Importance of Parasagittal Sensor Information in Tongue Motion Capture through a Diphonic Analysis

Salvador Medina[1,*], *Sarah Taylor*[2,*], *Mark Tiede*[3], *Alexander Hauptmann*[1], *Iain Matthews*[4]

[1]Carnegie Mellon University, USA
[2]University of East Anglia, UK
[3]Haskins Laboratories, USA
[4]Epic Games, USA

salvadom@cs.cmu.edu, s.l.taylor@uea.ac.uk, tiede@haskins.yale.edu,
alex@cs.cmu.edu, iain.matthews@epicgames.com

Abstract

Our study examines the information obtained by adding two parasagittal sensors to the standard midsagittal configuration of an Electromagnetic Articulography (EMA) observation of lingual articulation. In this work, we present a large and phonetically balanced corpus obtained from an EMA recording session of a single English native speaker reading 1899 sentences from the Harvard and TIMIT corpora. According to a statistical analysis of the diphones produced during the recording session, the motion captured by the parasagittal sensors has a low correlation to the midsagittal sensors in the mediolateral direction. We perform a geometric analysis of the lateral tongue by the measure of its width and using a proxy of the tongue's curvature that is computed using the Menger curvature. To provide a better understanding of the tongue sensor motion we present dynamic visualizations of all diphones. Finally, we present a summary of the velocity information computed from the tongue sensor information.

Index Terms: Tongue, Parasagittal, Electromagnetic Articulography, EMA, Articulatory Analysis

1. Introduction

The tongue is fundamental to shaping the sounds of speech. A dynamic model that describes the relationship between tongue motion and acoustic speech is key to applications such as animating talking heads, speech synthesis, acoustic-to-articulatory inversion and automatic speech recognition, and may inform methods for providing feedback during speech therapy.

Until recently, continuous speech and tongue datasets predominantly captured midsagittal deformations of the tongue and, consequently, acoustic-articulatory modelling has been approximated with 2-D dynamics. By not measuring parasagittal deformations, the tongue is assumed to be horizontal in the coronal plane, and this does not accurately represent the many degrees of freedom that enable human tongues to roll, twist and otherwise non-rigidly deform.

We present a novel multi-modal speech and tongue dataset which recorded midsagittal tongue sensors and two additional sensors placed parasagittally to capture complex 3-D tongue deformations. We consider this work to be a step forward towards obtaining a better representations of the tongue's surface since alternatives like volumetric Magnetic Resonance Imaging (MRI) scans and 3D ultrasound imaging have inappropriately slow sample rates. We will release our data which we believe forms the first large-scale EMA English dataset of continuous speech, and tongue, lips and jaw motion that includes two parasagittal sensors on the tongue.

From our data we present a diphone-level statistical analysis on the dynamics of the tongue during speech with a focus on the parasagittal motion. Specifically, we aim to determine the following: 1) To what extent and for which diphones is the lateral tongue actively controlled? 2) What are the characteristics of the lateral tongue width and curvature during continuous speech? We additionally present a visualization of the sensor motion of each diphone to provide a greater understanding on tongue dynamics during speech production.

2. Related Work

2.1. Tongue motion acquisition

Tongue motion has previously been acquired through X-ray imagery [1, 2], but radiation exposure makes large-scale data collection unfeasible. Real-time MRI [3, 4, 5] and 3D ultrasound [6] are safer options, but the resulting unregistered images make it challenging to track fiduciary points on the tongue over time, and these methods suffer from slow sampling rates. Although more intrusive, EMA can measure sensor position and orientation at fixed locations on the tongue with high spatial and temporal resolution and low error [7].

The MOCHA-TIMIT [8] corpus is a phonetically balanced dataset of 460 sentences read by two British English speakers. The articulatory data is captured in different modalities from Electropalatography (EPG), Laryngography and EMA in a midsagittal configuration. EMA was also used for capturing tongue motion in [9] for 320 utterances of Austrian German speech, to construct the mngu0 dataset [10] which contains 1354 utterances. In [11], Dutch and English speakers recited a short phrase and isolated words, while in [12], 3 Italian speaker were captured reading 500 Italian sentences providing approximately 2 hours of speech. EMA sensors are generally placed midsagittally along the tongue for capturing 2D deformation of the tongue tip, body and dorsum [7]. Although the parasagittal motions of the tongue contribute to speech production, they are largely overlooked during data collection.

There has been some prior work that considered lateral tongue motion [13] to study the production of /l/ in Australian English with the aide of two parasagittal sensors acquired at a rate of 100 Hz. The work presented by [14] included one

[*] Equal contribution.

parasagittal sensor to examine the contribution of lateral motion on the production of alveolar consonants in vowel-consonant-vowel syllables. Their findings indicate that lateral motion is fundamental for articulating the sound /z/. Two parasagittal sensors were included in the capture by [15] and [16] who respectively studied the articulation of Czech liquids in isolated nonsense words and English liquids in carrier sentences by Japanese speakers.

2.2. Tongue dynamics during speech production

The work in [2] analyzed patterns of deformations of the midsagittal edge of the tongue in transitions between lingual segments from X-Ray images. An analysis of tongue motion during emotive speech revealed that the vertical motion of the tongue dorsum is dampened during sad speech [5]. A study of vowel-consonant-vowel syllables in [17] revealed that tongue width is largest for palatal plosives and fricatives as the tongue widens as it is pressed against the hard palate, and smallest for velar plosives and fricatives, since the tongue body volume is largely retracted towards the velum. The work in [13] investigated tongue lateralization in the Australian production of /l/ and discovered that the lateral tongue is actively controlled rather than moving as a bi-product of tongue stretching. In [18], video recordings of the tongue during the articulation of an English passage revealed that bilateral movements are asymmetric and one side of the tongue typically moves ahead of the other depending on the speaker.

The majority of previous work performs analysis on isolated or nonsensical words, and there has been very little research into the 3-D tongue motion during continuous speech production. An exception to this is the work in [19] which presented a statistical technique for identifying critical, dependent and redundant roles played by the articulators during production of the English phonemes in the MOCHA-TIMIT corpus. They found that fricatives and affricates required the most number of critical articulators, and none were identified for the alveolar /l/. They additionally observed that the articulatory system comprised of three largely-independent components: the lip and jaw group, the tongue, and the velum.

3. Data

Our data consists of a single male English native speaker, reading 1899 sentences providing a total of 2.5 hours of speech audio. A subset of 720 sentences is from the Harvard set [20] which was read twice at a normal and fast pace. The remaining sentences were a subset of the TIMIT dataset [21].

Acoustics and articulatory movement were recorded using a Carstens AG501 EMA device. Sensors were attached to speech articulators using medical-grade cyanoacrylate glue. Three sensors were placed midsagittally on the tongue surface, one sensor on the tongue dorsum (TD), one on the tongue blade (TB), and one behind the tongue tip (TT). Two more sensors were parasagittally placed to the left (BL) and right (BR) of the tongue blade. Three additional sensors were placed on the lips, two were midsagittally attached on the upper (UL) and lower lips (LL) at the vermilion border, and one on the right corner (LC) of the lips. Additionally, two sensors were placed at the gingival border for the upper (UI) and lower (LI) medial incisors and between the canine and first premolar on the lower jaw (LJ). See Figure 1 for sensor placement.

The EMA sensor trajectories and single-channel acoustic data were synchronously captured at 250 Hz and 48 kHz respectively. Articulatory data was downsampled to 50 Hz, and corrected for head movement by rotating and translating to the occlusal plane using a reference biteplane.

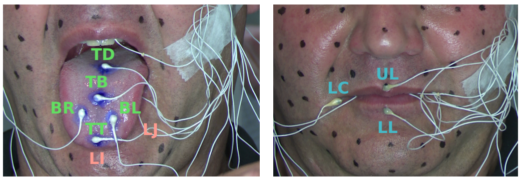


Figure 1: *EMA sensor configuration for tongue motion capture used in this work.*

An approximation of the surface of the palate was captured through three traces using one of the transducers glued to the tip of a wooden stem after removing the cotton swab as proposed by [22]. One trace followed the midsagittal curve, and two traces were captured through an alternating movement in a sagittal direction and the other trace in a coronal direction from the upper incisors to the posterior of the palate before the subject would feel any discomfort. We reconstructed the surface of the palate by fitting a plane to the traces using the 3-D software Blender.

In this work we focus only on the analysis of the parasagittal sensors of the tongue. The full processed and filtered data will be made publicly available for further research.

4. Importance of the Lateral Tongue

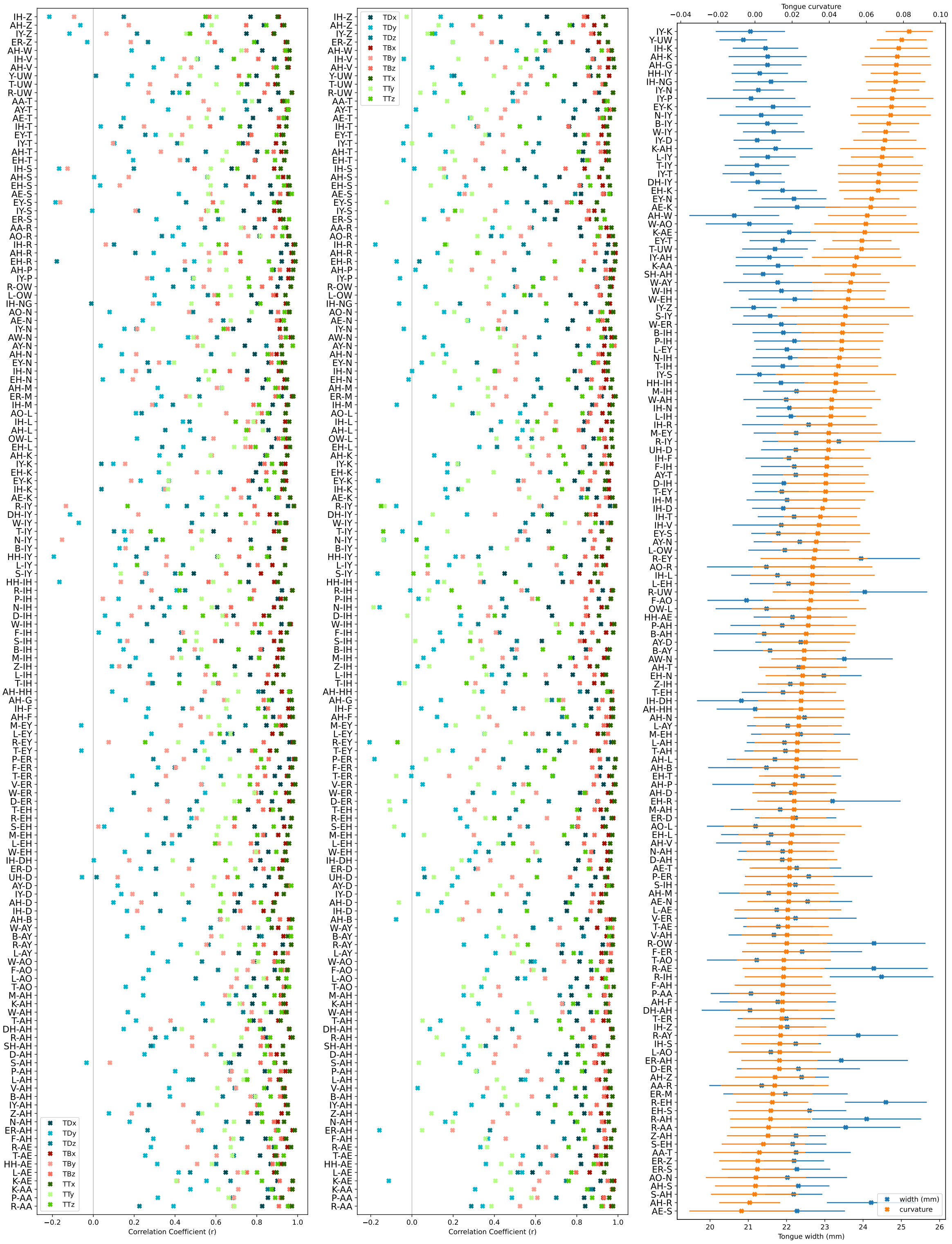
We use the Montreal Forced Aligner [23] to extract the diphone segments from the audio. For our analysis we ignore the diphones with silence or non-speech segments. This results in a total of 1,158 diphones from which 424 are consonant clusters. The remaining 734 diphones are distributed as follows: 305 vowel-consonant, 315 consonant-vowel, and 114 vowel-vowel. We filter out the consonant clusters and diphones with fewer than 86 examples resulting in 142 unique diphones which covers 60.4% of the non-consonant cluster data.

4.1. Relationship between mid and parasagittal sensors

We first investigate the extent to which the parasagittal sensors deform with respect to the midsagittal sensors to identify the sounds where the parasagittal deformations are largely independent from the midsagittal motion. We compute the Pearson correlation coefficient (r) for each midsagittal tongue sensor (TD, TB, TT) to each parasagittal sensor (BL, BR) independently for each of the x (anterior/posterior), y (left/right) and z (superior/inferior) axes. The complete set of correlations for BL and BR are respectively shown in Figures 2a and 2b.

We observe a high correlation of the parasagittal sensors with all the midsagittal sensors on the x-axis, demonstrating that during regular speech the surface of the tongue moves back and forth in a consistent manner. Moreover, we observe that the parasagittal sensors correlate most with TT, confirming the discoveries in [14], and are least correlated to the TB and TD sensors in the coronal plane with a prominent difference on the y-axis for particular diphones. Specifically, we observe very low and slightly negative correlations with TD and TB in the coronal plane for the diphones that end with the alveolars /z/, /s/, /d/ or /n/. We find this effect to be less prominent for alveolar /t/. The same effect can be seen in diphones ending with the front unrounded vowels /i/ and /ɪ/.

The results suggest that the lateral tongue is actively con-

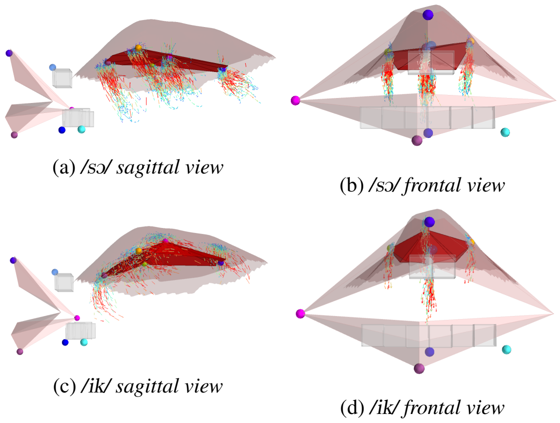


(a) *BL correlation*

(b) *BR correlation*

(c) *width and curvature*

Figure 2: *(a) and (b) Correlation of midsagittal tongue sensors to left and right parasagittal sensors shown for each diphone and axis. (c) Distance between the parasagittal sensors in mm and proxy curvature (BL-TB-BR) in the coronal plane for each diphone.*



(a) */sɔ/ sagittal view*

(b) */sɔ/ frontal view*

(c) */ik/ sagittal view*

(d) */ik/ frontal view*

Figure 3: *Data visualization of diphones /sɔ/ and /ik/. Sensors are color spheres. Sensor motion is represented by colored quivers (rainbow). The tongue pose is the mean of the mid-position of the second phone.*

trolled and does not move merely as a bi-product of midsagittal activity. Parasagittal sensors move independently of the mid-tongue sensors to the greatest extent in the coronal plane. This could be indicative of a) lateral curvature or b) a widening or narrowing of the superior surface to preserve tongue volume as it deforms. We further investigate this in the following sections.

4.2. Diphonic tongue width and curvature

We compute the 3-D Euclidean distance between the left and right parasagittal sensors as a proxy of the tongue width. Furthermore, to determine the extent of tongue roll and its relationship to the underlying speech. We compute the Menger curvature [24] as a measure of tongue curvature in the coronal plane for each diphone using three 2-D points corresponding to the y and z axes of BL, TB and BR. A negative value represents a curled upward tongue surface and a positive value indicates a curled downwards pose, while a zero value indicates a flat tongue. In Figure 3b we visualize a diphone with slightly negative curvature showing a close to flat tongue, while in Figure 3d we see an example with a high positive curvature.

The means and standard deviations of tongue width and curvature for each diphone can be found in Figure 2c shown in ascending order of curvature. We generally observe that tongue width negatively correlates with curvature ($r = -0.384$). This is intuitive since the sensors become closer as the edges of the tongue curl up. At the top of the graph we observe a cluster of diphones containing the velar consonants /k/, /g/, and /ŋ/ paired with vowels /i/, /ɪ/ and /ʌ/. These are associated with a relatively narrow tongue and large downwards curvature of the lateral tongue. They are followed by a cluster of diphones containing the vowel /i/ with a range of consonant contexts that have diverse places of articulation. However, outliers appear when /i/ is spoken in the context of the alveolar fricatives /ʃ/ and /ʒ/, where we observe that the tongue curvature is approximately halved. The diphones that contain /ʃ/ and /ʒ/ appear towards the bottom of the graph, although /ʒ/ is distributed more uniformly throughout the lower half. The outliers are therefore the result of co-articulation that stems from transitioning between a flat or upwards curled tongue to a downwards curvature and vice versa. This result indicates that parasagittal tongue motion is important for producing each of these sounds.

4.3. Dynamics of the parasagittal sensors

Our geometric analysis of the parasagittal sensors is indicative of the shape of the tongue, but tongue dynamics are lost. We present visualizations of all diphones as quiver plots and exemplar videos[1] for better understanding of the tongue sensor motion. Figure 3 shows the frontal and sagittal view of diphones /sɔ/ and /ik/. All the EMA sensors are represented by colored spheres. The images show the palate surface reconstruction. The lips and teeth are not a reconstruction but serve as reference for a better spatial understanding. The tongue's pose shown is the mean of the mid-poses from the second phone. The color-coded quivers represent the motion of the sensors from all the samples in the data for the given diphones. The sequence of colors from start to end are the colors of the rainbow from violet to red. In Figure 3a we can observe how /sɔ/ starts with the tongue tip close to the alveolar ridge (violet) followed by a rapid gesture that moves the tongue downwards (cyan) and back to a stationary position (red). In Figure 3c, we can appreciate how the curved transition of /ik/ begins with a quick constriction on the palate and ends with a low frontal tongue pose.

To gain insight into the tongue's motion statistics, we compute peak velocities of the five tongue sensors for all diphone samples and calculate the mean of the velocities for each diphone class. In our analysis, we found that the diphones with alveolar and post-alveolar fricatives /z/, /s/, and /ʃ/ show low mean peak velocity below 40 mm/s due to the long periods in which the tongue remains stationary. Alternatively, the diphones with the highest velocities above 180 mm/s require an open or close movement of the jaw such as /ʌr/, /kɔ/, /ɑt/, and /ɑk/.

5. Conclusion

We introduced a large and phonetically balanced corpus from a single English speaker from an EMA capture that includes 2.5 hours of speech and the articulation of the lips, jaw, and tongue with the addition of two parasagittal sensors to the traditional midsagittal configuration. We presented a correlation analysis at a diphonic level, demonstrating that both parasagittal sensors have a low correlation to the midsagittal sensors in the mediolateral direction which indicates that they contribute independently to speech production. The enriched information from the parasagittal sensors also allows us to determine an approximation to the width and curvature of the tongue from which we determined the characteristics of each diphone. We discovered that the vowel /i/ and alveolar consonant /ʃ/ exhibit co-articulatory effects when spoken in sequence. We have presented visualizations of the motions of all diphones in our data and made these publicly available. We believe our corpus will enable further research in continuous speech with a higher level of detail and the training of data-driven models for applications such as acoustic-articulatory inversion.

6. Acknowledgements

We would like to thank Kevin Munhall for inspiration and guidance, and Wei-rong Chen, Jaekoo Kang, and Boram Kim for their assistance in the data capture. Sarah Taylor was supported by the Engineering and Physical Research Council (Grant number EP/S001816/1).

[1] `https://salmedina.github.io/ContinuousTongueMotionAnalysis/`

7. References

[1] R. Harshman, P. Ladefoged, and L. Goldstein, "Factor analysis of tongue shapes," *The Journal of the Acoustical Society of America*, vol. 62, no. 3, pp. 693–707, 1977.

[2] K. Iskarous, "Patterns of tongue movement," *Journal of Phonetics*, vol. 33, no. 4, pp. 363–381, 2005.

[3] J. Woo, F. Xing, M. Stone, J. Green, T. G. Reese, T. J. Brady, V. J. Wedeen, J. L. Prince, and G. El Fakhri, "Speech map: A statistical multimodal atlas of 4D tongue motion during speech from tagged and cine MR images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 7, no. 4, pp. 361–373, 2019.

[4] W. Chen, D. Byrd, S. Narayanan, and K. S. Nayak, "Intermittently tagged real-time MRI reveals internal tongue motion during speech production," *Magnetic resonance in medicine*, vol. 82, no. 2, pp. 600–613, 2019.

[5] J. Kim, A. Toutios, S. Lee, and S. S. Narayanan, "Vocal tract shaping of emotional speech," *Computer Speech & Language*, vol. 64, p. 101100, 2020.

[6] S. Charles and S. M. Lulich, "Articulatory-acoustic relations in the production of alveolar and palatal lateral sounds in brazilian portuguese," *The journal of the Acoustical Society of America*, vol. 145, no. 6, pp. 3269–3288, 2019.

[7] T. Rebernik, J. Jacobi, R. Jonkers, A. Noiray, and M. Wieling, "A review of data collection practices using electromagnetic articulography," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 12, no. 1, 2021.

[8] A. Wrench, "The mocha-timit articulatory database," 1999.

[9] D. Schabus, M. Pucher, and P. Hoole, "The MMASCS multimodal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech," in *LREC*, 2014, pp. 3411–3416.

[10] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Interspeech 2011: 12th Annual Conference of the International Speech Communication Association*, 2011, pp. 1505–1508.

[11] G. Sivaraman, C. Y. Espy-Wilson, and M. Wieling, "Analysis of acoustic-to-articulatory speech inversion across different accents and languages." in *INTERSPEECH*, 2017, pp. 974–978.

[12] C. Canevari, L. Badino, and L. Fadiga, "A new italian dataset of parallel acoustic and articulatory data," in *INTERSPEECH*, 2015.

[13] J. Ying, J. A. Shaw, C. Carignan, M. Proctor, D. Derrick, and C. T. Best, "Evidence for active control of tongue lateralization in australian english /l/," *Journal of Phonetics*, vol. 86, p. 101039, 2021.

[14] W. F. Katz, S. Mehta, M. Wood, and J. Wang, "Using electromagnetic articulography with a tongue lateral sensor to discriminate manner of articulation," *The Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. EL57–EL63, 2017.

[15] P. Howson and A. Kochetov, "An EMA examination of liquids in czech," in *The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences*, no. 488, 2015.

[16] J. Moore, J. Shaw, S. Kawahara, and T. Arai, "Articulation strategies for English liquids used by Japanese speakers," *Acoustical Science and Technology*, vol. 39, no. 2, pp. 75–83, 2018.

[17] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Communication*, vol. 41, no. 2, pp. 303–329, 2003.

[18] B. Gick, M. Keough, O. Tkachman, and Y. Liu, "Lateral bias in lingual bracing during speech," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1903–1903, 2018.

[19] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695–710, 2009.

[20] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *STIN*, vol. 93, p. 27403, 1993.

[22] C. Neufeld and P. van Lieshout, "Tongue kinematics in palate relative coordinate spaces for electro-magnetic articulography," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 352–361, 2014.

[23] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[24] J.-C. Léger, "Menger curvature and rectifiability," *Annals of mathematics*, vol. 149, no. 3, pp. 831–869, 1999.