



Assessing ultrasound probe stabilization for quantifying speech production contrasts using the Adjustable Laboratory Probe Holder for UltraSound (ALPHUS)

Wei-Rong Chen^{a,*}, Michael C. Stern^b, D.H. Whalen^{a,c}, Donald Derrick^d, Christopher Carignan^e, Catherine T. Best^f, Mark Tiede^g

^a Child Study Center, Yale University, New Haven 06511, CT, USA

^b Department of Linguistics, Yale University, New Haven 06511, CT, USA

^c Speech-Language-Hearing Sciences, CUNY Graduate Center, New York 10016, NY, USA

^d New Zealand Institute of Language, Brain, and Behaviour, University of Canterbury, Christchurch, New Zealand

^e Department of Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom

^f The MARCS Institute, Western Sydney University, Penrith, New South Wales, Australia

^g Department of Psychiatry, Yale University, New Haven 06511, CT, USA

ARTICLE INFO

Article history:

Received 11 October 2023

Received in revised form 20 May 2024

Accepted 21 May 2024

Available online 10 June 2024

Keywords:

Ultrasound imaging of the tongue

Probe stabilization

Head correction

ABSTRACT

Ultrasound imaging of the tongue is biased by the probe movements relative to the speaker's head. Two common remedies are restricting or algorithmically compensating for such movements, each with its own challenges. We describe these challenges in details and evaluate an open-source, adjustable probe stabilizer for ultrasound (ALPHUS), specifically designed to address these challenges by restricting uncorrectable probe movements while allowing for correctable ones (e.g., jaw opening) to facilitate naturalness. The stabilizer is highly modular and adaptable to different users (e.g., adults and children) and different research/clinical needs (e.g., imaging in both midsagittal and coronal orientations). The results of three experiments show that probe movement over uncorrectable degrees of freedom was negligible, while movement over correctable degrees of freedom that could be compensated through post-processing alignment was relatively large, indicating unconstrained articulation over parameters relevant for natural speech. Results also showed that probe movements as small as 5 mm or 2 degrees can neutralize phonemic contrasts in ultrasound tongue positions. This demonstrates that while stabilized but uncorrected ultrasound imaging can provide reliable tongue shape information (e.g., curvature or complexity), accurate tongue position (e.g., height or backness) with respect to vocal tract hard structure needs correction for probe displacement relative to the head.

© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Introduction

In recent years, the techniques for ultrasound imaging of the tongue have experienced rapid growth in speech-related research (e.g. Tabain and Beare, 2018; Recasens and Rodríguez, 2016; Tiede et al., 2019; Stone, 1990; Chiu et al., 2019), clinical applications (e.g., visual biofeedback as intervention for speech disorders) (e.g. Preston et al., 2019; Preston et al., 2017; Roxburgh et al., 2015), and L2 learning (e.g. Gick et al., 2008; Chang, 2023). The use of ultrasound to study speech production was pioneered by Maureen Stone

and colleagues (e.g., Stone et al., 1983; Shawker et al., 1985; Stone, 2005). Ultrasound is useful for studying speech because it can visualize large extents of the tongue, providing high spatial resolution relative to sparsely sampled flesh-point tracking systems like electromagnetic articulography (EMA), and high temporal resolution relative to real-time magnetic resonance imaging (rtMRI). It is also relatively inexpensive and easy to use compared to EMA and rtMRI methods. Ultrasound research on speech generally involves holding a transducer (probe) under the chin in order to acquire real-time images of the tongue surface during speaking.

Stone et al. (1983) first noted the main analytical challenge in ultrasound research on speech, which is that ultrasound

* Corresponding author.

acquires images of the tongue relative to the probe surface, rather than to vocal tract hard structures. For research primarily interested in tongue shape alone the raw images are sufficient for analysis. However, most studies of speech articulation relate tongue position to the hard palate and other fixed structures of the upper vocal tract. If the probe moves relative to the hard palate (e.g., when the jaw opens to produce a low vowel like /a/) then the shape of the vocal tract (i.e., the distance function between the tongue and the palate) cannot be inferred from the ultrasound image alone. Moreover, if the probe slides off the initial placement, the desired target plane is lost and the acquired images are not comparable across sequences, thus resulting in inconspicuous errors in both tongue position and tongue shape. There have been many parallel threads of methodological development attempting to mitigate this issue, which can be grouped into two categories: (1) restriction and (2) correction methods.

1.1. Ultrasound imaging of the tongue: the restriction method

The restriction method aims to keep the relative position of the head and probe constant (e.g. [Gick et al., 2005](#); [Stone and Davis, 1995](#)). The head and transducer support system (HATS) ([Stone and Davis, 1995](#)) is perhaps the first successful ultrasound stabilization system based on this method. It completely immobilizes both the head and the probe using a large metal structure built into the recording room. It achieved sub-millimeter accuracy, but it is challenging to replicate the system, and the fact that the head is strictly immobilized limits its applications (e.g., participant discomfort during recording, limitations on co-recording with other devices, etc.).

More recently, many headset-style head-probe stabilization systems have been developed, with the aim to allow for head movement while keeping the probe to head position constant. For example, Articulate Instruments Ltd has marketed two stabilization headsets, the Ultrasound Stabilization Headset system (USH: [Scobbie et al., 2008](#)) and the UltraFit system ([Spreafico et al., 2018](#); [Matosova, 2016](#)). The metallic (aluminum) USH was developed earlier and was effective in providing rigid stabilization. It was then replaced by UltraFit, a more flexible and lightweight headset made of nylon.

The UltraFit provides less rigid probe-to-head fixation than USH due to its material but is more comfortable in prolonged usage. Those headsets are adjustable to fit a wide range of speakers, even with different head sizes, but the error (i.e., relative probe movement during usage) is mostly affected by the rigidity of the headset and the tightness of the headset fit to the head. [Pucher et al. \(2020\)](#) reported the error magnitudes (i.e., relative probe movements after head-movements were removed) for both UltraFit and USH, estimated by 95% quantile interval (97.5th percentile minus 2.5th percentile). When using the metallic USH, the largest magnitudes of error (excluding the highest and lowest 2.5% of data) among the three participants were measured at 12.9, 9.1, and 7.2 mm in posterior-anterior, inferior-superior, and left-right axes, respectively ([Pucher et al., 2020, p. 89](#)); unfortunately, the measurements for UltraFit were likely compromised by errors in head movement removal, as the authors explained, leaving only an implication that the largest error for UltraFit lies in the anterior-posterior direction (p. 90).

All restriction methods that immobilize the probe relative to the head also greatly restrict movement of the jaw, reducing naturalness during speaking. This results in a trade-off between naturalness of speaking and accuracy of tongue measurement; the tighter (more rigid) the head-probe stabilization, the more accurate the tongue image but less comfortable and less natural the speech, and vice versa. Further, small jaw movements can still be made in a very rigid stabilization system by depressing the soft tissue on the chin against the probe, but this depression can cause artificial deformations of tongue surface shape (e.g., [Stone and Davis, 1995](#); [Stone, 2005](#)).

1.2. Ultrasound imaging of the tongue: the correction method

The correction method compensates for head movement in data post-processing, rather than restricting it during data collection. The most comprehensive head movement correction method is the Haskins Optically Corrected Ultrasound System (HOCUS: [Whalen et al., 2005](#)). This method uses spatial tracking of points (e.g. using either an optical motion capture system or EMA) to align the rigid bodies of the probe and the head to re-orient tongue surface contours relative to the palate. Video-based correction algorithms have also been presented, such as Palatron ([Mielke et al., 2005](#)) and SOLLAR ([Noiray et al., 2020](#)), as less-accurate but cost-effective alternatives that only use a video camera to track fixed points on the rigid bodies of the head and the probe. Using the correction method, speakers can move their jaw relatively naturally while speaking, and accurate vocal tract areas can still be inferred by the researcher. Theoretically, if the tongue surface is acquired as a three-dimensional volume (i.e., 3D/4D ultrasound), motions in any direction can be compensated for. Currently, however, the most commonly used ultrasound systems acquire two-dimensional planar images only. For such 2D systems, the correction method has an assumption that the plane in which images are acquired is consistent within the speaker and, therefore, only the relative probe movements on the assumed acquiring plane can be corrected; those off the plane cannot. The same principle also applies to bi-planar (simultaneous mid-sagittal and coronal) recording in 3D/4D ultrasound.

For example, given that both the probe and the head have six degrees of freedom (see [Fig. 1](#)), if the probe is oriented to acquire midsagittal tongue shape (the most common usage), probe motions along the superior-inferior (z) and anterior-posterior (x) axes, as well as rotation around the left-right axis (pitch) are correctable because the relevant data are part of the targeted imaging plane. However, probe movement along the left-right (y) axis, or rotation around the x-axis (roll) or z-axis (yaw) is not correctable because the imaged data are no longer aligned with the target plane. Samples in which movement over the uncorrectable degrees of freedom (DOF) exceeds tolerable limits must be discarded.

1.3. The current study

To bridge the gap between the restriction and correction methods, this study presents and assesses an optimal approach to probe stabilization. This approach utilizes a 3D-printable probe stabilizer, named ALPHUS: Adjustable Labora-

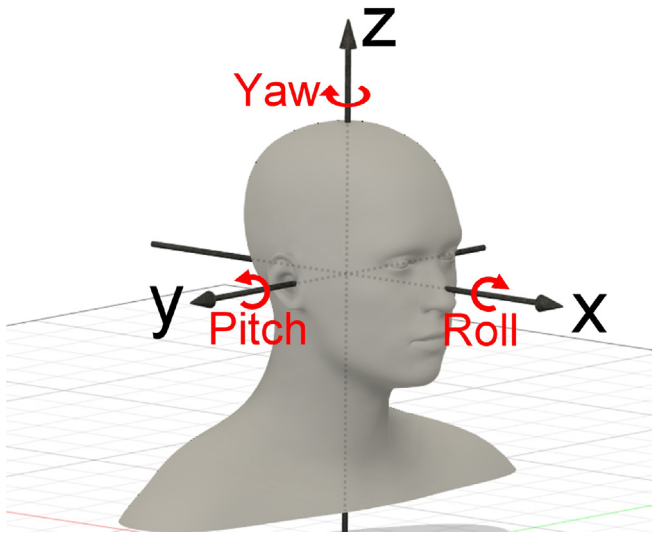


Fig. 1. Illustration of the degrees of freedom with respect to the speaker. x: posterior-anterior. y: left-right. z: inferior-superior.

tory Probe Holder for UltraSound. In contrast to conventional stabilization systems that sacrifice natural jaw opening in exchange for accuracy, we deliberately designed the stabilizer to be flexible in the pitch and z DOF (i.e., jaw opening) to allow for naturalistic speech production. By allowing speakers to open their jaw and then correcting for this movement in data post-processing, the current design enhances both the naturalness of speech as well as the recoverability of vocal tract shapes relevant for linguistic contrasts. In the upcoming sections of this paper, we will describe the stabilization approach in details, discuss probe movement in correctable and uncorrectable DOF and report accuracy measurements of the stabilization system. We will also present examples of English and Mandarin phonemic contrasts in ultrasound tongue contours to demonstrate the effects of stabilization and correction for probe-to-head movements.

2. Method

2.1. The design

The design of ALPHUS comprises two parts: the upper part, which is a cap placed roughly at the crown of the head, and the lower part that holds and stabilizes the probe under the chin in a perpendicular orientation to the y axis with respect to the speaker. Fig. 2 demonstrates the headset's application in sagittal tongue imaging with both closed (a) and open jaw (b). Following that, the probe was rotated 90 degrees for imaging the tongue surface in the coronal plane, as shown in Fig. 2c. The rightmost sub-figure (Fig. 2d) presents how the headset is configured to fit a five-year-old boy. Similar to the previous design (Derrick et al., 2018), these two parts are interconnected by elastic bands and buckles, allowing for the whole lower part to move downward while restricting any movements away from the midsagittal plane. While the upper part is a single object (Fig. 6), the lower part is more configurable, consisting of 17 sub-parts as illustrated in Fig. 3. Table 1 lists the

names of these parts (right column), required quantities (middle column), and item numbers (left column) referring to those in Fig. 3 and 5. An animation of assembling and disassembling the lower part of the headset can be found in the supplementary material (SUPP01) as well as on the Github repository.

2.2. Configuration and usage

The ultrasound probe is initially secured within a pair of clips, which are then fastened to the clip holder. Subsequently, for sagittal imaging, the probe is affixed to the leg using a hex rod (length = 33 mm), as depicted in Fig. 4a. Alternatively, for coronal imaging, it is attached with an L-shaped coronal plane adapter, as shown in Fig. 4b. The probe clips were individually designed for each specific probe, based on accurate probe dimensions acquired by 3D-scanning. The “leg” connects the clip holder with the main frame (i.e., the base and the arms) of the headset in seven possible different angles, from -15 to 15 degrees with an increment of five degrees (each angle is realized in a separate STL file). For sagittal imaging, using more negative angles results in imaging in more posterior regions in the vocal tract. The four small holes in the front of the clip holder are the reference points for the external point-tracking systems (the four points labelled as “Refs” in Fig. 4b). Object tracking markers (e.g., infrared emitters in optical tracking or coil sensors in EMA tracking) are to be placed in three of these holes for tracking the movements of the rigid body formed by the probe, clips and clip holder. At least three more tracking markers also need to be placed on the hard structure of the speaker's skull (e.g., mastoid processes, nasion, upper incisors, etc.) to simultaneously track head movements.

The speaker will then be instructed to sit straight while the experimenter adjusts the headset. The roll angle of the base (#4) of the headset is adjusted until the bubbles in the bubble levels (#17 in Fig. 3) are in the middle; then the speaker can relax and freely move the head and open the jaw during the experiment.

When the headset is to be used for young children, the width of the main frame can be narrowed by moving the two arms (#5 in Fig. 3) to the inner bracket, and the height and front-back location of the cheek adjusters (#10) can be adjusted according to the speaker's head size, as exemplified in Fig. 2d.

2.3. 3D-scanning of probe

The current system relies on 3D-scanning of the ultrasound probe to produce probe-specific clips. We no longer provide general purpose clips (for probes in unknown dimensions with the aid of modeling clay) as in the previous design (Derrick et al., 2018) because we found that the calibration process for head/probe-correction is almost impossible if the dimensions of the probe cannot be accurately acquired. As of the time of writing, we have supplied clips for 10 different ultrasound probes on the Github repository, covering the most commonly used ultrasound systems for speech research, and we will release additional clips to supplement this collection as needed.

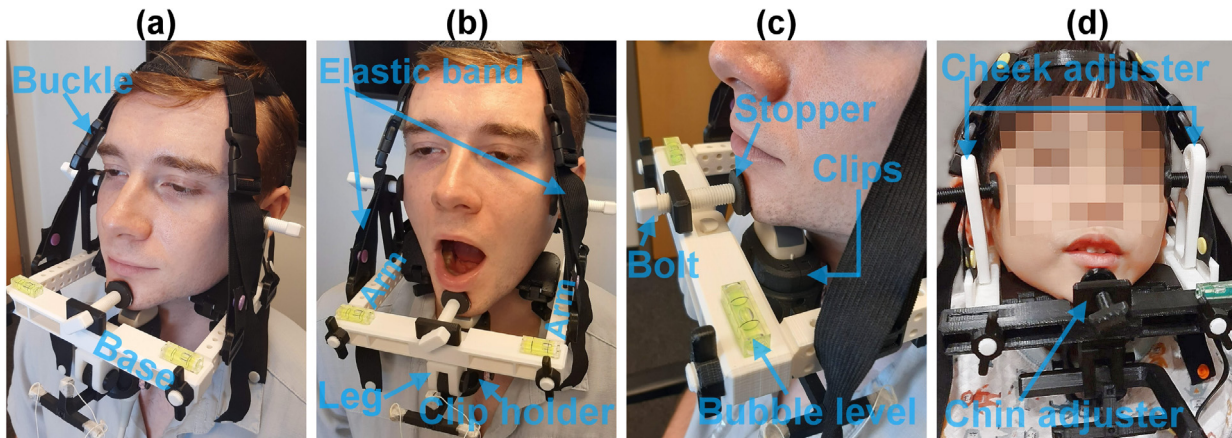


Fig. 2. Photo of the headset in place. (a) Configured for measuring in midsagittal plane. (b) Same as (a) with the jaw opened wide. (c) Configured for measuring in coronal plane. (d) Configured and mounted on a five-year-old boy.

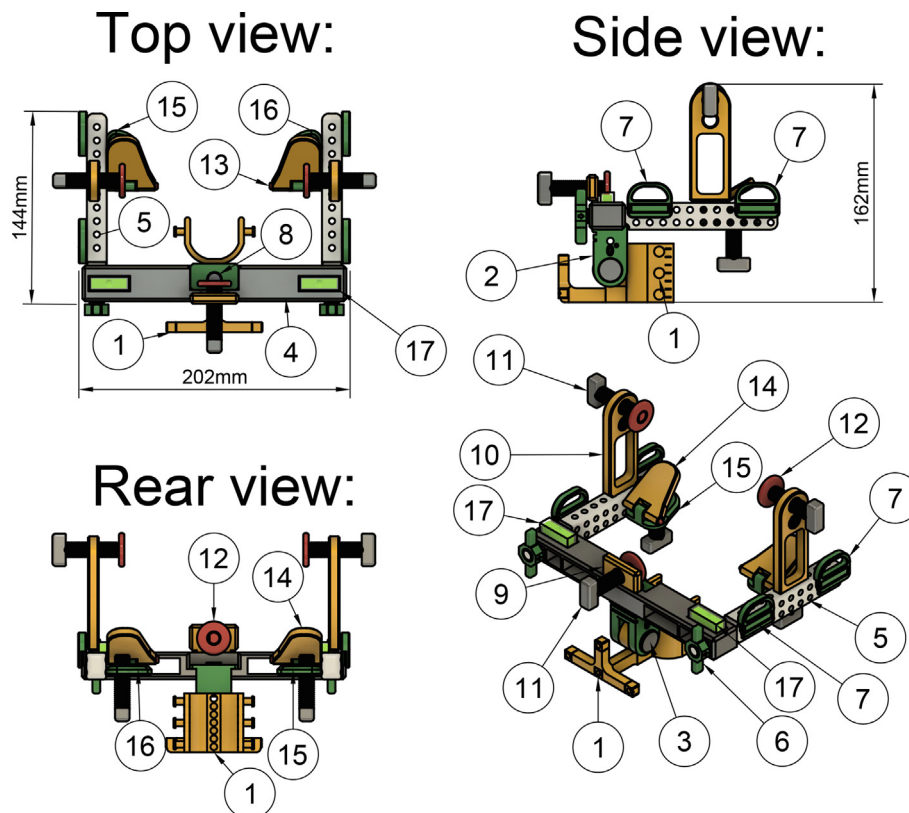


Fig. 3. The bottom part of the probe stabilizer in four view angles. The numbers label each unique part as listed in Table 1.

2.4. Fabrication

The fabrication of the entire headset can be mostly accomplished by 3D-printing the STL files obtained from the Github website. There are a few non-3D-printable components, including five pairs of knitted elastic straps with side-release buckles, two bubble levels (10x10x29 mm) and a few rubber bands (or knit elastic hair bands). These non-3D-printable parts can be procured from local or online fabric and crafts stores. To encourage broader usage, the current design has been optimized to allow the parts to be 3D-printed with

biodegradable Polylactic Acid (PLA) using the most basic type of 3D printer, the single extruder Fused Deposition Modeling (FDM) 3D printer. The whole headset weighs around 320 grams (excluding the probe).

3. Evaluation

We analyzed data from three experiments, involving co-collections of ultrasound and motion tracking with EMA for sagittal and coronal tongue imaging. These experiments were

Table 1

List of parts in the assembly. The second column (Qty) lists the quantities to be printed for each part. Asterisk (*) indicates the part (the bubble level) that is not 3D-printable and must be acquired from the marketplace.

Item	Qty	Part Name
1	1	Clip_holder
2	1	Leg
3	1	Hex_rod
4	1	Base
5	2	Arm
6	2	Nut
7	4	Ear
8	1	Leg_pin
9	1	Chin_adjuster
10	2	Cheek_adjuster
11	5	Bolt
12	3	Stopper
13	2	Jaw_flap_rod
14	2	Jaw_flap
15	1	Jaw_flap_base_R
16	1	Jaw_flap_base_L
17	2	Bubble level *

originally designed and carried out in a larger project for speech-related investigations. The three experiments were selected because they all have a wide coverage of speech sound inventories and were recorded in long sessions (i.e., 30–60 + minutes), providing a solid foundation for evaluating the stabilization system. In every experiment, speakers were recorded while reading aloud stimuli presented on a computer screen positioned approximately 1 m in front of the speaker's eyes. Ultrasound images were collected using a Telemed MicrUs EXT-1H system with a synchronization port and a micro convex probe (MC4-2R20S-3) at 100 frames per second (FPS). Kinematic data were simultaneously collected using a Carstens AG501 EMA system, sampled at 250 Hz. EMA sensors were attached to the three reference points on the probe clip holder for tracking probe movements ("Refs" in Fig. 4). Audio samples were recorded using a Sennheiser MKE-600 shotgun microphone with pre-amplified output split between Carstens recording (for synchronization with EMA) and a TASCAM US-600 USB audio interface with the sampling rate set at 44.1 kHz (for synchronization with ultrasound). The TASCAM stream recorded both audio and the Telemed frame synchronization signal as stereo audio files. During post-processing cross-correlation of the EMA and ultrasound-aligned audio was used to produce common alignment between audio, EMA, and ultrasound.

Head motion was tracked by EMA sensors placed at the left and right mastoid processes and the midline of the upper incisors. Sensors for tracking active articulators were attached to the tongue tip (TT; approximately 0.5 cm behind the apex of the tongue), tongue body (TB; the bite mark of the upper incisor on the tongue when the tongue is maximally protruded out of the mouth), jaw (midline of the lower incisors), and the vermillion borders of the upper (UL) and lower lips (LL). The head stabilizer headset was mounted on the speaker after all EMA sensors were attached.

The speaker's occlusal plane was established using a bite-plate, with three sensors for tracking, clasped between the teeth. All EMA data were subsequently aligned with this plane with an origin at the speaker's upper incisors (i.e., corrected for head motion). We also obtained a reference "clench" trial during which the speaker maintained a closed jaw, clenched teeth posture to serve as the baseline.

Ultrasound tongue contours were automatically traced and tracked by using the DeepEdge AI Tool for Ultrasound (Chen et al., 2020) (available at <https://github.com/WeirongChen/DeepEdge>). At each time point associated with an ultrasound video frame, the rotation matrix from the baseline probe position (at clench) to the current probe position was calculated, and then the inverse of that transformation was applied to the tongue contour to correct for the probe motions relative to the head, as described in Whalen et al. (2005).

EMA and ultrasound have two different coordinate systems: EMA uses a 3D mm coordinate system relative to the device (until remapped to the speaker), and tongue contours measured from ultrasound are in 2D pixel units. In order to co-register the two different coordinate systems a calibration procedure must be performed (at least once) for each probe used by the ultrasound system. The objective of the calibration procedure is to identify at least three positions of a sensor co-recorded in both EMA and ultrasound modalities, which supports rigid body transformation between the coordinate systems. We used the "water" method similar to the one described in Chen et al. (2017) to calibrate the coordinate system for the ultrasound probe. The method involves placing the ultrasound probe beneath a thin plastic container filled with water, followed by submerging an EMA sensor into the water to tracing the probe's mid-sagittal line. The EMA sensor's underwater location is visible in the ultrasound image, allowing

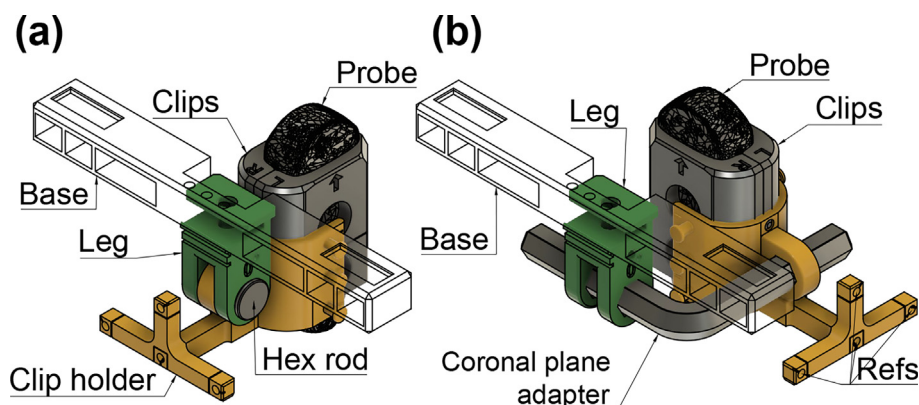


Fig. 4. Configurations of the orientations of probe and clip holder. (a) Oriented to midsagittal plane. (b) Oriented to coronal plane.

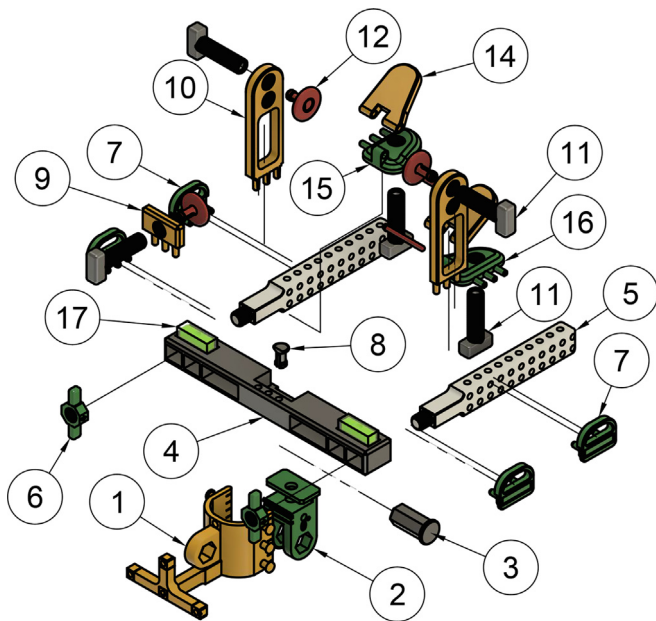


Fig. 5. Disassembled bottom part of the probe stabilizer.

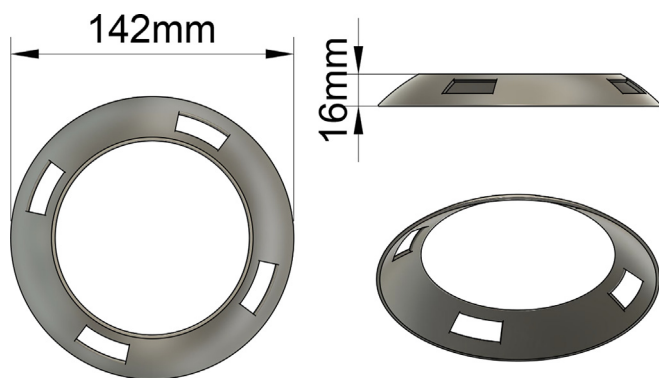


Fig. 6. The upper part of the probe stabilizer.

the EMA sensor's locations in both the ultrasound image and the EMA coordinate system to be co-registered.

To evaluate the accuracy of the stabilization system, we calculated the relative probe movement, defined as the displacement of the probe from the baseline (position during the "clench" trial) after head motion was removed. For each experiment, the distribution of these relative movements throughout the entire experiment indicates the stabilization errors.

To validate the process of head movement removal (i.e., head-correction) and to ensure that such a process does not introduce extra errors in the measurement of relative probe movement, we measured the distance between a reference sensor on the head (left mastoid) and another sensor on the probe (i.e., probe-to-head distance). Then, for each sample, we compared this measurement taken before and after head movement removal. The maximum values of the absolute differences between the probe-to-head distance measured before and after head-correction were 0.5, 0.75 and 0.26 mm for our three experiments, respectively, indicating negligible influence of the head-correction process on the accuracy of the measurement of relative probe movement.

3.1. Experiment I - Midsagittal plane

In the first experiment (Expt. 1), a male native speaker of Taiwan Mandarin (age 46) produced a list of onset-rime sequences in a carrier phrase [tʂə4 _ pa5] ("This is X." The numbers indicate lexical tones. 4: falling tone; 5: neutral tone). The target syllable (i.e., the "X" in the carrier phrase) had a consonant-vowel (CV) or consonant-vowel-nasal (CVN) structure, where the onset consonant was one of /p, m, f, t, n, l, k, x~h, tʂ, ʃ, z, ts, s/, the vowel was one of /i, a, u, ei, ie, ou, uo/, and the nasal coda /n/ or/ŋ/. This experiment was intended to record the (nearly) full phonetic inventories of Mandarin monosyllables, covering most of CV and CVN combinations in Mandarin. Tones were ignored, despite some reported differences in tongue height across Mandarin tones (e.g., Shaw et al., 2016). The list of stimuli consisted of 135 target syllables and the speaker repeated the list three times. Additionally, readings of the passage "The North Wind and the Sun" in both Mandarin and English versions were also recorded. The whole recording took one hour and six minutes, resulting in 456,816 total samples (acquired by EMA).

3.2. Experiment II - Midsagittal plane

In the second experiment (Expt. 2), a male native speaker of American English (age 28) was recorded. The experiment consisted of four blocks. The first block consisted of 32 sentences presented twice in random order. The sentences were designed to elicit a variety of vowel transitions (e.g., /a/ to /i/ transition in "Please a cop, Peter.") varying in the position of stress, as well as the number and syllabic identity of intervening consonants. Altogether, the sentences elicit good coverage of the range of vocal tract shapes produced during American English speech. During each presentation, the speaker produced the sentence five times at a normal speaking rate. The second block consists a subset of the sentences in the first block. The speaker was instructed to produce the sentence five times at a fast (twice normal) speaking rate in this block. In the third block, the speaker produced 26 isolated monosyllabic English words three times in random order. This set of words covers the vowel inventory of American English. Finally, in the fourth block, the speaker produced the "Comma Gets a Cure" passage (Honoring et al., 2000) four times at a normal speaking rate. This passage was designed to elicit good coverage of the English phoneme inventory across a variety of contexts and is commonly used for the examination of English pronunciation. The recording of Expt. 2 took 48 min and the total number of EMA samples analyzed for the current evaluation was 398688.

3.3. Experiment III - Coronal plane

The third experiment (Expt. 3) was designed to investigate nasal coda coarticulation and tongue grooving of sibilants in Taiwan Mandarin. The same speaker as in Expt. 1 was recorded and the same procedures of EMA and ultrasound co-collection applied. The experiment was started with a block for recording nasal coda coarticulation in the midsagittal plane, then the probe clip holder was rotated 90 degrees using the coronal plane adapter (Fig. 4b) for recording sibilant tongue

grooving in the coronal plane. For the purpose of the current paper, we only report the results of the latter block. The sibilant stimuli in Expt. 3 were CV combinations where the onset consonant was one of /tʃ, c, tʃ, tʃʰ, ʃ, z, ts, tsʰ, s/, and the vowel was one of /a, i, u, y, ɤ, ɤ, ye, ow, ej, aj, z, z/, excluding accidental gaps in Mandarin. The whole experiment took an hour and the sibilant block took around 30 min, resulting in 54144 samples.

4. Error measurements

4.1. Relative probe movements as error measurements

As mentioned above, if ultrasound tongue imaging is acquired with the restriction method (without correction), any probe movement introduces errors in measured tongue surface position relative to the head. Such error can be measured by comparing the probe position of the current frame to a baseline position (e.g., jaw tightly closed or teeth clenched). For example, assuming the probe position in a clench trial (after head movement is removed) is P_0 (as in Fig. 7) and the probe position of the current frame is P_1 , the direct error e_1 is calculated as $e_1 = |\vec{P}_1 - \vec{P}_0|$, and can also be decomposed into six degrees of freedom, namely position along x (posterior-anterior), y (left–right), and z (inferior-superior), and rotation around x (rx , roll), y (ry , pitch) and z (rz , yaw). This error constitutes the most direct measure of the accuracy of the probe stabilizer (e.g., Derrick et al., 2018; Pucher et al., 2020). The six DOF for the relative probe movements are defined with respect to the speaker, as illustrated in Fig. 1. With head/probe-correction, we are more concerned with errors in uncorrectable DOF, which are roll, yaw and y in midsagittal imaging and pitch, yaw and x in coronal imaging, while allowing for errors in the rest DOF.

4.2. Range of probe-to-head distance

An alternative measurement has been reported, calculating the consistency or changes in the distance between a fixed point on the head and another on the probe (e.g., the range or distribution of Euclidean distance between nose and probe in Pucher et al. (2020, Fig. 5 & 6) and Spreafico et al. (2018, Fig. 5)). This measurement is based on an assumption that the distance between the probe and the head is held constant if there are no relative probe movements, and any such movements will result in changes in this distance. However, this assumption is met only in some restricted conditions but not otherwise. Let the fixed point on the head be R_0 , and the distance from R_0 to that on the probe d_i , where $i = 0 \dots n$ ($i = 0$ when the probe is at the baseline position, P_0) for the i th frame (as denoted in Fig. 7). A change of d_i can be conceptually expressed as the $\Delta d = |d_i - d_0|$. As shown in Fig. 7a, when the probe moves along the circumference of the circle where the origin is R_0 and the radius d_0 (dash-dotted circle in Fig. 7a), this measurement (Δd) will be a constant zero, while the true error (e_1) can be large and increases as the angle θ between d_1 and d_0 is increasing. Δd only approaches e_1 when θ is close to zero (as in Fig. 7b). Given the most common orientation of the probe and head marker placement in ultrasound tongue imaging, this alternative measurement mostly captures

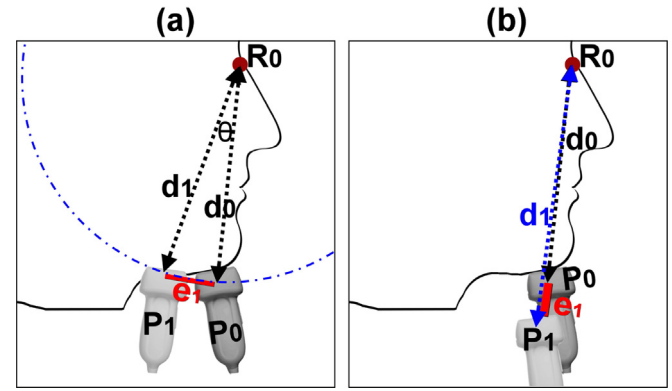


Fig. 7. Concepts of relative probe movements and error measurements. (a) Probe slides along a circle defined by the origin R_0 and radius d_0 . (b) Probe moves downward as the jaw opens. R_0 : A fixed point on the hard structure of the head. P_0 : baseline position of the probe (when the teeth are clenched). P_1 : current position of the probe. e_1 : absolute displacement from P_0 to P_1 (i.e., relative probe movement). d_0 : distance between R_0 and P_0 . d_1 : distance between R_0 and P_1 . θ : angle between d_1 and d_0 .

the error projected onto the inferior-superior axis, but underestimates the error on the posterior-anterior and left–right axes.

4.3. Robustness measurement

While measures of mean and standard deviation (or median and IQR) error are useful indications of overall accuracy, it can also be helpful to know the frequency and range of extreme errors when evaluating a system. For this purpose we additionally provide a measure computed as the mean of the Worst $k\%$ of observed measurement errors, where k is 5% ($MW5\%$) or 1% ($MW1\%$).

5. Results

5.1. Midsagittal plane

Fig. 8 summarizes the results of the evaluations for the first two experiments. The detailed values are listed in Table 2. The measurements marked with an asterisk (e.g., rx , rz , y) indicate those degrees of freedom (DOF) in which movement is uncorrectable in 2D ultrasound images; the objective is thus to minimize the movements over these DOF but allow for movement in the remaining (correctable) DOF. In both Expt. 1 and Expt. 2, the magnitudes of relative probe movement were negligible in the three uncorrectable DOF: 95% of the samples along these DOF were smaller than 2 mm in the y axis and less than 2° in roll and yaw. Even in the worst 1% of cases ($MW1\%$), movements were only slightly greater than 2 mm in y and still less than 2° in roll and yaw.

On the other hand, a large extent of movement was observed in the correctable pitch and z DOF. The movements in another correctable DOF, x , were slightly larger than those in the uncorrectable y DOF but still minimal (3–4 mm) in the two experiments. This is because the chin adjuster (#9) and the interconnected bolt and stopper (e.g., the piece that runs against the chin in Fig. 2) prevents the probe from posterior-anterior sliding. The movements along correctable DOF can then be applied to the obtained tongue contours to register them consistently with vocal tract hard structure. Overall, the stabilizer succeeded in limiting movement in the uncorrectable

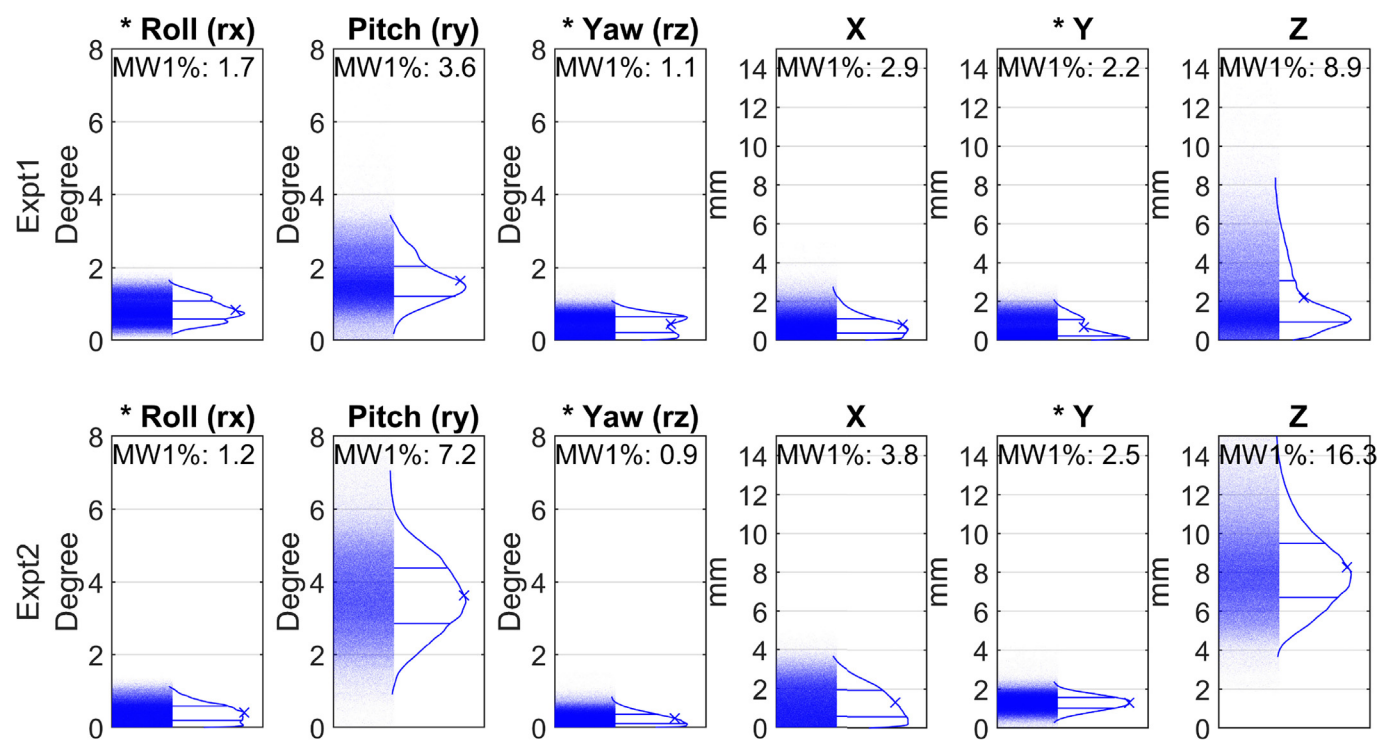


Fig. 8. Relative probe movements in the midsagittal plane. Columns labelled with an asterisk indicate uncorrectable degrees of freedom. In each panel, the curved line indicates the probability distribution of the samples, the two horizontal lines the third (upper) and the first (lower) quartile, and the cross on the curved line the mean. Actual samples were plotted with randomized x-values to the left of the curved line, creating shaded bands that visualize the density of the data.

Table 2
Relative probe movements in the mid-sagittal plane for the first two experiments EX1 and EX2. The unit for roll, pitch and yaw is degree (angle) and that for x, y and z is mm. SD: standard deviation of the samples; Max: absolute maximum value; 95%, 99% and 99.9%: .95, .99 and .999 quantile values, respectively; MW5% and MW1%: the mean of the worst (highest) 5% and 1% of samples, respectively. Asterisks indicate uncorrectable degrees of freedom.

Expt	Measure	Mean	SD	Max	95%	99%	99.9%	MW5%	MW1%
Expt1	*Roll (rx)	0.8	0.3	2.1	1.4	1.6	1.8	1.5	1.7
	Pitch (ry)	1.6	0.7	7.3	2.8	3.3	4.1	3.2	3.6
	*Yaw (rz)	0.4	0.3	1.6	0.9	1	1.3	1	1.1
	X	0.8	0.6	5	1.9	2.5	3.3	2.3	2.9
	*Y	0.7	0.5	2.9	1.6	2	2.4	1.8	2.2
	Z	2.2	1.8	14.2	5.8	7.6	10.8	7	8.9
Expt2	*Roll (rx)	0.4	0.3	1.4	0.9	1.1	1.2	1	1.2
	Pitch (ry)	3.6	1.1	9.1	5.4	6.6	8	6.1	7.2
	*Yaw (rz)	0.3	0.2	1.5	0.6	0.8	1	0.7	0.9
	X	1.3	0.9	5	2.9	3.5	4.1	3.3	3.8
	*Y	1.3	0.4	4.2	1.9	2.3	2.9	2.2	2.5
	Z	8.3	2.2	20.7	12.4	14.9	18.2	13.9	16.3

DOF (roll, yaw, and y), while allowing natural movements in the correctable DOF (pitch, x, and z) in order to encourage natural-istic speaking.

Furthermore, an inquiry arises regarding whether the probe movement remains consistent across utterances or may be reduced on shorter timescales, as raised by [Lulich and Pearson \(2019\)](#). To address this, we re-plotted [Fig. 8](#) by utterance, resulting in [Fig. 9](#). Each data point (blue circle) in [Fig. 9](#) indicates the mean of the worst (largest) 1% (MW1%) of samples in each utterance, and the indications of distributions (blue curve, horizontal lines and blue cross) for the six DOF were drawn from the distributions of utterances. In these two experiments, each utterance was a short sentence or phrase, with median duration of 14.5 and 10.5 for Expt. 1 and Expt. 2, respectively. Only the utterances within +/- 10% of the median duration were included in [Fig. 9](#), resulting in 109 and 102 utterances for

Expt. 1 and Expt. 2, respectively. The results showed that there were wide ranges of dispersion across utterances around 1 cm in the z DOF and 4° in pitch. The results also showed that the scale of the probe movements does not appear to be smaller on shorter timescales, as the means of MW1% across utterances in [Fig. 9](#) are comparable to the values of MW1% across the entire experiment in [Fig. 8](#).

5.2. Coronal plane

[Fig. 10](#) and [Table 3](#) display the results of the third experiment, where the probe was configured to acquire tongue images in the coronal plane. The results are quantitatively similar to those in Expt. 1 (with the same speaker of the same language). However, note that the set of uncorrectable DOF is different in this experiment relative to the midsagittal

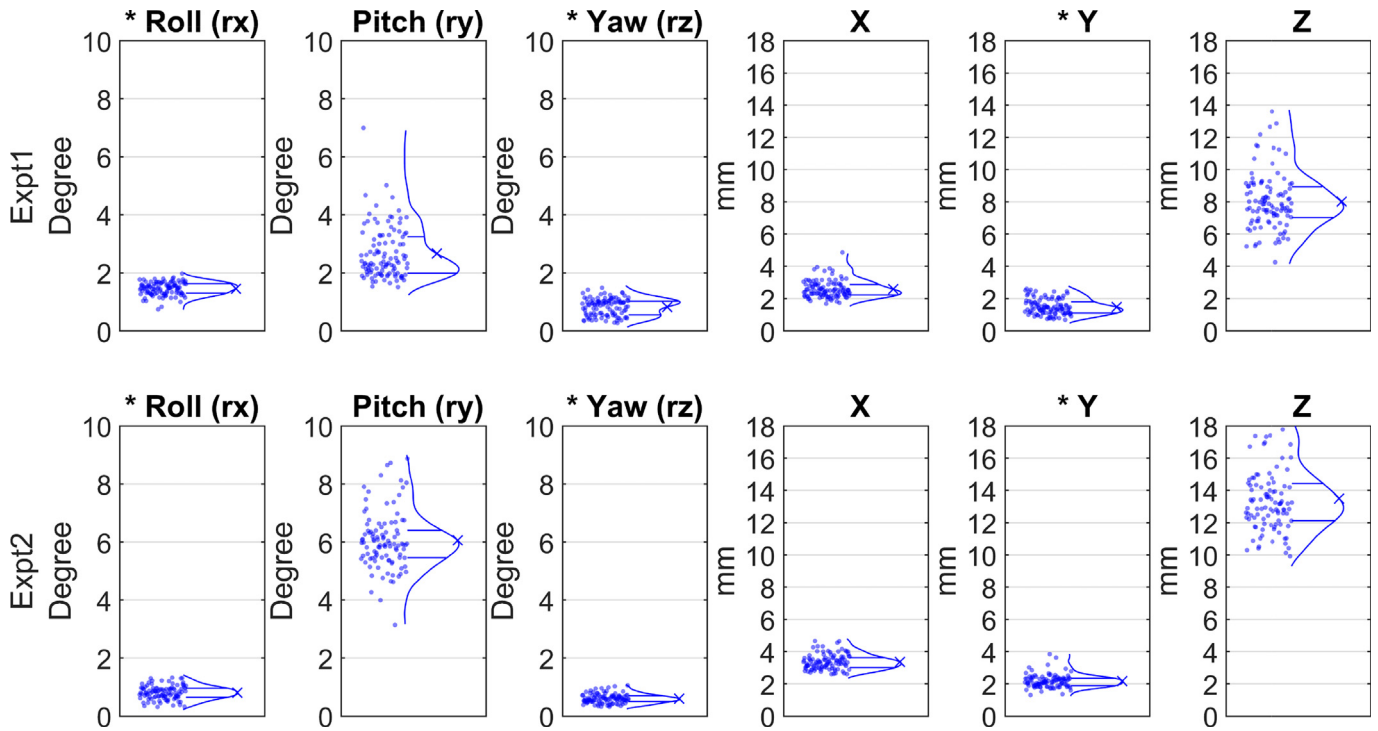


Fig. 9. Relative probe movements in the midsagittal plane; same as Fig. 8 but plotted by utterance. Each data point represents the mean of the largest 1% of relative probe movements in each utterance. The medians of utterance duration were 14.5 and 10.5 s for Expt. 1 and Expt. 2, respectively. The number of utterances was 109 for Expt. 1 and 10.2 for Expt. 2.

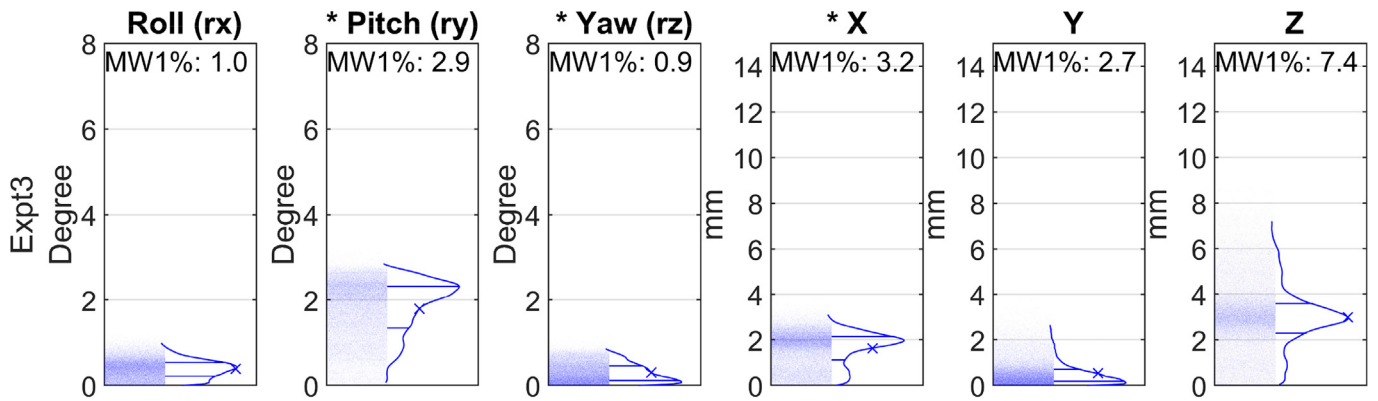


Fig. 10. Relative probe movements in the coronal plane. Denotations same as the previous figure.

Table 3

Relative probe movements in the coronal plane for the third experiment. Units and column denotations are the same as in Table 2.

Expt	Measure	Mean	SD	Max	95%	99%	99.9%	MW5%	MW1%
Expt3	Roll (rx)	0.4	0.2	1.2	0.8	0.9	1.1	0.9	1
	*Pitch (ry)	1.8	0.7	3.2	2.6	2.8	3.1	2.7	2.9
	*Yaw (rz)	0.3	0.2	1.2	0.7	0.8	1	0.8	0.9
	*X	1.6	0.7	3.9	2.6	3	3.4	2.8	3.2
	Y	0.5	0.5	3.8	1.7	2.4	3.2	2.1	2.7
	Z	3	1.4	8.8	5.7	6.9	7.8	6.3	7.4

experiments Expt. 1 and Expt. 2. Recall that the uncorrectable DOF are those which fall outside of the intended imaging plane and the correctable DOF the rest. These DOF differ depending on the orientation of the probe. When the probe is oriented for coronal imaging, the uncorrectable DOF are ry , rz , and x . Also recall that the system is designed to be unrestricted in the z (vertical) and ry (pitch) DOF in order to support naturalistic

jaw movement during speech. While movements in the z axis remain correctable for coronal plane recordings, movements over the x and ry DOF are no longer correctable in coronal plane recordings. In this experiment, results indicate that errors were negligible in rz (yaw) ($< 1^\circ$) and relatively small in the x and ry DOF: 95% of samples were less than 2.6 $^\circ$ in ry and 2.6 mm in x . The mean of the worst 1% of samples (MW1%)

was less than 3° in *ry* (pitch) and around 3 mm in *x*. In the correctable DOF, again, the movements were large in the *z* axis as expected, indicating that natural jaw opening was not restricted. The probe movements were negligible in the other two correctable DOF (*rx* and *y*), which are not affecting the naturalness in speech production.

5.3. Consistency of probe-to-head distance

An alternative measurement that primarily reflects the *z* DOF is to measure the range or consistency of the distance between probe and head fixed positions, denoted as the Δd measurement here (Fig. 7). We replicated such a measurement as reported in Pucher et al. (2020, Fig. 5, 6 & 9) by calculating the distribution of the Euclidean distance between the sensor placed at the upper incisor and the sensor at the probe.

Fig. 11 shows that results for the three experiments, derived from the same samples as in Fig. 8 and 10. The numbers at the top of each column indicate the 95% quantile interval (97.5th percentile minus 2.5th percentile). The values of 95% quantile interval of Δd for the three experiments were 5.7, 8.1 and 3.9 mm, respectively. All of them are less than any of the measurements (excluding mean and SD) of relative probe movements in the *z* axis for the corresponding experiment as in Table 2 and 3, indicating underestimation of the true errors. We further calculated the correlations between Δd and relative probe movements in *x*, *y*, and *z*, separately for each experiment. As shown in Table 4, the pattern of variations in Δd highly correlates with those in the probe movements in the inferior-superior (*z*) axis but not in the other two DOF, indicating that the measurement of Δd is biased toward vertical probe movements.

5.4. Corrected ultrasound tongue edge contours

In this section, we present the results of the ultrasound tongue contours after the head/probe movements were compensated for, using the proposed probe stabilization system. Figs. 12 and 13 demonstrates the comparisons of the uncorrected (left columns) and corrected (right columns) ultrasound tongue contours taken from the first (Expt. 1) and second (Expt. 2) experiments, respectively. Each of the thin lines indicates a repetition of a vowel. The thick curve above all thin lines indicates the location of the hard palate traced by an EMA sensor. The circle and triangle symbols in the left panels indicate the means (across repetitions) of the most posterior and most anterior points in the contours, while those symbols in the right panels show the time-synchronized positions of EMA sensors for each individual token. Each sample was extracted at the acoustic midpoint of the vowel.

5.4.1. Taiwan Mandarin

Fig. 12 show the phonemic contrasts of /a/ vs. /a/ (first row), /a/ vs. /ou/ (second row) and /i/ vs. /ei/ (third row) in Taiwan Mandarin. As shown in the first row of Fig. 12, the uncorrected tongue contours (left panel) do not demonstrate a clear distinction between /a/ and /a/. After the tongue contours were corrected by adjusting for the relative probe movements, the contrast between the two vowels becomes distinct. The mean absolute differences in probe position (in head space) between

/a/ and /a/ were: 0.24°, 1.72°, 0.29°, 2.17 mm, 0.06 mm, and 5.27 mm for roll, pitch, yaw, *x*, *y*, and *z*, respectively.

The second row presents the results of correction for the contrast between /a/ and /ou/. In the uncorrected representation (left panel), the two vowels exhibit distinct tongue shape throughout. Post-correction (right panel), there is a considerable overlap in the anterior part of the tongue, while the contrast between the two vowels is primarily observed in the posterior part of the tongue.

The third row displays corrections for /i/ and /ei/. Here, the uncorrected and corrected tongue contours for /i/ and /ei/ lack clear distinction. This is expected since high front vowels like /i/ and /ei/ typically involve minimal jaw movement, and thus stabilizing the probe with the headset is sufficient to maintain the contrast in tongue positions between /i/ and /ei/.

5.4.2. American English

Fig. 13 demonstrates the phonemic contrasts of /a/ vs. /ɛ/ vs. /i/ (first row), /ʌ/ vs. /ɔ/ (second row), and /i/ vs. /ɪ/ (third row) in American English.

In the first row, the contrasts among the three vowels were seemingly distinguishable in the uncorrected tongue contours (left panel) by a small margin. However, after correction (right panel), the differences in tongue positions are more salient and represent the full extent of the contrasts between the vowels – the plausible vowel contrasts in uncorrected contours appear to be an underestimation.

In the second row, the distinction between /ʌ/ and /ɔ/ is moderately observable in the uncorrected representations (left panel) primarily in the anterior part of the tongue, with a certain degree of overlapping. Post-correction (right panel), the tongue shapes between the two vowels are unambiguously separated in the anterior region.

Lastly, similar to the results of /i/ and /ei/ in Taiwan Mandarin shown above, the third row of Fig. 13 does not show clear distinction in tongue shape between uncorrected (left panel) and corrected (right panel) representations for /i/ and /ɪ/ in American English, probably due to slight differences in jaw position between the two high front vowels.

6. Discussion

6.1. Open source, modulation and adaptation

The 3D-printed ultrasound probe stabilizer presented here was derived from the one published in Derrick et al. (2018) and revised since then based on the feedback from our experiments. One of the aims of these updates has been to improve the probe stabilizer's adaptability. The current design makes most components modular and adjustable to better adapt to speaker physiology, machine-intrinsic characteristics, and application-specific requirements. For example, we learned from our experiments that the pitch angle of the probe with respect to the main frame of the headset needs to be adjusted in order to obtain ultrasound images with the most optimal view (e.g., more anterior or more posterior view, depending on the purposes of the experiment). In the current design, adjusting

¹ Note that the vowels /ou/ and /ei/ are diphthongs in Mandarin. These two examples are intended to show the effect of head/probe-correction but not to demonstrate linguistically meaningful distinctions.

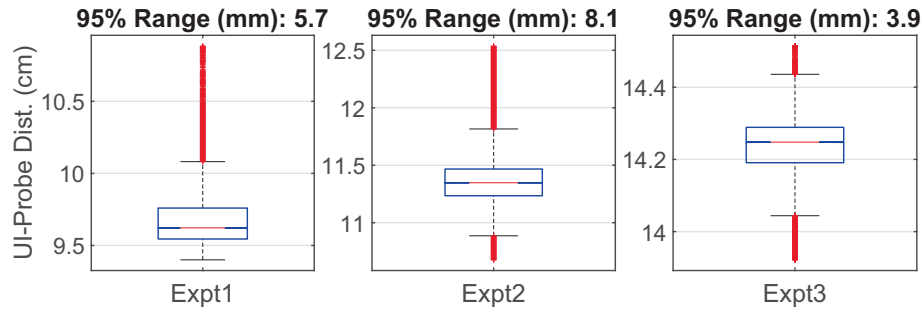


Fig. 11. Consistency (measured by the interval between the 2.5th and 97.5th percentiles) of probe to head distance (Δd) in Expt. 1 (left), Expt. 2 (middle) and Expt. 3 (right).

Table 4

Pearson correlation coefficients for the linear correlations between Δd and one of the relative probe movements in x, y, and z (columns) in each experiment (rows).

Expt	x	y	z
Expt. 1	-0.04	0.37	0.93
Expt. 2	0.2	-0.1	0.94
Expt. 3	0.17	0.06	0.8

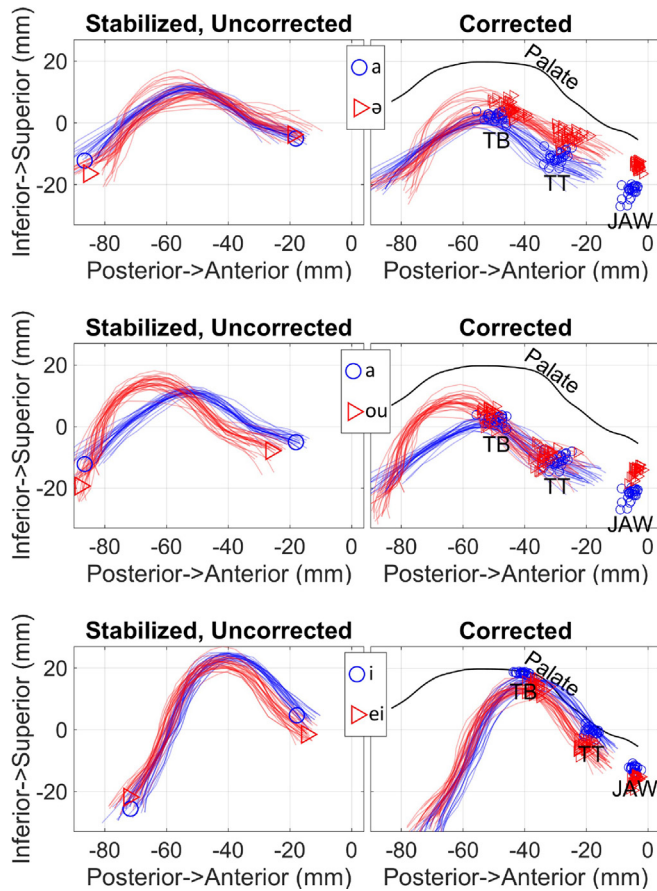


Fig. 12. Extracted tongue contours from Expt. 1 in uncorrected (left column) and corrected (right column) representations, for the contrasts of /a/ vs. /a/ (first row), /a/ vs. /ou/ (second row) and /i/ vs. /ei/ (third row) in Taiwan Mandarin. The mean probe positions in head spaces between /a/ and /a/ (first row) differed by 1.7° in pitch and 5.27 mm in z.

probe pitch angle can be done by switching between the “legs” (#2 in Fig. 3) in one of seven available angles.

Two other important updates include the ability to adapt to young children by re-positioning the arms (#5), cheek

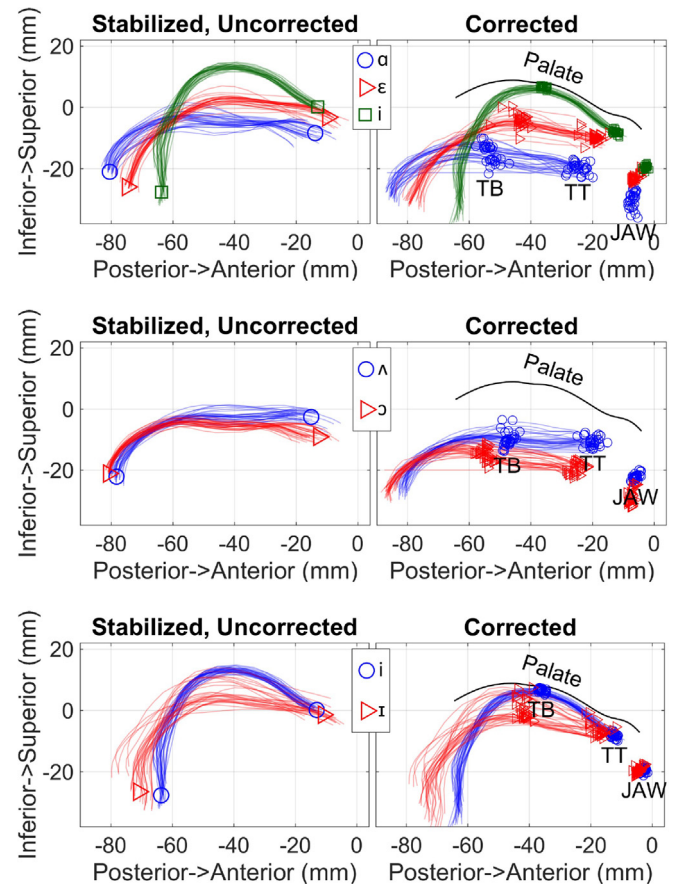


Fig. 13. Extracted tongue contours from Expt. 2 in uncorrected (left column) and corrected (right column) representations, for the contrasts of /a/ (circles) vs. /ε/ vs. /i/ (first row), /a/ vs. /a/ (second row), and /i/ vs. /i/ (third row) in American English.

adjusters (#10), and jaw flaps (#14), and the ability to change from midsagittal view to coronal view by replacing the hex rod (#3) with the coronal plane adapter within an experiment without repositioning the participant (Fig. 4).

For mid-sagittal imaging, it is crucial to align the probe to the speaker's mid-sagittal plane. This is equivalent to making the base (#4) perpendicular to this plane, since the base and probe are at a 90° angle. A common practice is that, at the beginning of the experiment setting, the experimenter adjusts the probe alignment based on visual inspection. The two bubble levels (#17) on the base were designed to aid such alignment. Assuming that when the speaker maintains an upright posture, the angle between the speaker's mid-sagittal plane

and the gravity vector is close to zero degrees, adjusting the roll angle of base until the bubbles are roughly at the center of the levels will facilitate the correct positioning of the probe during the setup of the stabilizer.

To ensure its sustainability, we made the 3D-printing codes open source and freely accessible on Github (<https://github.com/WeirongChen/ALPHUS>), and we hope that future updates can include more inputs from the community. Users can download and modify individual piece to fit their specific needs.

6.2. Validation and accuracy

The evaluation of the stabilization system replies primarily on assessing probe movements relative to the head. Such movements, if not compensated for, directly translate into errors of ultrasound tongue contours in the same magnitude. Our results indicate that, with the setting for midsagittal imaging, the relative probe movements were mostly (95% of samples) below 2 mm or 2° over uncorrectable degrees of freedom (DOF). They were up to 12 mm or 3° over correctable DOF (Table 2). For coronal imaging, 95% of the relative probe movements were lower than 3 mm or 3° in uncorrectable DOF and up to 6 mm in the correctable DOF (3). However, as shown in Table 2, the commonly used 95% criterion often underestimate the errors of interest; the actual error ranges (Max) were roughly twice the 95% quantile values. Therefore, we calculated the mean of the worst 1% of samples, *MW1%*, which includes the extreme values; these values increased only slightly (< 1 mm) from the 95% quantile values in the uncorrectable DOF, demonstrating the robustness of the system.

For comparison, Pucher et al. (2020, p. 89) reported that 95% of such movements (measured with a similar method) for the metallic UHS system were up to 13 mm in the anterior-to-posterior (correctable) DOF and 7 mm in the left-to-right (uncorrectable) DOF. An earlier restriction method by fixing the probe in a rigid mechanical arm and stabilizing the speaker's head with a metallic headrest demonstrated that the range of movement (maximum – minimum) was up to 4 mm (y) or 3° (roll) in uncorrectable DOF and 9 mm (z) or 6.5° (pitch) in correctable DOF (Gick et al., 2005, p. 512). More discussions regarding the accuracy of immobilizing ultrasound probe can be found in Stone (2005).

It is plausible to assume that the errors induced by head and probe movements may be smaller if measured within shorter time scales. However, the results demonstrated that the range of probe movements, as estimated by the mean of worst (largest) 1% of samples (*MW1%*) are comparable between those measured across experiments (Fig. 8) and those within a trial (Fig. 9), thus not supporting such an assumption.

We also assessed the consistency (range) of probe-to-head distance (Δd), an alternative evaluation proposed by Pucher et al. (2020) that reflects jaw opening but largely overlooks probe sliding underneath the chin (see Section 4.2 and Fig. 7). The data indicate that the range of probe-to-head distance underestimates the actual movements in all three experiments. Correlation analyses (Table 4) showed that Δd is heavily biased toward vertical probe movements and largely unrelated to lateral and posterior-anterior movements, as expected. Such underestimation is more prominent in Pucher et al. (2020), which reported that the 95% quantile interval of

probe-to-head distance was 2.7 mm and that of posterior-anterior movements was 33.6 mm for their speaker 2 with UltraFit. This is probably due to the differences in the design: the elastic bands connecting the top and bottom parts in our system allow the jaw (and probe) to move downward freely, which Δd is mostly related to, while the rigid arm structure in UltraFit restricts the vertical probe movement but the probe sliding is not as restricted, thus causing the largest movements in the posterior-anterior dimension, which is not captured by the consistency of probe-to-head distance. It is then inadvisable to use this measurement as an appropriate error metric for evaluating the accuracy or reliability of an ultrasound probe stabilizer.

6.3. Correction for probe-to-head movements

For 2D midsagittal ultrasound imaging of the tongue, the *uncorrectable* degrees of freedom (DOF) of the probe are those that move out of the imaging plane (i.e., lateral translation along y, rotation around anterior/posterior x (roll) and rotation around inferior/superior z (yaw)), and the *correctable* DOF are those that move within the plane (i.e., translation along x and z, and rotation around y (pitch)). To identify (as error screening signals) uncorrectable DOF probe movement and to compensate for correctable DOF movement, the positions of both the probe and the speaker's head must be tracked and synchronized with ultrasound recordings in 3D using point source tracking methods (e.g., EMA or Optotrak) as outlined in Whalen et al. (2005). However, the most prominent of these DOF for purposes of compensation is vertical (inferior/superior) translation along z induced by jaw displacement, which has the largest range of observed motion in our stabilization system. This can also be obtained using less accurate but more affordable video-based feature tracking algorithms requiring only a video camera as described in Mielke et al. (2005) and Noiray et al. (2020). When two cameras or a mirror angled at 45° is used, potentially all correctable DOF can be recovered.

Our results showed that correction for probe-to-head movement is necessary when the positional contrasts under investigation involve non-high or height-contrastive vowels, as demonstrated in Section 5.4; such correction may be less important when only high vowels are of interest (e.g., /i/ vs. /ei/ or /i/ vs. /ɪ/, as in the bottom panels of Fig. 12 and 13), due to bracing and the minimal jaw opening required for these vowels. Correction is also unnecessary when the focus of investigation relies on the analyses of tongue shape that are translation and rotation invariant, such as the degree of tongue curvature (e.g., Ménard et al., 2011; Dawson et al., 2016), the number of inflectional points (e.g., Preston et al., 2019) or the degree of tongue grooving in a coronal view (e.g., Whalen et al., 2011).

However, recent approaches to analysis of ultrasound tongue shapes using generalized additive models (GAM) rely on a conversion to polar coordinates in which the radius is the dependent variable and the angle an independent smoothing factor (e.g., Al-Tamimi and Palo, 2023; Coretta, 2019; Heyne et al., 2019); because this approach assumes that both radius and angle are based on a constant coordinate system related to speaker palatal hard structure, it requires both stabilization

of the probe and correction of the tongue contours for the probe-to-head movement, especially if non-high vowels are involved.

7. Conclusion

In the system described in this paper, movement is relatively unconstrained along the inferior-superior (z) axis in order to permit naturalistic jaw movement. Since this movement is correctable in post-processing, we are still able to accurately recover vocal tract shapes despite the large magnitude of this movement. While movement along the posterior-anterior (x) axis is also correctable (for sagittal imaging), we deliberately limit probe sliding in this DOF using the chin adjuster because it might result in missing the area of interest due to probe-intrinsic view angle limits.

Systems like our headset and AAA's UltraFit have the benefit of producing more natural speech by allowing flexibility. Such flexibility can easily result in more than 1 cm of error in the tongue contours, as shown in Pucher et al. (2020) as well as in the current study. Simply using those probe stabilizers does not correct for such error. Evidence demonstrates that, even when a probe stabilizer is used, relative probe movement as small as 2° in pitch or 5 mm in the vertical dimension, as in Expt. 1, can obscure the phonemic contrast between low and mid vowels (Fig. 12).

Ultrasound applications relying on accurate tongue position or vocal tract information, such as tongue posture comparisons or silent speech interfaces, suffer from insufficient and potentially biased inputs if probe movement information is ignored, and can thus result in inaccurate outputs. For such applications, correction for the probe movement relative to the head should be applied to the tongue contours in conjunction with a probe stabilizer. On the other hand, for applications where the shape of the tongue is of primary concern (e.g., tongue curvature or grooving) or investigations that only involves speech sounds requiring minimal jaw opening (e.g., high vowels, fricatives, etc.), acquiring tongue images with ultrasound and a probe stabilizer alone may be sufficient. Researchers should be aware of this limitation and choose appropriate stabilization and/or correction methods for their research questions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Work funded by NIH DC-002717 granted to Haskins Laboratories and to Yale University.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.wocn.2024.101339>.

References

- Al-Tamimi, J., & Palo, P. (2023). Dynamics of the tongue contour in the production of guttural consonants in levantine arabic. In *20th International Congress of Phonetic Sciences (ICPhS)*.
- Chang, Y. h. S. (2023). Effects of production training with ultrasound biofeedback on production and perception of second-language english tense-lax vowel contrasts. *Journal of Speech, Language, and Hearing Research*, 66, 1479–1495. https://doi.org/10.1044/2023_JSLHR-22-00587.
- Chen, W. R., Tiede, M., & Chen, S. (2017). An optimization method for correction of ultrasound probe-related contours to head-centric coordinates. *Ultrahfest VIII*, 75–76.
- Chen, W. R., Tiede, M., Whalen, D. (2020). DeepEdge: automatic ultrasound tongue contouring combining a deep neural network and an edge detection algorithm. In *The 12th International Seminar on Speech Production (ISSP 2020)*.
- Chiu, C., Wei, P. C., Noguchi, M., & Yamane, N. (2019). Sibilant fricative merging in taiwan Mandarin: An investigation of tongue postures using ultrasound imaging. *Language and Speech*, 63, 877–897. <https://doi.org/10.1177/0023830919896386>.
- Coretta, S. (2019). Assessing mid-sagittal tongue contours in polar coordinates using generalised additive (mixed) models. *OSF preprint*. <https://doi.org/10.31219/osf.io/q6vzb>.
- Dawson, K. M., Tiede, M. K., & Whalen, D. H. (2016). Methods for quantifying tongue shape and complexity using ultrasound imaging. *Clinical Linguistics & Phonetics*, 30, 328–344. <https://doi.org/10.3109/02699206.2015.1099164>.
- Derrick, D., Carignan, C., Chen, W. R., Shujau, M., & Best, C. T. (2018). Three-dimensional printable ultrasound transducer stabilization system. *The Journal of the Acoustical Society of America*, 144, EL392–EL398. <https://doi.org/10.1121/1.5066350>.
- Gick, B., Bernhardt, B., Bacsfalvi, P., & Wilson, I. (2008). Ultrasound imaging applications in second language acquisition. *John Benjamins, Amsterdam. Studies in bilingualism*, 309–322.
- Gick, B., Bird, S., & Wilson, I. (2005). Techniques for field application of lingual ultrasound imaging. *Clinical Linguistics & Phonetics*, 19, 503–514. <https://doi.org/10.1080/02699200500113590>.
- 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. <https://doi.org/10.3389/fpsyg.2019.02597>.
- Honorof, D.N., McCullough, J., Somerville, B., 2000. Comma gets a cure. *Diagnostic passage*.
- Lulich, S. M., & Pearson, W. G. (2019). Three-/four-dimensional ultrasound technology in speech research. *Perspectives of the ASHA Special Interest Groups*, 4, 733–747. https://doi.org/10.1044/2019_PERS-SIG19-2019-0001.
- Matosova, A. (2016). *Ultrafit. Thesis unpublished master thesis. Free University of Bozen*.
- Mielke, J., Baker, A., Archangeli, D., & Racy, S. (2005). *Palatron: a technique for aligning ultrasound images of the tongue and palate. Coyote Papers 14: Working Papers in Linguistics: Linguistic Theory at the University of Arizona*.
- Ménard, L., Aubin, J., Thibault, M., & Richard, G. (2011). Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model. *Folia Phoniatrica et Logopaedica*, 64, 64–72. <https://doi.org/10.1159/000331997>.
- Noiray, A., Ries, J., Tiede, M., Rubertus, E., Laporte, C., & Ménard, L. (2020). Recording and analyzing kinematic data in children and adults with SOLLAR: Sonographic & optical linguo-labial articulation recording system. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 11, 1–25. <https://doi.org/10.5334/labphon.241>.
- Preston, J. L., McAllister, T., Phillips, E., Boyce, S., Tiede, M., Kim, J. S., & Whalen, D. H. (2019). Remediating residual rhotic errors with traditional and ultrasound-enhanced treatment: A single-case experimental study. *American Journal of Speech-Language Pathology*, 28, 1167–1183. https://doi.org/10.1044/2019_AJSLP-18-0261.
- Preston, J. L., McAllister Byun, T., Boyce, S. E., Hamilton, S., Tiede, M., Phillips, E., Rivera-Campos, A., & Whalen, D. H. (2017). Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors. *JoVE*, e55123. <https://doi.org/10.3791/55123>.
- Preston, J. L., McCabe, P., Tiede, M., & Whalen, D. H. (2019). Tongue shapes for rhotics in school-age children with and without residual speech errors. *Clinical Linguistics & Phonetics*, 33, 334–348. <https://doi.org/10.1080/02699206.2018.1517190>.
- Pucher, M., Klingler, N., Luttenberger, J., & Spreafico, L. (2020). Accuracy, recording interference, and articulatory quality of headsets for ultrasound recordings. *Speech Communication*, 123, 83–97. <https://doi.org/10.1016/j.specom.2020.07.001>.
- Recasens, D., & Rodríguez, C. (2016). A study on coarticulatory resistance and aggressiveness for front lingual consonants and vowels using ultrasound. *Journal of Phonetics*, 59, 58–75. <https://doi.org/10.1016/j.wocn.2016.09.002>.
- Roxburgh, Z., Scobbie, J. M., & Cleland, J. (2015). Articulation therapy for children with cleft palate using visual articulatory models and ultrasound biofeedback. In *18th International Congress of Phonetic Sciences (ICPhS)*. International Phonetic Association.
- Scobbie, J. M., Wrench, A. A., & Van Der Linden, M. L. (2008). Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. In *Proceedings of the 8th International Seminar on Speech Production* (pp. 373–376).
- Shaw, J. A., Chen, W. R., Proctor, M. I., & Derrick, D. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research*, 59, S1566–S1574. https://doi.org/10.1044/2015_JSLHR-S-15-0031.

- Shawker, T. H., Stone, M., & Sonies, B. C. (1985). Tongue pellet tracking by ultrasound: development of a reverberation pellet. *Journal of Phonetics*, 13, 135–146. [https://doi.org/10.1016/S0095-4470\(19\)30741-7](https://doi.org/10.1016/S0095-4470(19)30741-7).
- Spreafico, L., Pucher, M., & Matosova, A. (2018). UltraFit: A speaker-friendly headset for ultrasound recordings in speech science. *Interspeech*, 2018, 1517–1520. <https://doi.org/10.21437/Interspeech.2018-995>.
- Stone, M. (1990). A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *The Journal of the Acoustical Society of America*, 87, 2207–2217. <https://doi.org/10.1121/1.399188>.
- Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, 19, 455–501. <https://doi.org/10.1080/02699200500113558>.
- Stone, M., & Davis, E. P. (1995). A head and transducer support system for making ultrasound images of tongue/jaw movement. *Journal of Acoustical Society of America*, 98, 3107–3112. <https://doi.org/10.1121/1.413799>.
- Stone, M., Sonies, B. C., Shawker, T. H., Weiss, G., & Nadel, L. (1983). Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. *Journal of Phonetics*, 11, 207–218. [https://doi.org/10.1016/S0095-4470\(19\)30822-8](https://doi.org/10.1016/S0095-4470(19)30822-8).
- Tabain, M., & Beare, R. (2018). An ultrasound study of coronal places of articulation in central arrernte: Apicals, laminals and rhotics. *Journal of Phonetics*, 66, 63–81. <https://doi.org/10.1016/j.wocn.2017.09.006>.
- Tiede, M., Chen, W.r., Whalen, D. (2019). Taiwanese Mandarin sibilant contrasts investigated using coregistered ema and ultrasound. In Calhoun, S., Escudero, P., Tabain, M., Warren, P. (Eds.), *International Congress of Phonetic Sciences (ICPhS)* (pp. 427–431).
- Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., Vatikotis-Bateson, E., & Hailey, D. S. (2005). The haskins optically corrected ultrasound system (hocus). *Journal of Speech, Language, and Hearing Research*, 48, 543–553. [https://doi.org/10.1044/1092-4388\(2005/037\)](https://doi.org/10.1044/1092-4388(2005/037)).
- Whalen, D. H., Shaw, P., Noiray, A., & Antony, R. (2011). Analogs of tahlitan consonant harmony in english cvc syllables. In W. S. Lee & E. Zee (Eds.), *17th International Congress of Phonetic Sciences (ICPhS XVII)* (pp. 2129–2132).