音声研究　第 29 巻　58–71 頁
2025年

Journal of the Phonetic Society of Japan
Vol. 29, pp. 58–71, 2025

特集論文

Quantifying Between-Speaker Variation in Ultrasound Tongue Imaging Data

Takayuki Nagamine*

超音波舌撮像データにおける話者間特性の定量化手法

SUMMARY: This article outlines a quantitative, between-group comparison of tongue shapes using *ultrasound tongue imaging*, one of the vocal tract imaging techniques widely used in articulatory phonetics research. This article first provides a brief overview of ultrasound tongue imaging, followed by a description of a cross-speaker normalisation method based on bite plane rotation. It then outlines ultrasound data recording and analysis workflow with a case study illustrating data analysis using Principal Component Analysis (PCA). This paper demonstrates that the bite plane rotation, coupled with statistical normalisation methods, allows for a reliable tongue-shape comparison by establishing a common coordinate system across speakers.

Key words: articulatory phonetics, ultrasound tongue imaging, Articulate Assistant Advanced, bite-plane rotation, Principal Component Analysis (PCA)

1. Introduction

This article describes *ultrasound tongue imaging*, one of the vocal tract imaging techniques that is widely used in contemporary articulatory phonetics research (Kochetov 2020). Ultrasound tongue imaging is a non-invasive, cost- and time-effective tool that allows for direct access to articulation. In this article, I aim to complement the existing ultrasound tongue imaging tutorials that provide general overviews (e.g., Gick et al. 2008; Wilson 2014) by providing concrete data acquisition and analysis methods and illustrating them through a case study.

The focus of the current article is to demonstrate a quantitative, population-level comparison of multiple speakers' tongue shapes using ultrasound tongue imaging. In the section below, I start by providing the research context and how articulatory data could complement the existing findings. I then provide a discussion of issues and solutions to tongue shape comparison using ultrasound. I then explain a typical workflow of an ultrasound experiment based on the standard practice in our lab, focussing on the experiment preparation and data analysis. Finally, a case study illustrates a quantitative analysis of tongue shape using Principal Component Analysis (PCA).

The foundation of the quantitative analysis presented here is the combination of (1) bite plane rotation and (2) statistical within-speaker normalisation, allowing researchers to align multiple speakers' tongue shape onto a common coordinate system. PCA is a data dimensionality reduction technique that can be useful to capture articulatory dimensions that are salient in the data.

This article assumes ultrasound research using the Articulate Assistant Advanced (AAA) software (version 220.5.1; Articulate Instruments 2022), which is one of the most widely used software for ultrasound research. Data processing, analysis and visualisation are done via R version 4.3.2 (R Core Team 2023). Data and codes used in this article, as well as a list of useful resources, are publicly available in the online supplementary materials at https://osf.io/3sdhf/.

1.1 Research Context

It is widely well-known that L1 Japanese speakers have difficulty producing English /l/ and /ɹ/ accurately. Previous studies show that L1 Japanese speakers' production of English /l/ and /ɹ/ is influenced by or even substituted with the Japanese liquid category /r/, canonically realised as alveolar tap or flap [ɾ] (e.g., Riney, Takada & Ota 2000). This may be because L1 Japanese speakers are less sensitive to the third formant (F3) frequencies, the important acoustic cue in the English liquid contrast (Iverson et al. 2003). Whereas English

* Postdoctoral Research Fellow, Speech, Hearing and Phonetic Sciences, University College London（ポスドク研究員，聴覚・音声言語学研究科，ユニバーシティカレッジロンドン）

/ɹ/, in particular, is characterised by notably low F3 frequencies that provide a reliable acoustic cue to contrast with English /l/, L1 Japanese speakers instead rely on F2 to make a distinction between English /l/ and /ɹ/ in both perception (e.g., Iverson et al. 2003) and production (e.g., Saito & van Poeteren 2018). The reliance on F2 in production suggests that L1 Japanese speakers redeploy articulatory strategies for Japanese /ɾ/ (e.g., front-back dimension) to produce English liquids (Saito & van Poeteren 2018).

Despite a rich amount of previous research, it remains unclear *how* exactly L1 Japanese speakers are different from L1 English speakers in articulation. While acoustic events are the immediate product of adjustments in vocal tract configurations (Iskarous & Pouplier 2022), articulation of English liquids cannot be easily inferred based solely on the acoustic signals. The two canonical configurations for English /ɹ/, 'retroflex' and 'bunched' tongue shapes, for instance, are not readily distinguishable by lower formants (Zhou et al. 2008). It may be the case that English speakers use speaker-specific articulatory strategies to achieve a common acoustic output of low F3 that characterises English /ɹ/, suggesting that it can be challenging to infer the exact articulatory properties for English /ɹ/ from the acoustic signals alone (Mielke, Baker & Archangeli 2016).

Uncovering how L1 Japanese speakers differ from L1 English speakers in the articulation of English /l/ and /ɹ/ is of both theoretical and pedagogical interest, given the persistent nature of difficulty for L1 Japanese speakers in producing English /l/ and /ɹ/. I frame this article around the comparisons of tongue shapes for English /l/ and /ɹ/ between L1 Japanese and L1 English speakers. In the following section, I briefly outline the overview of ultrasound tongue imaging techniques, before explaining the methods in detail.

2. Ultrasound Tongue Imaging

2.1 Ultrasound Research Methods

Ultrasound tongue imaging provides a near-holistic midsagittal image of the tongue with high spatiotemporal resolutions. Typically, an ultrasound probe (or transducer) is placed underneath a participant's chin, which contains a piezoelectric crystal emitting a high-frequency ultrasound (Stone 2005). The ultrasound travels through the tongue and reflects back once it reaches the air just above the tongue surface due to the change in density between the tongue and the air, which then is received by the probe (Stone 2005). This

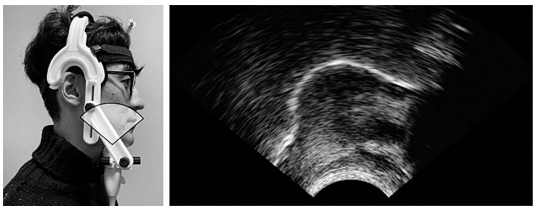

Figure 1 Left: A side-profile view of a participant wearing an ultrasound stabilisation headset with the ultrasound probe placed underneath the chin. The overlaid fan schematises the area scanned by ultrasound. Right: A midsagittal tongue image recorded with the Articulate Assistant Advanced software. The speaker is producing a tongue-tip-up retroflex /ɹ/. Tongue tip points to the right and the tongue posterior to the left. The mandible shadow is imaged as the dark black area towards the right in the fan, obscuring the tongue tip image.

results in a thick white curve just above the tongue surface as shown in Figure 1, enabling us to infer shapes, positions, and movements of the tongue. Note that ultrasound does not travel through hard structures like bones, including the mandible and the hyoid bone, and they appear as a black/dark shadow in the ultrasound tongue images (see the black shadow towards the right edge of the fan in Figure 1).

Ultrasound achieves quite a high sampling rate, which is suitable for capturing quick movements of the tongue. A faster frame rate is necessary to capture fast lingual movements in segments such as alveolar taps or flaps [ɾ]; although machines in previous research achieve approximately up to 30 fps (Derrick & Gick 2011; Yamane, Howson & Po-Chun (Grace) 2015), it is possible with more recent machines to achieve a much higher rate of ca. 80 fps or even greater (Kirkham et al. 2023; Kochetov 2020; Nagamine 2023). A recent co-registration study demonstrates a high degree of accuracy in the spatial and temporal alignment between ultrasound and electromagnetic articulography (EMA; Kirkham et al. 2023).

Furthermore, ultrasound imposes little discomfort for the speaker compared to other existing methods such as EMA and real-time magnetic resonance imaging (MRI), allowing for experiments with a wider range of participant populations. It is possible to obtain a clear image of the tongue by just holding the probe by hand and placing it underneath the participant's chin without any stabilisation measure, which is useful in the field of speech therapy when restraints need to be

minimised as much as possible or probe stabilisation is not practically feasible (Klein et al. 2013; Preston et al. 2017). Ease in visualising the tongue is also an advantage in pronunciation teaching classrooms, in which time and resources are often limited. An increased number of studies report the benefit of ultrasound visual feedback in L2 pronunciation teaching and learning (e.g., Antolík, Pillot-Loiseau & Kamiyama 2019; Bryfonski 2023).

Finally, the increased portability of the recent ultrasound machines enables us to conduct articulatory research in the field with a participant population that is not easily accessible through laboratory-based experiments. The ultrasound machine can be only slightly larger than a smartphone, such as the Telemed MicrUS system that operates with the Articulate Assistant Advanced software. A recent study demonstrates its utility in fieldwork for ultrasound data collection as part of a social sciences festival at a local market in a small town in the UK (Nance et al. 2023).

To summarise, ultrasound is a non-invasive, relatively easy approach to studying lingual articulation compared to other existing methods. The high spatio-temporal resolution and increased portability are particularly suitable for a wide range of recording settings outside laboratories, including language classrooms and in the field. Despite the ease of data acquisition, however, data processing and analysis usually require an extensive amount of time, effort and consideration, as discussed below.

2.2 Probe Stabilisation

The rich dimensionality of midsagittal tongue surface data obtained from ultrasound, as opposed to flesh-tracking methods such as EMA, introduces greater demands at the data processing phase. While a holistic midsagittal view of the tongue surface can visualise both global and local lingual movements, it is often quite challenging to partition different parts of the tongue and identify linguistically meaningful movements based solely on ultrasound images (Davidson 2006; Stone 2005). In addition, a lack of fixed anatomical structure in ultrasound images makes it challenging to infer the exact position of the tongue in the vocal tract (Gick et al. 2008; Stone 2005). Individuals differ substantially from one to another in their vocal tract anatomy, including the length and width of the tongue as well as articulatory strategies to produce certain sounds, including English /ɹ/ (Slud et al. 2002). These overall suggest that it is not appropriate to directly compare tongue shapes obtained from different subjects, making it difficult to perform a simple cross-speaker comparison.

Alternatively, a fruitful approach to articulatory investigation using ultrasound is to compare *within-speaker* articulatory strategies across multiple subjects, given the remarkably high degree of within-speaker consistency found in speech production (Johnson, Ladefoged & Lindau 1993). As a consequence, previous ultrasound research attempts to quantify measures that capture within-speaker variability in midsagittal tongue movement. Strycharczuk and Scobbie (2017), for example, quantify the degree of GOOSE-fronting in British English by measuring the tongue position for /uː/ relative to the reference vowel /iː/ in each speaker. Similarly, Kirkham and Nance (2017) quantified the degree of tongue root advancement in each subject to investigate how Twi-English bilingual speakers realise the [ATR] and [TENSE] features in their vowel production. These derived measurements can be statistically normalised (e.g., using *z*-scores) within each speaker, facilitating cross-speaker comparison in subsequent statistical analyses.

In addition to the considerations above, reliable quantitative analysis of tongue contours using ultrasound is based on at least two implicit assumptions. First, at the data acquisition stage, it is crucial to ensure that the probe is stabilised so that it does not move substantially within one recording session (Gick et al. 2008). While tongue images can be easily obtained by a hand-held probe underneath the speaker's mandible, this inevitably imposes additional noise in data due to the movement of the probe itself, making it impossible for researchers to tease it apart from tongue movement (Derrick et al. 2018; Stone 2005). Quantitative analysis attempts to minimise measurement errors by obtaining multiple observations and employing statistical tests, and it is therefore important to ensure that multiple observations from multiple speakers recorded with ultrasound can be reliably compared.

One way of minimising measurement errors in ultrasound recording is to stabilise the probe relative to the head wherever possible. Various stabilisation techniques have been proposed and used, including a stand on the table (Stone 2005), a headrest on which the participant could rest their head (Derrick et al. 2018), and a wearable headset (Spreafico, Pucher & Matosova 2018). The Haskins Optically Corrected Ultrasound System (HOCUS) combines ultrasound with optical tracking to correct the probe movement relative to the participant's head movement (Whalen et al. 2005). Recent stabilisation headsets are made

of light plastic instead of more rigorous materials like metals (Articulate Instruments 2008) and thus impose relatively small degrees of discomfort on the research participants who wear the headset while maintaining measurement accuracy (Pucher et al. 2020; Spreafico, Pucher & Matosova 2018). While the need of probe stabilisation may cancel out the benefit of the ease in imaging tongue shape (see Section 2.1) and some headsets (e.g., Ultrafit) constrain the participant's jaw movement to some degree, the discomfort that a plastic headset may impose on participants is fairly minimal considering that they are free to move their head. An example of a plastic headset is seen in the left picture in Figure 1.

2.3 Establishing a Common Coordinate System across Speakers

Once probe stabilisation has been achieved, another consideration is to align tongue splines in a common coordinate system across speakers and repetitions relative to fixed, passive articulators (Scobbie et al. 2011; Scobbie, Stuart-Smith & Lawson 2012). While this process is not necessary when the researcher's interest lies solely in comparing tongue shapes irrespective of the tongue position, such as Dorsum Excursion Index (DEI; Zharkova 2013), establishing a common coordinate system across speakers' tongue splines aids linguistically-meaningful interpretations of tongue movements regarding magnitude, orientation, location and relative timing (Scobbie et al. 2011; Westbury 1994). The coordinate transform is incorporated in a common workflow in EMA, in which the locations of the lingual sensors can be normalised across speakers by the use of reference sensors attached to the bite plate, as well as the speaker's nasion, upper and lower incisors and both mastoids (Rebernik et al. 2021). Similarly, the holistic midsagittal view of the entire vocal tract captured by real-time MRI provides many possible structures that could be used as reference points, including the upper and lower incisors (e.g., Maekawa 2023).

Ultrasound, on the other hand, does not usually provide relative positional information of the tongue in the vocal tract because it does not record as many anatomical landmarks that can serve as reference points as other methods such as EMA or MRI (Gick et al. 2008; Stone 2005; Zharkova 2013). In addition, the probe angle is usually determined in order to ensure clear visibility of the tongue surface, meaning that the horizontal and vertical axes from the ultrasound images bear no consistent linguistically-meaningful interpretations such as 'frontness' or 'height' of the tongue (Scobbie, Stuart-Smith & Lawson 2012). Stone (2005) proposes the use of a dental cast by taking each speaker's dental impression and using it as a reference to determine the probe position for multiple recording sessions for a single speaker. This method, however, relies heavily on the researcher's qualitative assessment of the probe position, which may suffer from reduced precision in probe placement. Although it is also possible to rotate the tongue curves based on the horizontal axis defined by the tongue positions for the peripheral vowels, this may add extra difficulty in interpreting the resulting diagrams that may be inconsistent with findings based on EMA data (Scobbie, Stuart-Smith & Lawson 2012).

A possible and practical solution to the lack of reference points in ultrasound is to establish a *post-hoc* common coordinate system to normalise tongue positions across speakers using the speaker's bite plane, obtained through the use of a simple thin plastic plate (seen in the top left image in Figure 2). The plastic plate is 40 mm wide and approximately 60 mm long with a 2-mm thickness. The plate, made of biocom-

Figure 2 **Top left: A bite plate made of biocompatible plastics, provided to Lancaster Phonetics Lab by courtesy of Dr Eleanor Lawson (University of Strathclyde). Bottom left: A speaker biting the bite plate by inserting the end 'A' in the mouth. The dashed circle indicates where the upper incisor makes contact against the barrier (i.e., 'B' on the bite plate). Right: Bite plane measurement superimposed on an ultrasound image. The arrow indicates the point at which the tongue shows deformation in shape (corresponding to 'A' in the top left image). Bite plane measurement traces the flat tongue surface observed anterior to the deformation point. Note that the label 'B' in the right image is in parenthesis because it does not indicate the exact location of the incisors as it was not measured precisely here.**

patible plastics by Dr Eleanor Lawson at University of Strathclyde, has a small 'hump' at approximately 45 mm from the end that barriers the upper front teeth, which maintains consistent across speakers the distance between the upper front teeth and the end of the plate inside the mouth ('B' in the top left image in Figure 2). The speaker bites this plate by inserting the longer end of the plate into the mouth (i.e., 'A' in the top left image in Figure 2) to the extent that the upper front teeth make contact with the barrier (illustrated in the bottom left image in Figure 2). When the participant pushes their tongue against the plate, the bite plane can then be imaged in ultrasound as a flat surface from the point where the tongue shape shows some deformation to the tongue anterior (see the image on the right in Figure 2).

The bite-plane normalisation involves 'offsetting' and 'rotating' the tongue contours against the bite plane once recording is completed and tongue contours are estimated/tracked. Bite plane offsetting and rotation is illustrated in Figure 3. Here, as an example, I compare the midsagittal tongue shapes of the English vowel /æ/ in the word *ham*, extracted at the vowel midpoint, produced multiple times by two speakers: one L1 Japanese speaker (Speaker A) and one L1 English speaker (Speaker B).

The tongue shapes and the bite planes in the solid line in Figure 3 show *offsetting* in which the origin of the coordinate system is aligned at zero (i.e., the point where the bite plate meets the tongue in the mouth, causing tongue deformation). At this point, the angle of the bite plane is still different between the two speakers, suggesting that it is not possible to perform reliable interpretations of tongue shape along conventional dimensions such as 'tongue height' or 'tongue retraction'. The dotted lines, on the other hand, show the tongue

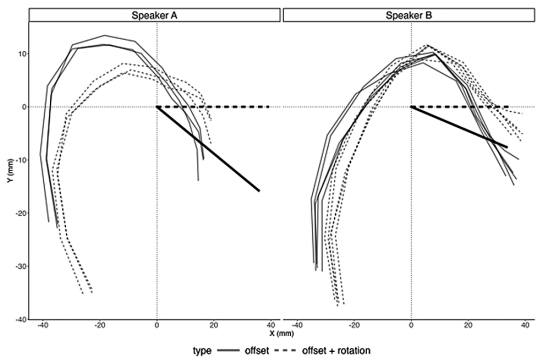

Figure 3 Example tongue contours after offsetting (solid line) and offsetting/rotating (dotted line). Thicker lines represent each speaker's bite plane.

contours that are both *offset* and *rotated* against the bite plane, in which not only the coordinate origin is aligned but also the bite plane is rotated to be horizontal. This way, a common coordinate system can be established for both speakers with shared *x*- and *y*-axes (Scobbie et al. 2011).

Note that AAA has various options for exporting the tongue contours, including exporting the *x/y* coordinates with/without offsetting and rotating as well as scaling the size of the tongue relative to the length of the measured bite plane being one. The origin of the tongue position offsetting/rotation also needs to be decided between *knot 0* and *knot 1*. *Knot 0* corresponds to the point where the dotted and solid lines meet in the L-shaped fiducial spline that can be defined on the AAA software (i.e., the point 'A' in the right image in Figure 2). *Knot 1*, on the other hand, is the other end of the solid line. It is possible to define the point 'B' as the speaker's incisor location, although it was not measured precisely in Figure 2. A quick illustration of these exporting options is provided in the online supplementary materials.

During the data collection, the participant's bite plane can be recorded by asking them to bite the plastic bite plate by inserting the longer end (i.e., 'A' in the top left image in Figure 2) into the mouth and placing the upper incisors at the small bump on the plate that acts like a stopper (i.e., 'B' in the top left image in Figure 2; see also the bottom left image in Figure 2). Participants then push their tongues up against the flat surface under the plate, which visualises a flat occlusal plane with some deformation of the tongue (see the right image in Figure 2).

At the data analysis stage, the tongue contours need to be offset and rotated relative to the bite plane for each participant for cross-speaker tongue shape comparisons (e.g., Strycharczuk & Scobbie 2017). In AAA, this can be implemented by defining the bite plane tracing as the 'fiducial' (i.e., reference) template, superimposing it to all ultrasound frames (so that both estimated tongue contours and bite plane tracing appear simultaneously for all ultrasound frames), and selecting 'offset and rotate' in the menu window for exporting the data. Fiducial templates can be defined at the data processing stage, described further in Section 3.2 below.

To summarise, ultrasound offers researchers access to lingual articulation easily because of its non-invasiveness, high spatiotemporal resolution and portability. Processing midsagittal tongue images obtained with ultrasound, however, requires some methodological considerations. In addition to the need for head stabili-

sation during recording, it is suggested that a common coordinate system be established across speakers using bite plane measurement for each speaker as a way of cross-speaker normalisation. More information about the bite-plane rotation can be found in Scobbie et al. (2011) and Strycharczuk and Scobbie (2017). See also Scobbie, Stuart-Smith and Lawson (2012) for more detailed discussions on the tongue curve rotation methods.

3. Example Ultrasound Recording and Analysis Workflow

As shown in the previous sections, a reliable cross-speaker comparison can be facilitated by incorporating probe stabilisation and bite-plane measurement in an ultrasound workflow. In this section, I provide further details on these procedures by showing an example workflow for ultrasound recording and providing detailed explanations for considerations that need to be made.

3.1 Data Recording

This section mainly outlines the experiment preparation stage given its importance for a reliable data analysis. This includes headset fitting, probe position adjustment, recording parameter setting and bite plane measurement.

3.1.1 Fitting Probe Stabilisation Headset

Once participants arrive at the recording venue and complete all necessary paperwork and briefing procedures, I fit the ultrasound headset on the participant's head in order to stabilise the ultrasound probe relative to head movement. I use UltraFit, a light-weight plastic ultrasound probe stabilisation headset (see Figure 1) that is commercially available from Articulate Instruments Ltd. (Spreafico, Pucher & Matosova 2018). I make sure that the headset is fitted straight and tight enough for the probe to be stabilised but not too tight for participants to feel discomfort.

3.1.2 Adjusting Probe Position

Once the headset is successfully fitted, the position of the probe needs to be adjusted. I apply the ultrasound gel to the probe surface at this point so that the probe makes firm contact with the lower chin, as any pocket of air between the probe and the participant's chin can make the image quality poorer.

The probe position adjustment is carried out in the following order. First, I determine the probe position along the midsagittal plane by adjusting it from the participant's front. Second, looking at the probe from the side, I adjust the probe angle; in most cases, this is to make sure that the probe points straight up, which usually results in a good image quality provided a proper field-of-view setting (see Section 3.1.3 below). At the same time, I adjust the height of the probe so that the probe makes direct contact with the participant's chin, when the Articulate Assistant Advanced (AAA) software should start displaying the midsagittal tongue image.

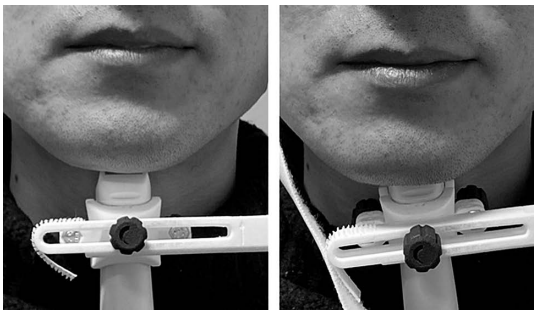

Figure 4 Comparisons of probe angle seen from front. While the probe stands straight up in the left picture, it is clearly tilted pointing off midsagittal in the picture to the right, caused by lifting the probe too much.

Note that it is also necessary at this point to check from the participant's front again whether the probe still points straight up without any tilting, as a tilted probe does not scan a midsagittal tongue shape appropriately, especially when the tongue goes far from the probe origin for e.g., a high front vowel /i/. When the probe is found to be tilted, the probe is likely being lifted up too much and thus applying too much pressure against the participant's chin, which is illustrated in Figure 4.

Although how clearly the tongue can be imaged varies from one participant to another, a rule of thumb in gauging the optimal probe angle is to check the position of the mandible and hyoid bone, both of which appear as black shadows in ultrasound images (Preston et al. 2017; see also the dark structures on both fan edges in Figure 2). When the tongue is recorded with the tongue tip pointing to the right, the mandible shadow is seen towards the right of the fan whereas the hyoid bone shadow is seen towards the left. It is suggested to orient the probe so that both shadows appear on each edge of the fan, or at least one of the shadows, as these structures are the few of the reference anatomical landmarks that can be imaged in ultrasound images (Preston et al. 2017).

3.1.3 Recording Parameter Setting

At this stage, recording parameters may also need to be adjusted on the AAA software. Crucial settings include *field of view* (FOV) and *depth*. A larger FOV results in a wider fan-shaped window capturing a wider area of midsagittal tongue surface. FOV is in a trade-off relationship with the frame rate, where a larger FOV typically results in reduced frame rate (Stone 2005). AAA software achieves approximately 80 fps even with a 100% FOV setting (corresponding to a 101.2° FOV using a 20-mm radius convex probe in our Telemed MicrUS system), which should be adequate for capturing most of the lingual sounds including alveolar taps/flaps [ɾ] (Recasens & Rodríguez 2016).

The depth setting determines how far the ultrasound travels from the probe and how long the probe needs to wait to receive the reflected ultrasound (Stone 2005). In short, this changes the size of the tongue for each participant displayed on the screen. I typically use the depth setting of 80 mm, which is optimal for most participants. For some participants, however, such as tall male speakers who have larger anatomical structures, I sometimes need to adjust the depth to 90 mm or even larger. Appropriate depth setting can be gauged by checking how well ultrasound captures the tongue shape for the high front vowel /i/.

Finally, before recording the stimuli, participants' bite plane can be recorded (see Section 2.3. for details). The palate shape can also be recorded by having the participant swallow water at this stage. Overall, all these preparatory procedures should take up to 15 to 20 minutes depending on the researcher's experience and the anatomical characteristics of the participants. Note that it might not be possible to obtain the best quality images across all speakers, as it partly depends on factors related to participants such as the size of the lower jaw, allowing for a proper probe placement, as well as the presence of a beard.

3.2 Tongue Contour Tracing/Estimation

Ultrasound data obtained through the AAA software typically involves multiple short recordings of midsagittal tongue movement and synchronised audio recordings. While it might be possible to conduct subjective and qualitative judgements on tongue movement solely based on visual inspections of ultrasound images, it is often desirable to analyse the tongue data quantitatively through various data visualisation and/or statistical methods. For this, an ultrasound data analysis workflow usually begins with tongue contour estimation/tracking to allow for further quantitative analysis at the later stages.

Tongue shape analysis is usually conducted on tongue contours extracted from ultrasound images. Common software includes GetContours (Tiede 2021) and EdgeTrack (Li, Kambhamettu & Stone 2005). Most of the tongue contour extraction methods are implemented while relying on either the edge detection technique, in which pixel differences (i.e., differences in brightness) are used to estimate the tongue surface, or the whole image processing technique, in which whole ultrasound images are compressed into several numeric values via directly applying data dimensionality reduction techniques to ultrasound images (Wrench & Balch-Tomes 2022). Edge detection based on pixel differences is, however, prone to substantial estimation error due to the presence of noise or blurred images where tongue contours cannot be reliably determined (Wrench & Balch-Tomes 2022). Similarly, it is not immediately easy to interpret the results of the whole image processing technique, especially in understanding what each of the dimensions represents in light of the original midsagittal tongue dimension.

One recent approach for faster and more reliable tongue contour estimation involves the *DeepLabCut* (DLC) toolkit (Mathis et al. 2018). DLC is a markerless pose estimation algorithm that has been used to estimate the position and movement of different body parts based on deep neural networks (Mathis et al. 2018). DLC can be implemented via the AAA software for ultrasound tongue surface estimation, and this achieves a faster and more accurate tongue surface estimation compared to the existing features in AAA or other software (Wrench & Balch-Tomes 2022).

The current DLC models implemented in AAA estimate 11 key points along the tongue surface for each ultrasound frame, with two points capturing each of the key areas of the tongue surface including tongue root, tongue body, tongue dorsum, tongue blade and tongue tip as well as the epiglottic vallecula (i.e., a small depression between the root of the tongue and the epiglottis). Although some parts of the tongue (e.g., tongue tip and tongue root) can be difficult to see in ultrasound images when they are obstructed by hard structures, DLC *estimates* (rather than *tracks*) the position of these obscured parts based on the rest of the tongue available in the image using pre-trained deep neural networks (Wrench & Balch-Tomes 2022). The 11 sets of *x*/*y* coordinates are then exported for further data processing and statistical analysis. The bite plane information is expressed in the same manner, such that the origin and the end of the bite plane are expressed using the *x*/*y* co-

ordinate, although the bite plane does not usually need to be exported for data analysis.

4. Case Study

In order to demonstrate further stages of a typical ultrasound data analysis workflow, I present a case study involving a brief ultrasound analysis in which I aim to answer the research question in the context of the current paper. The research question is '*How do L1 Japanese speakers differ from L1 English speakers in articulation for English /l/ and /ɹ/?'* Codes and data to reproduce this analysis are publicly available in the OSF repository at https://osf.io/3sdhf/.

4.1 Participants

The data to be analysed here are obtained from 38 speakers, including 12 L1 North American English speakers (10 female, 2 male), with a mean age of 29.7 years (SD = 6.05) and 26 L1 Japanese speakers (16 female, 10 male), with a mean age of 19.7 years (SD = 1.05). This is a subset of a larger corpus consisting of 55 speakers (41 L1 Japanese speakers and 14 L1 North American English speakers), chosen on the basis of clarity of ultrasound tongue image. L1 North American English speakers grew up either in the US (n = 8) or in Canada (n = 4) using English primarily up until the age of 13, all of whom identify as fluent L1 speakers of North American English. Three Canadian English speakers are originally from Ontario, two of whom reported that they speak French fluently. One Canadian speaker was originally from Poland, with reported fluency in French and Polish. American English speakers came from all over the country without a particular concentration of their origin, two of whom reported that they fluently speak French and Chinese (without specifying particular varieties) respectively. One American English speaker reported working fluency in eight additional languages. All L1 English speakers resided in the UK at the time of the recording for their work and postgraduate study.

All L1 Japanese speakers were undergraduate students enrolled at universities located near the cities of Kobe and Nagoya in Japan at the time of recording. As a result, the majority of them were originally from Aichi (n = 9) or Hyogo (n = 7) while the rest came from both Eastern (e.g., Chiba and Shizuoka) and Western (e.g., Kagawa, Shimane, Osaka, Kagoshima, Fukui, Gifu and Shiga) Japan. They all reported that they were monolingual Japanese studying English as a foreign language in Japan who receive instruction primarily through the school curriculum. The mean length of their English study was 9.19 years (SD = 2.02). They did not have extensive experience in study abroad, with the mean length of overseas experience being 0.58 months (SD = 1.08). One of them reported that she speaks Korean fluently, whereas eight others reported experience in studying additional foreign languages including Chinese, French, Korean and Swedish. No participants reported any hearing or speaking impairments.

4.2 Procedure

The experiments took place between October and December 2022. Each session consisted of a speech production experiment involving ultrasound recording and a speech perception experiment. Audio and midsagittal tongue images were recorded simultaneously in a quiet room at universities in Japan for L1 Japanese speakers and in a sound-attenuated room in the UK for L1 North American English speakers. Audio signals were pre-amplified and digitized at 44.1 kHz with 16-bit quantisation using a Sound Device USB-Pre2 audio interface. Ultrasound data were recorded using the Telemed MicrUS system with a 20-mm radius convex probe, synchronised with the audio signals and recorded on a laptop computer via the AAA software version 220.5.1 (Articulate Instruments 2022). The recording parameters were consistent within-speaker but varied across speakers to obtain the most optimal tongue images within a range of probe frequency between 2–4 MHz, depth between 80–90 mm and field of view settings between 80–100% (91.6°–101.2° FOV), resulting in a frame rate of ca. 80 frames per second.

Ethics approval was obtained from Lancaster University, Kobe Gakuin University and Meijo University. All participants were compensated for their time and participation with 2,000 Japanese Yen or 15 British Pound Sterling in the form of cash or vouchers commensurate with the regulations at each of the recording venues.

4.3 Data Processing

The data to be analysed here are the tongue shapes for intervocalic English /l/ and /ɹ/ from a minimal pair *'believe'* and *'bereave'*, produced between three to five times by the 38 speakers mentioned above. This results in a total of 297 tokens obtained from L1 North American English speakers (n = 53 for /l/; n = 57 for /ɹ/) and L1 Japanese speakers (n = 94 for /l/; n = 93 for /ɹ/).

While it can be difficult to image the high front vowel /i/ using ultrasound due to the distance between the ultrasound probe and the tongue surface, the two

English words were chosen nevertheless to allow for a cross-linguistic comparison of liquid consonant articulation in intervocalic position between English and Japanese as presented in Nagamine (2023). In fact, including high vowels provides an additional advantage that the clear fronting/raising movement of these vowels makes it easier for us to assess the degree of tongue retraction/lowering often involved in English liquid articulation.

Segmentation was automatically conducted at the phone level via Montreal Forced Aligner (MFA) version 2.0.6 (McAuliffe et al. 2017) and it was then adjusted wherever necessary using Praat (Boersma & Weenink 2022). MFA performs overall well on tokens in which the liquid consonants were realised as approximants, but less so when English /l/ and /ɹ/ were substituted by an alveolar tap [ɾ] in some of L1 Japanese speakers' data.

Following the data processing protocol above (see Section 3.2), I estimated tongue contours using DLC/AAA and tracked each speaker's bite plane, which was then superimposed onto all the ultrasound frames. I then exported the offset/rotated tongue contours within a defined interval. Tongue shapes were extracted at 11 equidistant points during the vowel~liquid~vowel intervals in these words (i.e., ***b<u>elie</u>ve*** and ***b<u>erea</u>ve***), with 0% corresponding to the onset of the first vowel /ɪ/ and 100% to the offset of the second vowel /i/, at 10% increments.

No hand correction was performed to the tongue contour estimations on the AAA software as DLC achieves a reasonably high accuracy in tongue contour estimation as long as the whole tongue shape is clearly imaged. For this study, I estimated tongue contours using the DLC ResNet50 model, which, in theory, can achieve tongue contour estimation accuracy equivalent to that of human coders (Wrench & Balch-Tomes 2022). DLC applies tongue contour estimation on a frame-by-frame basis without considering temporal continuity between frames; this, together with its capability of accurate tongue contour estimation, prevents the common problem of 'drifting' in tongue contour tracking (Wrench & Balch-Tomes 2022).

Rather than manually correcting tongue contours, therefore, tokens exhibiting tracking errors were simply removed from the analysis. Tongue contour tracking errors were identified through visual inspections of the ultrasound images displayed in the AAA software and data visualisation using R. This study only includes tongue shapes from speakers whose tongues were clearly imaged. The tongue plots generated in R (included in the online supplementary materials) further confirm that no tokens show significantly erroneous tongue contour estimations. After these considerations, I retained 38 speakers in this study, excluding 17 speakers from the initial pool of 55 who participated in recording.

4.4 Data Analysis

In this case study, I describe a data analysis procedure using Principal Component Analysis (PCA). PCA is a data dimensionality reduction technique that can extract a small number of major abstract patterns based on correlations between data points in the raw articulatory data (Johnson 2008; Mokhtari et al. 2007; Stone 2005; Turton 2017). PCA is particularly useful for ultrasound tongue imaging, in which systematically identifying main lingual variation has been a major challenge in data analysis (Davidson 2006), and it has been used in previous ultrasound research to quantify articulatory characteristics of lateral allophony across dialects of British English (Turton 2017), palatalised and non-palatalised consonants in Scottish Gaelic (Nance & Kirkham 2022), and vowel-to-vowel coarticulations of consonants in German (Hoole & Pouplier 2017).

PCA identifies orthogonal axes along which the greatest amount of variance can be found in the data, resulting in *eigenvectors* (principal components: PCs) expressing the direction of the variation, and *eigenvalues* indicating the amount of variance for each eigenvector (Hoole & Pouplier 2017). These values are then used to compute *PC loadings* and *PC scores*, with the former expressing the relative weighting of each PC and the latter showing how much each PC can be associated with each token in the data set (Johnson 2008).

Another advantage of PCA is that the identified PCs can be projected back to the original measurement unit; in this case, it is possible to show variations associated with each PC on the midsagittal tongue shape (Johnson 2008). While the derived PCs are abstract, the reconstructed midsagittal tongue shapes facilitate linguistically meaningful interpretations of tongue movements on the front-back or high-low dimensions (Johnson 2008; Stone 2005).

The current study mostly follows the protocol for the PCA analysis in Nance and Kirkham (2022), whose analysis codes are also publicly available. Prior to the PCA analysis, the *x*/*y* coordinate values along the tongue contours are within-speaker *z*-normalised to facilitate cross-speaker comparisons. I then compute PC scores based on all tokens (i.e., *x*/*y* coordinates

for all timepoints) for all speakers using the *princomp* function in R. To interpret the PCs, I reconstruct the variations expressed with each PC based on the PC loadings and the standard deviation scores for each PC score against the mean tongue curve. Finally, I compare the PC scores between English /l/ and /ɹ/ across the two speaker groups. Because visual inspections of spectrograms suggested that the English liquids occurred at approximately the 30% point during the vowel~liquid~vowel interval, I report the PC scores at the 30% time point as a proxy of tongue shapes for English /l/ and /ɹ/ in the following results section. The codes for this analysis are available in the online supplementary materials.

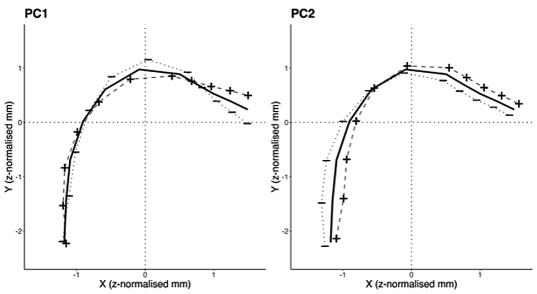

Figure 5 Reconstructed midsagittal tongue shape based on variation explained by PC1 (tongue dorsum lowering: left) and PC2 (tongue fronting: right).

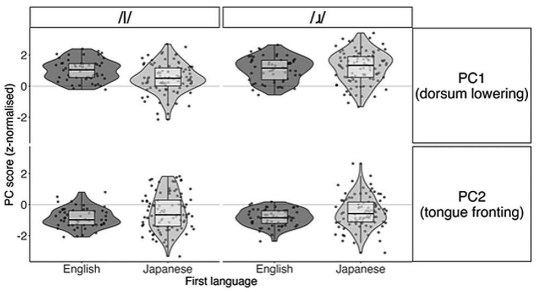

Figure 6 Violin plots and individual data points showing distributions of the PC scores. Zero corresponds to the mean tongue in Figure 5 and is indicated in a thin horizontal line.

4.5 Results

The PCA analysis identifies four principal components (PCs) that account for over 5% of variance in the data: PC1 (44.95%), PC2 (20.86%), PC3 (9.94%) and PC4 (8.34%), with a cumulative sum being 84.09%. The four PCs are retained here because the proportion of variance exceeds the threshold of 5%, following the recommendation by Baayen (2008). PC1 represents variation in tongue dorsum, whereas PC2 in tongue fronting. PC3 captures variation around the tongue posterior and tongue root, and PC4 captures a slight variation in tongue front. Although PC1 also seems to suggest tongue tip raising, it is unclear at this stage whether this reflects actual tongue movement along this dimension or if it is due to measurement noise, arising from difficulty in clearly imaging tongue tip using ultrasound. Due to the limitation of space, the current analysis focusses on the first two PCs only. Visualisation of reconstructed tongue shapes along each dimension is available in the online supplementary materials.

In order to interpret the dimension captured by the first two PCs in a linguistically meaningful way, tongue curves have been reconstructed showing the maximum and minimum tongue shapes represented by each PC in Figure 5. The thick contour in Figure 5 represents the mean tongue, with the variation in tongue shape expressed by adding and subtracting the value of each loading multiplied by the standard deviation of PC scores. The left panel in Figure 5 suggests that PC1 captures tongue dorsum lowering, in which higher PC1 values represent lower tongue dorsum. The right panel in Figure 5 shows the variation captured by PC2, corresponding to the degree of tongue fronting: Higher PC2 values indicate more advanced tongue positions.

Finally, the distributions of the PC scores are juxtaposed in Figure 6 for L1 English speakers (dark grey) and L1 Japanese speakers (light grey) by segment (column) and PC (row). Overall, although the two speaker groups show similar mean PC scores, L1 Japanese speakers' distribution is wider than that of L1 English speakers, suggesting a more variable tongue shape for L1 Japanese speakers. Along the PC1 dimension, L1 English speakers show positive PC1 values ($M = 1.02$, $SD = 0.68$ for /l/; $M = 1.09$, $SD = 0.78$ for /ɹ/), suggesting that their tongue dorsum is lowered for both /l/ and /ɹ/ (i.e., corresponding to the '+' curve in the left panel in Figure 5). L1 Japanese speakers, on the other hand, show a slightly lower mean PC1 score with a wider distribution for English /l/ ($M = 0.52$, $SD = 0.92$) than L1 English speakers. For English /ɹ/, while the two speaker groups are similar in the mean PC1 value, L1 Japanese speakers show a slightly wider distribution judging from Figure 6, and the standard deviation ($M = 1.21$, $SD = 1.02$).

A similar tendency can be seen for PC2. For both English /l/ and /ɹ/, L1 English speakers show negative PC2 values ($M = -0.85$, $SD = 0.63$ for /l/; $M =$

−0.84, *SD* = 0.55 for /ɹ/), indicating that their tongue shape is retracted along the tongue posterior and lower along the tongue anterior (i.e., the '−' curve in the right panel in Figure 5). Although L1 Japanese speakers show negative mean PC2 values similar to L1 English speakers, the PC2 distributions are much wider, indicating a more variable tongue shape (*M* = −0.49, *SD* = 1.19 for /l/; *M* = −0.49, *SD* = 1.02 for /ɹ/).

To summarise, the PCA analysis identifies four major components of the tongue shape for English /l/ and /ɹ/, with the first two PCs explaining approximately 66% of the variation in the data. PC1 corresponds to tongue dorsum lowering whereas PC2 can be interpreted as tongue fronting. Visual inspection of the distributions of the PC scores suggests that L1 English speakers' articulation is more consistent than L1 Japanese speakers, with an overall tendency of a lower and retracted tongue shape for L1 English speakers.

5. General Discussion and Conclusion

The objective of this article is to illustrate a workflow of ultrasound tongue imaging analysis, guided by the research context of L1 Japanese speakers' production of English /l/ and /ɹ/. The research question asks what articulatory differences can be identified between L1 English and L1 Japanese speakers. I demonstrate that an appropriate data collection protocol combined with bite-plane rotation and Principal Component Analysis (PCA) facilitates a systematic quantitative articulatory analysis.

The main findings from the case study, visualised in Figure 6, demonstrate that L1 Japanese speakers exhibit greater variability in articulatory strategies for English /l/ and /ɹ/ in terms of (1) tongue dorsum height and (2) tongue retraction. L1 English speakers, on the other hand, show consistently lower tongue dorsum and retracted tongue shape. This finding agrees with the previous descriptions of English liquid articulation that commonly involves tongue retraction (Alwan, Narayanan & Haker 1997; Narayanan, Alwan & Haker 1997; Stevens 2000).

This paper shows that a combination of bite plane rotation and PCA is a powerful tool for quantitative analysis of midsagittal tongue shape using ultrasound. It is often challenging to systematically partition regions of the tongue surface due to the lack of anatomical landmarks in ultrasound images despite the rich dimensionality of the data (Davidson 2006; Stone 2005). Although it is possible to rely on metrics that do not depend on the tongue position for data analysis (e.g., Curvature Index, Stolar & Gick 2013; Dorsum Excursion Index, Zharkova 2013), it is often more meaningful to be able to explain articulatory variation in the data in phonetic dimensions such as 'front-back' or 'high-low' (Scobbie et al. 2011; Stone 2005). The PCA identifies the major variation in the data in a data-driven manner and the identified principal components can be used to reconstruct the tongue shape as illustrated in Figure 5.

The flexibility in subsequent statistical analysis is one of the strengths of PCA. Given that PCA summarises the dimension of data variability into numeric values (PC scores), as shown in Figure 6, further statistical analysis can be conducted using e.g., linear mixed-effect models (Nance & Kirkham 2022) to formally test the effect of variables of interest. In the context of the current analysis, the PC scores along each dimension can be a dependent variable, predicted by fixed effects including the language group (two levels: L1 English vs L1 Japanese), segment (two levels: /l/ vs /ɹ/), and the interaction between them. As suggested by an anonymous reviewer, each participant's proficiency effect could also be incorporated as a fixed effect. Note also that it is also possible to directly model differences in tongue contours using Generalised Additive Mixed-effects Models (Al-Tamimi & Palo 2024; Strycharczuk, Lloyd & Scobbie 2024; Wieling 2018). While a previous study demonstrates that both GAMMs and PCA yield similar results for a dynamic analysis of tongue shape (Al-Tamimi & Palo 2024), it could be argued that the flexibility in the choice of subsequent statistical analysis can be an advantage of the PCA approach.

The current analysis, however, has been conducted for illustrative purposes only and thus further methodological considerations are necessary for a more formal analysis. For example, the decisions as to where the tongue shape is extracted need to be justified further, such as at the maximal constriction during consonant production (e.g., Léger, King & Ferragne 2023) or at the midpoint during a region of interest (e.g., Kirkham & Nance 2017). Also, the within-speaker normalisation using *z*-score is one of many available normalisation methods, such as range normalisation. Recent developments in the DLC implementation of the AAA software also allow tongue shape to be normalised across speakers based on the distance between the short tendon and the tongue surface (Strycharczuk, Lloyd & Scobbie 2024).

More importantly, while the current paper demonstrates the usefulness of the PCA analysis, it also illustrates some methodological challenges involved in data interpretation. PCA is purely a data-driven approach

with no *a priori* physiological foundations, meaning that different data sets may result in different PCs (Stone 2005). This can be clearly shown by comparing the PCs identified in the current paper and in Nagamine (2023), which were derived from similar data sets despite a few methodological differences (e.g., dynamic vs static analysis, the number of participants and prompts). Whereas Nagamine (2023) identifies tongue retraction (PC1) and tongue height (PC2), which are the fundamental dimensions in describing tongue shape (e.g., Johnson 2008), the current study does not offer such clear-cut decompositions of the tongue contours. Rather, PC2 in the current study may be similar to Turton's (2017) PC1 which corresponds to the clear-dark allophony of laterals in British English.

Note also that a reliable PCA analysis presupposes that the entire tongue surface is clearly visible in the image. In the case study, I decided not to perform hand correction on the estimated tongue contours, but I would emphasise that the decision to perform hand correction should be made carefully on a case-by-case basis.

The entirely data-driven nature of PCA underscores the importance of appropriate data collection, including tongue image quality, probe stabilisation and bite plane rotation as described in this paper. However, with all these considered appropriately, ultrasound analysis using the bite plane rotation, coupled with appropriate statistical methods such as *z*-score normalisation and PCA, offers a promising avenue for a reliable between-speaker comparison (Stone 2005). Ultrasound is one of the most accessible methods for articulatory research due to its low cost, non-invasiveness, portability and easier set-up procedure. This makes it particularly useful for use not just in a laboratory setting but also in language classrooms and fieldwork settings, offering insights into a wide range of research contexts by visualising the (once) invisible tongue movement.

Acknowledgements

This is part of the author's PhD research at Lancaster University, UK (funded by the Graduate Scholarship for Degree-Seeking Students by the Japan Student Services Organization (JASSO) and the 2022 Research Grant by the Murata Science Foundation). I am grateful to Prof. Claire Nance and Dr Sam Kirkham for their guidance and support throughout my PhD, to Prof. Noriko Nakanishi, Prof. Yuri Nishio and Dr Bronwen Evans for their help with data collection, and to Dr Alan Wrench for insightful discussions regarding bite plane normalization. Thanks also to Prof. Tatsuya Kitamura and Prof. Noriko Yamane for the invitation to contribute to the special issue; to anonymous reviewers for their constructive comments, and to all the participants for taking part in the study.

References

Al-Tamimi, J. and P. Palo (2024) "Retraction of the whole tongue induced by pharyngealisation in Levantine Arabic: A between-subject account using static and dynamic PCA and GAMMs." In I. Wilson, A. Mizoguchi, J. Perkins, J. Villegas and N. Yamane (eds.) *Ultrafest XI: Extended Abstracts*, 79–83. Zenodo. doi:10.5281/ZENODO.12578650

Alwan, A., S. Narayanan and K. Haker (1997) "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics." *The Journal of the Acoustical Society of America* 101(2), 1078–1089. doi:10.1121/1.417972

Antolík, T. K., C. Pillot-Loiseau and T. Kamiyama (2019) "The effectiveness of real-time ultrasound visual feedback on tongue movements in L2 pronunciation training: Japanese learners' progress on the French vowel contrast /y/-/u/." *Journal of Second Language Pronunciation* 5(1), 72–97. doi:10.1075/jslp.16022.ant

Articulate Instruments (2008) *Ultrasound Stabilisation Head-Set: Users Manual Revision 1.5*. Articulate Instruments.

Articulate Instruments (2022) Articulate Assistant Advanced version 220 [Computer software]. Articulate Instruments.

Baayen, R. H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511801686

Boersma, P. and D. Weenink (2022) Praat: Doing phonetics by computer (Version 6.2.09) [Computer software]. https://www.fon.hum.uva.nl/praat/ (accessed February 15, 2022)

Bryfonski, L. (2023) "Is seeing believing?: The role of ultrasound tongue imaging and oral corrective feedback in L2 pronunciation development." *Journal of Second Language Pronunciation* 9(1), 103–129. doi:10.1075/jslp.22051.bry

Davidson, L. (2006) "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance." *The Journal of the Acoustical Society of America* 120(1), 407–415. doi:10.1121/1.2205133

Derrick, D., C. Carignan, W. Chen, M. Shujau and C. T. Best (2018) "Three-dimensional printable ultrasound transducer stabilization system." *The Journal of the Acoustical Society of America* 144(5), EL392–EL398. doi:10.1121/1.5066350

Derrick, D. and B. Gick (2011) "Individual variation in English flaps and taps: A case of categorical phonetics." *Canadian Journal of Linguistics/Revue canadienne de linguistique* 56(3), 307–319. doi:10.1353/cjl.2011.0024

Gick, B., B. Bernhardt, P. Bacsfalvi and I. Wilson (2008) "Ultrasound imaging applications in second language acquisition." In J. G. Hansen Edwards and M. L. Zampini (eds.) *Studies in Bilingualism*, Vol. 36, 309–322. Amsterdam: John Benjamins Publishing Company. doi:10.1075/sibil.36.15gic

Hoole, P. and M. Pouplier (2017) "Öhman returns: New horizons in the collection and analysis of imaging data in speech production research." *Computer Speech & Language* 45, 253–277. doi:10.1016/j.csl.2017.03.002

Iskarous, K. and M. Pouplier (2022) "Advancements of phonetics in the 21st century: A critical appraisal of time and space in Articulatory Phonology." *Journal of Phonetics* 95, 101195. doi:10.1016/j.wocn.2022.101195

Iverson, P., P. K. Kuhl, R. Akahane-Yamada, E. Diesch, Y. Tohkura, A. Kettermann and C. Siebert (2003) "A perceptual interference account of acquisition difficulties for non-native phonemes." *Cognition* 87(1), B47–B57. doi:10.1016/S0010–0277(02)00198–1

Johnson, K. (2008) *Quantitative Methods in Linguistics*. Malden, MA: Wiley-Blackwell.

Johnson, K., P. Ladefoged and M. Lindau (1993) "Individual differences in vowel production." *The Journal of the Acoustical Society of America* 94(2), 701–714. doi:10.1121/1.406887

Kirkham, S. and C. Nance (2017) "An acoustic-articulatory study of bilingual vowel production: Advanced tongue root vowels in Twi and tense/lax vowels in Ghanaian English." *Journal of Phonetics* 62, 65–81. doi:10.1016/j.wocn.2017.03.004

Kirkham, S., P. Strycharczuk, E. Gorman, T. Nagamine and A. Wrench (2023) "Co-registration of simultaneous high-speed ultrasound and electromagnetic articulography for speech production research." In R. Skarnitzl and V. Jan (eds.) *Proceedings of the 20th International Congress of Phonetic Sciences*, 942–946. Guarant International.

Klein, H. B., T. McAllister Byun, L. Davidson and M. I. Grigos (2013) "A multidimensional investigation of children's /r/ productions: Perceptual, ultrasound, and acoustic measures." *American Journal of Speech-Language Pathology* 22(3), 540–553. doi:10.1044/1058-0360(2013/12-0137)

Kochetov, A. (2020) "Research methods in articulatory phonetics I: Introduction and studying oral gestures." *Language and Linguistics Compass* 14(4), e12368. doi:10.1111/lnc3.12368

Léger, A., H. King and E. Ferragne (2023) "Is rhoticity on the tip of your tongue? Tongue shapes for English /r/ in French learners with ultrasound." In R. Skarnitzl and J. Volín (eds.) *Proceedings of the 20th International Congress of Phonetic Sciences*, 2741–2745. Guarant International.

Li, M., C. Kambhamettu and M. Stone (2005) "Automatic contour tracking in ultrasound images." *Clinical Linguistics & Phonetics* 19(6–7), 545–554. doi:10.1080/02699200500113616

Maekawa, K. (2023) "Articulatory characteristics of the Japanese /r/: A real-time MRI study." In R. Skarnitzl and J. Volín (eds.) *Proceedings of the 20th International Congress of Phonetic Sciences*, 992–996. Guarant International.

Mathis, A., P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis and M. Bethge (2018) "DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning." *Nature Neuroscience* 21(9), 1281–1289. doi:10.1038/s41593-018-0209-y

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner and M. Sonderegger (2017) "Montreal Forced Aligner: Trainable text-speech alignment using kaldi." *Interspeech 2017*, 498–502. doi:10.21437/Interspeech.2017-1386

Mielke, J., A. Baker and D. Archangeli (2016) "Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/." *Language* 92(1), 101–140. doi:10.1353/lan.2016.0019

Mokhtari, P., T. Kitamura, H. Takemoto and K. Honda (2007) "Principal components of vocal-tract area functions and inversion of vowels by linear regression of cepstrum coefficients." *Journal of Phonetics* 35(1), 20–39. doi:10.1016/j.wocn.2006.01.001

Nagamine, T. (2023) "Dynamic tongue movements in L1 Japanese and L2 English liquids." In R. Skarnitzl and J. Volín (eds.) *Proceedings of the 20th International Congress of Phonetic Sciences*, 2442–2446. Guarant International.

Nance, C., M. Dewhurst, L. Fairclough, P. Forster, S. Kirkham, T. Nagamine, D. Turton and D. Wang (2023) "Acoustic and articulatory characteristics of rhoticity in the North-West of England." In R. Skarnitzl and J. Volín (eds.) *Proceedings of the 20th International Congress of Phonetic Sciences*, 3573–3577. Guarant International.

Nance, C. and S. Kirkham (2022) "Phonetic typology and articulatory constraints: The realisation of secondary articulations in Scottish Gaelic rhotics." *Language* 98(3), 419–460.

Narayanan, S. S., A. A. Alwan and K. Haker (1997) "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals." *The Journal of the Acoustical Society of America* 101(2), 1064–1077. doi:10.1121/1.418030

Preston, J. L., T. McAllister Byun, S. E. Boyce, S. Hamilton, M. Tiede, E. Phillips, A. Rivera-Campos and D. H. Whalen (2017) "Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors." *Journal of Visualized Experiments* 119, 55123. doi:10.3791/55123

Pucher, M., N. Klingler, J. Luttenberger and L. Spreafico (2020) "Accuracy, recording interference, and articulatory quality of headsets for ultrasound recordings." *Speech Communication* 123, 83–97. doi:10.1016/j.specom.2020.07.001

R Core Team (2023) R: A language and environment for statistical computing (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/ (accessed October 31, 2023)

Rebernik, T., J. Jacobi, R. Jonkers, A. Noiray and M. Wieling (2021) "A review of data collection practices using electromagnetic articulography." *Laboratory Phonology* 12(1), 6. doi:10.5334/labphon.237

Recasens, D. and C. Rodríguez (2016) "A study on coarticulatory resistance and aggressiveness for front lingual consonants and vowels using ultrasound." *Journal of Phonetics* 59, 58–75. doi:10.1016/j.wocn.2016.09.002

Riney, T. J., M. Takada and M. Ota (2000) "Segmentals and global foreign accent: The Japanese flap in EFL." *TESOL Quarterly* 34(4), 711–737. doi:10.2307/3587782

Saito, K. and K. van Poeteren (2018) "The perception–production link revisited: The case of Japanese learners' English /ɹ/ performance." *International Journal of Applied Linguistics* 28(1), 3–17. doi:10.1111/ijal.12175

Scobbie, J., E. Lawson, S. Cowen, J. Cleland and A. Wrench (2011) "A common co-ordinate system for mid-sagittal articulatory measurement." *QMU CASL Working Papers* 20, 1–4.

Scobbie, J., J. Stuart-Smith and E. Lawson (2012) "Back to front: A socially-stratified ultrasound tongue imaging study of Scottish English /u/." *Italian Journal of Linguistics/Rivista di Linguistica* 24(1), 103–148.

Slud, E., M. Stone, P. J. Smith and M. Goldstein Jr. (2002) "Principal components representation of the two-dimensional coronal tongue surface." *Phonetica* 59(2–3), 108–133. doi:10.1159/000066066

Spreafico, L., M. Pucher and A. Matosova (2018) "UltraFit: A speaker-friendly headset for ultrasound recordings in speech science." *Interspeech 2018*, 1517–1520. doi:10.21437/Interspeech.2018-995

Stevens, K. N. (2000) *Acoustic Phonetics*. Cambridge, MA: The MIT Press.

Stolar, S. and B. Gick (2013) "An index for quantifying tongue curvature." *Canadian Acoustics* 41(1), 11–15.

Stone, M. (2005) "A guide to analysing tongue motion from ultrasound images." *Clinical Linguistics & Phonetics* 19(6–7), 455–501. doi:10.1080/02699200500113558

Strycharczuk, P., S. Lloyd and J. M. Scobbie (2024) "Apparent time change in the articulation of onset rhotics in Southern British English." In R. Skarnitzl and J. Volín (eds.) *Proceedings of the 20th International Congress of Phonetic Sciences*, 3602–3606. Guarant International.

Strycharczuk, P. and J. M. Scobbie (2017) "Fronting of Southern British English high-back vowels in articulation and acoustics." *The Journal of the Acoustical Society of America* 142(1), 322–331. doi:10.1121/1.4991010

Tiede, M. (2021) GetContours version 3.5 [Computer software]. https://github.com/mktiede/GetContours (accessed March 1, 2025)

Turton, D. (2017) "Categorical or gradient? An ultrasound investigation of /l/-darkening and vocalization in varieties of English." *Laboratory Phonology* 8(1), 13. doi:10.5334/labphon.35

Westbury, J. R. (1994) "On coordinate systems and the representation of articulatory movements." *The Journal of the Acoustical Society of America* 95(4), 2271–2273. doi:10.1121/1.408638

Whalen, D. H., K. Iskarous, M. K. Tiede, D. J. Ostry, H. Lehnert-LeHouillier, E. Vatikiotis-Bateson and D. S. Hailey (2005) "The Haskins optically corrected ultrasound system (HOCUS)." *Journal of Speech, Language, and Hearing Research* 48(3), 543–553. doi:10.1044/1092-4388(2005/037)

Wieling, M. (2018) "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English." *Journal of Phonetics* 70, 86–116. doi:10.1016/j.wocn.2018.03.002

Wilson, I. (2014) "Using ultrasound for teaching and researching articulation." *Acoustical Science and Technology* 35(6), 285–289. doi:10.1250/ast.35.285

Wrench, A. and J. Balch-Tomes (2022) "Beyond the edge: markerless pose estimation of speech articulators from ultrasound and camera images using DeepLabCut." *Sensors* 22(3), 1133. doi:10.3390/s22031133

Yamane, N., P. Howson and W. Po-Chun (Grace) (2015) "An ultrasound examination of taps in Japanese." In The Scottish Consortium for ICPhS 2015 (ed.) *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5. The International Phonetic Association.

Zharkova, N. (2013) "Using ultrasound to quantify tongue shape and movement characteristics." *The Cleft Palate-Craniofacial Journal* 50(1), 76–81. doi:10.1597/11-196

Zhou, X., C. Y. Espy-Wilson, S. Boyce, M. Tiede, C. Holland and A. Choe (2008) "A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/." *The Journal of the Acoustical Society of America* 123(6), 4466–4481. doi:10.1121/1.2902168

（Received Apr. 17, 2024, Accepted Mar. 13, 2025,
e-Published Apr. 30, 2025）