-based method calculates pixel difference for each column of pixels in the data, providing a localized measure of change. Image data are not spatially downsampled, as they are in McMillan & Corley (2010).

Whole-image pixel difference outputs a time series of distance values between ultrasound frame pairs of length *n-1* where *n* is the length of the sequence of frames provided. Such a time series is shown in Figure 1, where frame-by-frame pixel difference values above about 500 (a floor due to the noise typical of ultrasound imaging) indicate movement of the imaged portion of the tongue. Scanline-based pixel difference, which is not shown in Figure 1, outputs arrays of shape (*s*, *n*-1), where *n* is the length of the frame sequence as before and s is the number of scanlines in the data, effectively consisting of one vector of pixel distance values for each scan line.

2.2. Optical flow

A second method included in SATKit for characterizing frame-by-frame difference, *optical flow* characterizes the direction and magnitude of apparent motion between a pair of images. For each image pair, a field of vectors is computed which describes the "flow" of pixel brightness patterns (Horn & Schunck 1981). Optical flow has previously been applied to the analysis of ultrasound imaging data for medical (e.g., Danilouchkine *et al.* 2009) and phonetic research (Moisik *et al.* 2014; Poh & Moisik 2019).

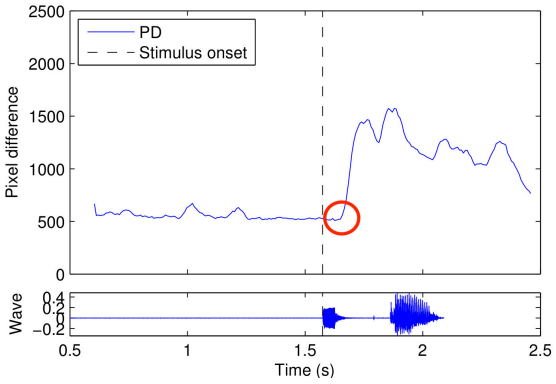

Figure 1: *Whole-image pixel difference for an utterance [kɔt] 'caught' with waveform. Dashed line indicates go-signal (1kHz beep); onset of articulation (circled) precedes acoustic word onset.*

Optical flow does not depend on the visibility or tracking of specific structures for analysis, but rather captures holistic patterns of motion: reasonable results can be obtained provided there is enough frame-by-frame consistency (i.e., the differences between frames are small). Because of this, optical flow is especially well-suited to analysis of laryngeal ultrasound (Moisik *et al.* 2014; Poh & Moisik 2019), where tracking the movement of specific structures of the larynx using contour methods is infeasible.

SATKit implements an optical flow method similar to that described in Moisik *et al.* (2014), but using dense optical flow (Farnebäck 2003), which is summarized in Figure 2. For a pair of frames of the same shape (*m*, *n*), a flow vector consisting of angle and magnitude measurements is calculated for each pixel. The vectors in the resulting flow field of shape (*m, n,* 2) are averaged to obtain a *consensus vector* for the entire field. Utilities are provided in SATKit for further processing of this signal: consensus vectors can be decomposed into velocity signals projected onto horizontal, vertical, or arbitrary oblique axes; and cumulative trapezoidal integration of velocity signals can be used to estimate displacement of rigid structures visible in the ultrasound's field of view. Time series for consensus vectors and derived measures can be calculated across all frames in a recording for each successive pair of frames (Figure 3).

SATKit provides both a batch-processing utility and a graphical user interface that allows for real-time visual inspection of the results of an optical flow analysis and for the adjustment of analysis parameters (region of interest, which component of the field is analyzed, *etc.*). The user interface also allows data to be exported in conventional formants (such as CSV) for further analysis with other software.

2.3. Dimensionality reduction

Dimensionality reduction characterizes the dimensions of variation in a data set by creating a small number of informative features from a much larger number of features. In SATKit, the especially large number of pixels in ultrasound images are reduced to patterns of covariation in pixel intensity, a method which has also seen use in some prior studies of lingual and laryngeal articulation (Hueber *et al.* 2007; Hoole & Pouplier, 2017; Mielke *et al.* 2017; Lin & Moisik, 2019). This approach is particularly useful for characterizing similarity and difference in articulation in speaker-specific terms.

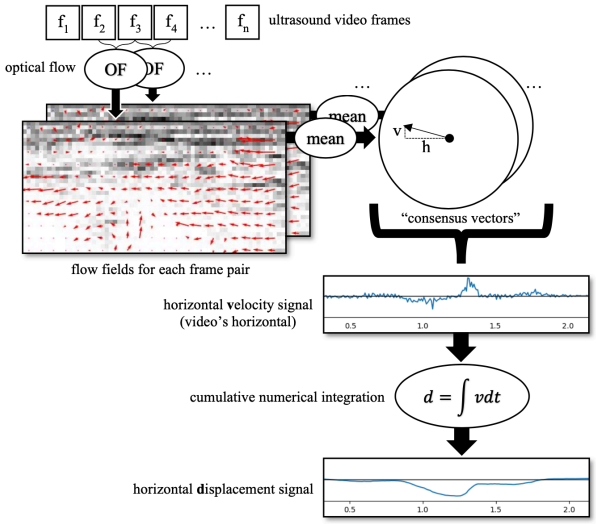

Figure 2: *Schematic of SATKit optical flow calculation and derived measurements.*

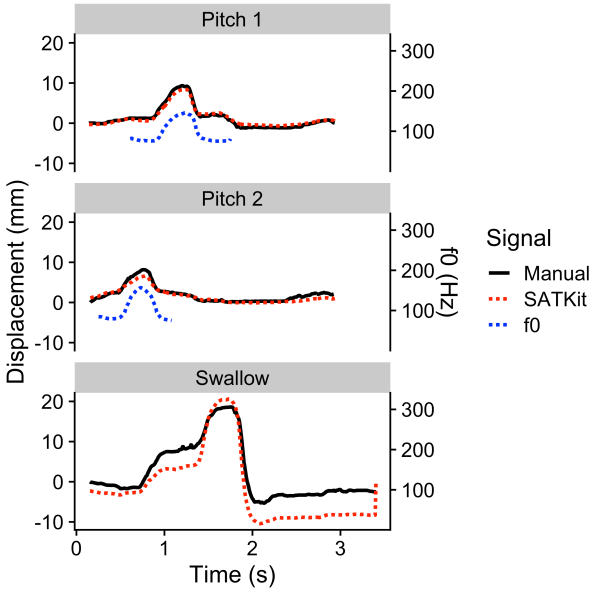

Figure 3: *Comparison of the vertical laryngeal displacement signal from SATKit, a manually validated displacement signal, and f_0.*

SATKit implements principal components analysis (PCA) using the scikit-learn package (Pedregosa *et al.* 2011), with the pixels of an image set used as basis data. Utilities are included for inputting and outputting data from standardized caches, interpolation of raw scanline data to physical dimensions, and region of interest selection from raw scanline or interpolated data. Edge enhancement and filtering operations from Carignan (2014) are included to reduce imaging noise and improve the performance of the PCA, including a speckle-reducing anisotropic diffusion filter specifically designed for ultrasound applications (Yu & Acton 2002).

PCA yields principal components (PCs), uncorrelated dimensions which are rank-ordered by the proportion of variation in the basis data they explain. PC scores, which characterize each observation in the basis data in terms of its position on the new dimensions, are also produced. While dimensionality reduction over sufficiently large and diverse

articulatory image sets typically captures linguistically relevant variation in the new reduced-dimensionality space (Johnson *et al.* 1993; Nix *et al.* 1996), it can be difficult to interpret PCs without further visualization of the uncovered patterns of covariation. SATKit thus includes tools for generating so-called *eigentongues* (Hueber *et al.* 2007) or eigenlarynges for each PC, which visualize the by-pixel covariation captured by each PC in the shape of the basis data as an aid to determining the linguistic relevance of each PC.

Figures 4-5 show a PCA case study from Faytak *et al.* (2020). Figure 4 shows representative raw and filtered tokens of a Mandarin Chinese speaker's [n] and [ŋ] syllable codas and the eigentongue for the first PC for all filtered tokens of these phones; this eigentongue captures variation between the alveolar (deep red pixels) and velar (deep blue pixels) places of articulation. The first two principal components for this speaker (Figure 5, left) capture nasal place (PC1) and coarticulatory influence of the preceding vowel (PC2). The clustering of data in a second speaker's PC1-PC2 space (Figure 5, right) show that the latter completely merges the two nasal codas after /i/ and partially collapses the distinction after non-high vowels, whereas the first speaker maintains the distinction in all environments.

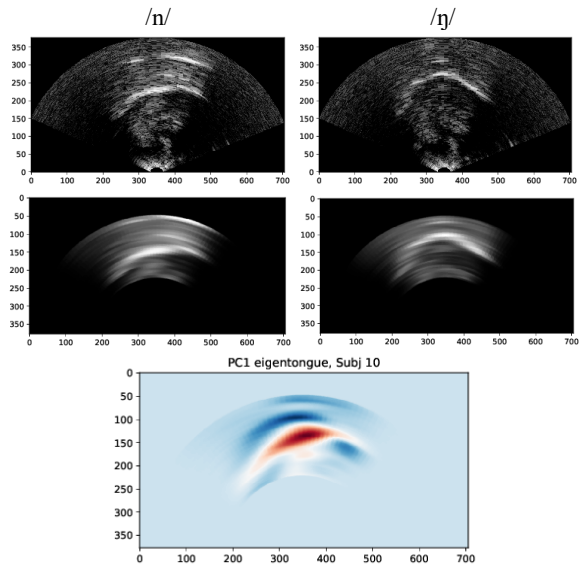

Figure 4: *Representative raw (top) and filtered (middle) tokens of [n] and [ŋ] from one speaker, and the eigentongue which captures variation in pixel brightness in all tokens of both nasals (bottom).*

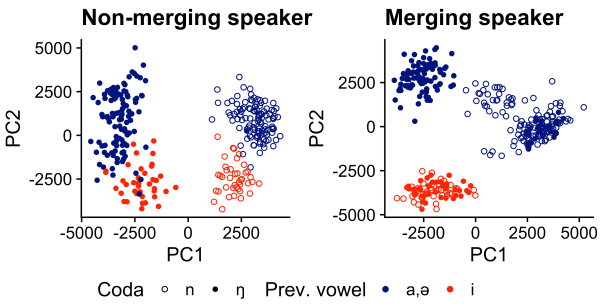

Figure 5: *Reduced-dimensionality spaces for the speaker shown in Figure 4 (left) and a second speaker with coda neutralization (right). Polarity and magnitude of PC scores is arbitrary; axes are arranged for clarity.*

3. Discussion and ongoing development

In this paper, we have presented the Speech Articulation Toolkit (SATKit) and highlighted its methods for whole-image analysis of the pixels in ultrasound imaging data. These methods are intended as a useful complement to contour tracking methods, which are broadly used in the speech sciences at present, but are not well-suited to all research questions that ultrasound data may be collected for. We hope that the common development of several non-contour methods in a single package will facilitate their further use and encourage further validation of their suitability to data beyond what is discussed here. For instance, while we have focused here on applications to ultrasound data, these methods are in theory applicable to other two-dimensional image-based articulatory data (face video, MRI) with extensions that may be included in future updates.

SATKit is under active development and features will be added to all methods described here, in particular the dimensionality reduction utilities. Simple machine learning models for detection of articulatory states based on linear discriminant analysis are likely to be included in a future release, along the lines of those implemented in Carignan (2014) and used in, for example, Mielke *et al.* (2017) and Shaw *et al.* (2020). Furthermore, while contour segmentation and tracking are beyond the scope of this package, we ultimately plan for SATKit to include additional utilities for quantifying contour shape and deformation from imported contour data.

4. Acknowledgements

We thank Alan Wrench for technical advice concerning AAA, and the organizers and attendees of the virtual ISSP for fostering productive discussion in a difficult time.

5. References

Carignan, C. (2014). TRACTUS (Temporally Resolved Articulatory Configuration Tracking of UltraSound) [MATLAB software suite]. Retrieved from http://christophercarignan.github.io/TRACTUS.

Danilouchkine, M., Mastik, F., & van der Steen, A. (2009). A study of coronary artery rotational motion with dense scale-space optical flow in intravascular ultrasound. *Physics in Medicine and Biology*, 54(6), 1397–1418.

Drake, E., Schaeffler, S., and Corley, M. (2013). Does prediction in comprehension involve articulation? Evidence from speech imaging. In *11th Symposium of Psycholinguistics (SCOPE)*, Tenerife.

Eshky, A., Ribeiro, M., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J., & Wrench, A. (2018) Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In *Proceedings of INTERSPEECH 2018*, Hyderabad.

Farnebäck, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. In J. Bigun and T. Gustavsson (Eds.): *Scandinavian Conference on Image Analysis 2003*, 363–370. Berlin: Springer.

Faytak, M., Liu, S., & Sundara, M. (2020). Nasal coda neutralization in Shanghai Mandarin: Articulatory and perceptual evidence. *Laboratory Phonology*, 11(1), 23.

Heyne, M., Derrick, D., & Al-Tamimi, J. (2019). Native language influence on brass instrument performance: An application of generalized additive mixed models (GAMMs) to midsagittal ultrasound images of the tongue. *Frontiers in Psychology*, 10, 2597.

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

Horn, B., & Schunck, B. (1981). Determining optical flow. *Artificial Intelligence*, 17(1), 185–203.

Hueber, T., Aversano, G., Cholle, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., & Stone, M. (2007). Eigentongue feature extraction for an ultrasound-based silent speech interface. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP '07*, Honolulu.

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *JASA*, 94(2), 701-714.

Kochetov, A. (2020). Research methods in articulatory phonetics I: Introduction and studying oral gestures. *Language and Linguistics Compass*, 14(4), e12368.

Koppenhaver, S. L., Hebert, J. J., Parent, E. C., & Fritz, J. M. (2009). Rehabilitative ultrasound imaging is a valid measure of trunk muscle size and activation during most isometric sub-maximal contractions: a systematic review. *Australian Journal of Physiotherapy*, 55(3), 153-169.

Laporte, C., & Ménard, L. (2018). Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. *Medical Image Analysis*, 44, 98-114.

Li, M., Kambhamettu, C., & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics*, 19(6–7), 545–554.

Lin, J., & Moisik, S. (2019). The lingual voice quality settings of Standard Singapore English and Singapore Colloquial English. In Calhoun, S., Escudero, P., Tabain, M., & Warren, P. (eds.), *Proceedings of ICPhS 19*.

McMillan, C. & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition*, 117(3), 243–260.

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

Moisik, S., Lin, H., & Esling, J. H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *JIPA*, 44(1), 21–58.

Nix, D. A., Papcun, G., Hogden, J., & Zlokarnik, I. (1996). Two cross-linguistic factors underlying tongue shapes for vowels. *JASA*, 99(6), 3707-3717.

Palo, P. (2019). Measuring pre-speech articulation. Dissertation, Queen Margaret University.

Palo, P., Moisik, S., & Faytak, M. (2021). SATKit: Speech Articulation Toolkit [Python software suite]. Retrieved from https://github.com/giuthas/satkit.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Poh, D. & Moisik, S. (2019). An acoustic and articulatory investigation of citation tones in Singaporean Mandarin using laryngeal ultrasound. In Calhoun, S., Escudero, P., Tabain, M., & Warren, P. (eds.), *Proceedings of ICPhS 19*.

Ribeiro, M., Sanger, J., Zhang, J., Eshky, A., Wrench, A., Richmond, K., & Renals, S. (2021). TaL: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos. *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*. Shenzhen.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology*, 10(1), 1.

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

Slud, E., Stone, M., Smith, P., & Goldstein Jr., M. (2002). Principal components representation of the two-dimensional coronal tongue surface. *Phonetica*, 59(2-3), 108-133.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, 19(6-7), 455-501.

Vasseljen, O., Fladmark, A., Westad, C., & Torp, H. (2009). Onset in abdominal muscles recorded simultaneously by ultrasound imaging and intramuscular electromyography. *Journal of Electromyography and Kinesiology*, 19(2), e23–e31.

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

Yu, Y., & Acton, S. T. (2002). Speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 11(11), 1260-1270.