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

Matthew Faytak (UCLA), **Pertti Palo** (independent), **Scott Moisik** (NTU Singapore)

Overview. We introduce here the Speech Articulation Toolkit (SATKit), a bundle of open-source Python 3 methods for direct quantitative analysis of the pixels in ultrasound imaging data. The methods are immediately compatible with data collected using Articulate Assistant Advanced, with more general compatibility with any raw scanline or scan-converted data as a short-term goal. The methods are also, in theory, applicable to other image-based articulatory data (video of the external face, MRI) with modifications. We aim for SATKit to provide analysis methods complementary to or in place of contour extraction (e.g. Li et al. 2005).

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 captures the presence of change - including tongue contour movement and changes in interior musculature - in an ultrasound signal and is particularly well-suited to gauging the onset of articulation as a complement to reaction times measured from acoustics. SATKit provides implementations 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 localised measure of change.

Optical flow. Optical flow characterizes the direction and magnitude of apparent motion between a pair of images. For each pair, a vector field is computed that describes the "flow" of pixel brightness patterns (Horn & Schunck 1981). Optical flow captures holistic patterns of motion, provided differences between frames are small, and is especially well-suited to analysis of laryngeal ultrasound (Moisik et al. 2014; Poh & Moisik 2019), where tracking the structures of the larynx is infeasible. SATKit implements an optical flow method similar to that described in Moisik et al. (2014), but using dense optical flow. Vectors in resulting flow fields are averaged to obtain time-varying *consensus vectors* for the entire field or a region of interest. The resulting consensus vectors can be decomposed into velocity signals projected onto horizontal, vertical, or oblique axes. Numerical integration of these signals can be used to estimate displacement of entire structures (Figure 2).

Eigendecomposition. Dimensionality reduction can also be used to extract patterns of covariation in pixel brightness across a set of images (Hueber et al., 2008; Hoole and Pouplier, 2017; Mielke et al., 2017; Lin & Moisik, 2019). This *eigendecomposition* approach is particularly useful for rapid exploratory analysis of the ultrasound signal and in characterizing speaker-specific variation. SATKit implements eigendecomposition using principal component analysis (PCA); utilities are provided for filtering or selecting regions of interest and storing frame data in a standard array format which forms the basis data for the PCA. Additional utilities are provided to reshape and rescale the resulting eigenvectors into eigentongues (Hueber et al, 2008) or eigenlarynges to aid interpretation of the principal components as dimensions of variability. Figure 3 shows representative frames for a speaker's [n] and [ŋ] codas and an eigentongue which captures variation between these two places of articulation.

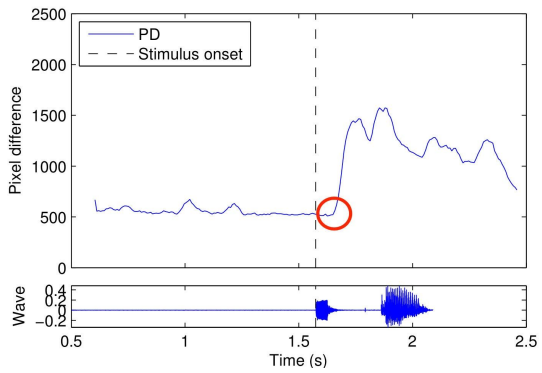

Figure 1: Whole-image pixel difference for an utterance of "caught", with corresponding waveform at bottom. Dashed line indicates go-signal (1kHz beep); onset of articulation (circled) precedes acoustic word onset.

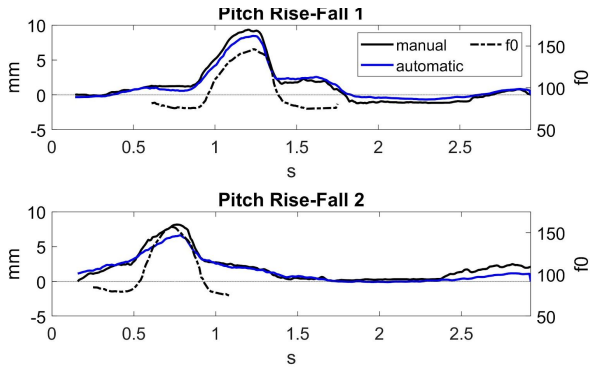

Figure 2: Comparison of larynx displacement signals (and f_o; dashed line) obtained manually using a custom tool in MATLAB 2019a (black line) and automatically with SATKit (blue line).

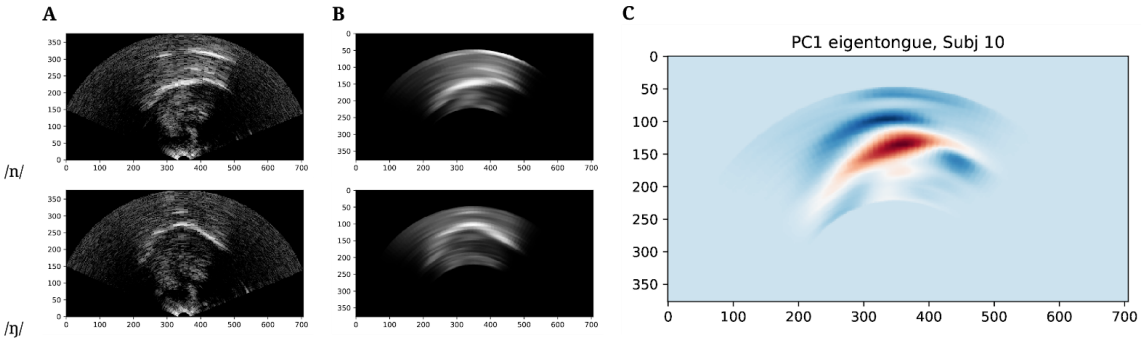

Figure 3: Raw single-trial data (A), filtered and gated data (B), and an eigentongue (C) capturing variation among tokens of [n] (red) and [ŋ] (blue) (Faytak et al, under revision).

References: Faytak, M. et al (under revision). Nasal coda neutralization in Shanghai Mandarin: Articulatory and perceptual evidence. **Hoole, P. & Pouplier, M.** (2017). Öhman returns: New horizons in the collection and analysis of imaging data in speech production research. Comp Speech & Lang 45. **Horn, B. & Schunck, B.** (1981). Determining optical flow. Artificial Intelligence 17(1). **Hueber, T. et al** (2007). Eigentongue feature extraction for an ultrasound-based silent speech interface. ICASSP '07. **Li, M. et al** (2005). Automatic contour tracking in ultrasound images. CL&P 19(6–7). **Lin, J. & Moisik, S.** (2019). The lingual voice quality settings of Standard Singapore English and Singapore Colloquial English. ICPhS 19. **Mielke, J. et al.** (2017). The articulatory dynamics of pre-velar and pre-nasal /æ/-raising in English: An ultrasound study. JASA 142(1). **Moisik, S. et al** (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). JIPA 44(1). **Palo, P.** (2019). Measuring pre-speech articulation. Dissertation, Queen Margaret Univ. **Poh, D. & Moisik, S.** (2019). An acoustic and articulatory investigation of citation tones in Singaporean Mandarin using laryngeal ultrasound. ICPhS 19.