MARKERLESS TONGUE POSE ESTIMATION IN CROSS-LINGUISTIC SPOKEN AND MIMED SPEECH USING ULTRASOUND

Annslin Maria Lukose, Amelia Gully, George Bailey

University of York, York, UK
annslinmaria46@gmail.com, amelia.gully@york.ac.uk, george.bailey@york.ac.uk

ABSTRACT

Studies of articulatory variation in mimed speech, a proxy for alaryngeal speech, have been restricted to monolingual speakers. This study presents a single-subject case study that investigates cross-linguistic phonological influence of vowels in a 25-year-old female Malayalam-English bilingual speaker, producing spoken and mimed speech. Ultrasound tongue imaging is used to acquire articulatory data, and novel techniques for data analysis are used including pose estimation with DeepLabCut™ and geometric morphometric analysis (GMM) for quantitative measurement of tongue variation. The results reveal significant differences in tongue kinematics across languages and conditions, without direct indication of phonological transfer. However, within-speaker articulatory variation is evident across the independent variables. These findings will aid in the development of silent speech interfaces for voice restoration in bilingual laryngectomy patients.

Keywords: Ultrasound tongue imaging, pose estimation, Geometric morphometrics, mimed speech

1. INTRODUCTION

A bilingual individual is someone who, in their everyday lives, uses two (or more) languages or dialects [11]. Based on the Speech Learning Model (SLM), the sounds comprising the L1 and L2 phonetic subsystems are interconnected, and this interaction occurs because the L1 and L2 sounds exist in a common "phonological space" [10]. As per the revised Speech Learning Model (SLM-r), one reason for the difference in learning outcomes of L1 and L2 is that, since the sounds of L2 are automatically related to the phonetic inventory of L1, L2 sounds are initially replaced by L1 sounds. This would then result in cross-linguistic influence or phonological transfer, which refers to “the ways in which a person’s knowledge of the sound system of one language can affect that person’s perception and production of speech sounds in another language” [13].

Due to their absence of larynxes, laryngectomy patients can only mime speech, and they rely on different voice restoration methods to achieve communication with their alaryngeal speech. Some studies have compared articulation across normal, whispered and mimed speech, for monolingual English [9] [20], and French [4] speakers. The current study is the first investigation to study articulatory variation in spoken and mimed speech conditions in two languages spoken by a single speaker.

With the implementation of machine learning tools, speech can be predicted from articulation in a technique called articulatory-to-acoustic (forward) mapping [5], which was further utilised to develop ‘Silent Speech Interface’ systems (SSI). SSI involves recording motions of the articulators in the absence of audible acoustic signals and automatically synthesising speech based on those movements [6], [7]. Laryngectomy patients are suitable for ultrasound-based SSI since neither glottal excitation nor vocal tract airflow are necessary [7]. Ultrasound tongue imaging (UTI) is a non-invasive and safe technique [3]. Additionally, with regard to cost of equipment, safety, portability and structures visualised, UTI is a more cost-effective technique compared to electromagnetic articulography (EMA) [16].

The objective of the current study is to obtain ultrasound images of the tongue to examine the phonological transfer or cross-linguistic articulatory variation of target vowels in a bilingual speaker speaking in two different conditions, one mimicking alaryngeal speech. An additional goal is to understand within-speaker articulatory variation across languages and conditions. The potential future application of this study is the improvement of ultrasound-based silent speech interfaces for bilingual laryngectomy patients. This paper is a single-subject case study that investigated speech samples collected in mimed speech (approximating alaryngeal speech) and spoken speech conditions from a Malayalam-English bilingual speaker, language variations which are not well studied particularly in this context. Malayalam is a south Indian language with five short and long vowels /ɪ e a o u i: a: o: u:/ [14], and two diphthongs (/ai/ and /au/). Indian English has a vowel system quite similar to the RP variety of British English, with [e] and [o] (monophthongs) replacing the lexical sets FACE and GOAT (diphthongs), respectively [22], which might be due to phonological transfer from L1 to L2.

In order to achieve the research objectives, tongue motion patterns during spoken and mimed production were collected using UTI. This was followed by assessing the cross-linguistic phonological influence of vowels in these speaking conditions and estimating within-speaker articulatory variation using state-of-the-art techniques such as pose estimation of the tongue using DeepLabCut™ (DLC) [21] and geometric morphometric (GMM) analysis. DLC was first implemented on tongue and lip movements by Wrench and Balch-Tomes [23]. Unlike manual labelling or semi-automatic labelling used in Articulate Assistant Advanced version 219.08 [2], DLC is an innovative method for obtaining tongue splines using machine learning, where the tongue contours are automatically landmarked, making numerical measurement of the articulatory kinematics possible.

2. METHODS

2.1. Data collection

A 25-year-old female native Malayalam speaker, whose second language was English, participated in this study.

Four long vowels (/i:/, /e:/, /o:/, and /u:/) at two vowel positions (initial and medial) in both Malayalam and English were investigated for this research, in keeping with similar studies (e.g., [4]). Final vowel position was excluded due to an insufficient number of meaningful words across both languages. Monosyllabic and bisyllabic words with vowels around alveolars and velars were used as stimuli, to ensure that tongue movements were clearly visible in the mimed condition, where audio cues were not available. A clap was also introduced before the target vowel while recording in the mimed speech condition in order to aid with phoneme segmentation during the analysis process. Two words each at initial and medial vowel positions were selected, and three repetitions of each of these words were recorded in both spoken and mimed conditions. A summary of the data set is displayed in Table 1. In total, there were 96 tokens.

Vowels	4: /i:/, /e:/, /o:/ and /u:/
Positions	2: Initial and medial
Words	2 words for each vowel at each position
Conditions	2: Spoken and mimed speech
Repetitions	3 repetitions of each word for each condition

Table 1: A summary of the data set

The midsagittal plane of the tongue for each repetition was captured with ultrasound tongue imaging (UTI), using the software Articulate Assistant Advanced (AAA), version 219.08 [2]. A MC4-2R20S-3 microconvex-array ultrasound transducer with a probe frequency ranging from 2 MHz to 4 MHz was used to collect ultrasound recordings of the tongue. The UltraFit [19] headset was used to hold the ultrasound probe in place. Audio was recorded directly into AAA using a Rode smartLav+ wired lavalier microphone at 22050 Hz sampling frequency.

The participant completed the recordings separately for each language, with a break after the ultrasonic recordings for one language. Considering the comfort of the participant, the probe was adjusted between languages. For each language, every word was recorded three times in the spoken speech followed by three times in the mimed condition.

2.2. Pose estimation

After obtaining the ultrasonic videos of the tongue for each repetition of the words, DeepLabCut™ was used to obtain tongue splines for each recording. An illustration of the pose estimated tongue contour obtained using DLC is shown in Figure 1 (first repetition of the word ‘ache’ in English in the spoken condition). Figure 1 depicts 11 key points along the surface of the tongue corresponding to the vallecula, tongue root (2x), tongue body (2x), tongue dorsum (2x), tongue blade (2x) and tongue tip (2x).

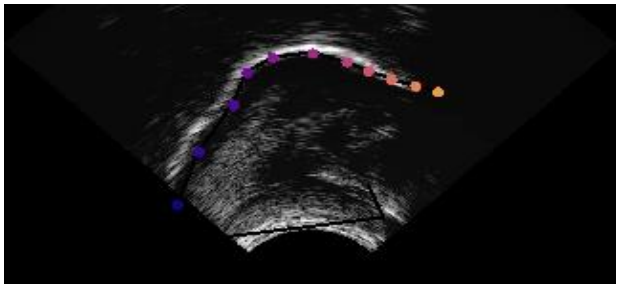

Figure 1: Pose estimated tongue contour for the word ‘ache’ in English in spoken speech condition

2.3. Data analysis

The midpoint method [8] was chosen to measure the target vowels.

In this research study, GMM was applied to the dataset in order to facilitate comparison of tongue shapes. GMM, developed by Bookstein, 1996a (as cited in [1]), is a statistically robust framework for quantifying and comparing shapes. The first step is Generalised Procrustes Analysis (GPA), a superimposition method [1] that minimises

differences in shape [15] by eliminating non-shape variation in the arrangement of landmarks [1] including scale, translation, and rotation. This has the added benefit of normalising any differences in position between recording sessions that may have occurred due to movement of the ultrasound probe. The next step is to quantify the variation in tongue shape across both languages and conditions using principal component analysis (PCA). For GPA-normalised shape data, PCA detects the axes of greatest shape variation in the dataset [15]. The first principal component (PC1) indicates the most shape variation in the data set, the second principal component specifies the second most shape variation in the data set, and so on. PCA has been used to quantify shape in the vocal tract [12], but its application to ultrasound-derived data is new.

3. RESULTS AND DISCUSSION

3.1. Visual analysis

Observations of Figure 2 suggests that the position of the tongue for both conditions vary noticeably across vowels, with obvious tongue height and tongue advancement deviations. For both the languages, most of these variances were seen in vowels /o:/ and /u:/ for both initial vowel position and medial vowel position. A small variation in tongue advancement could be observed for vowels /e:/ and /i:/ in the initial vowel position and medial vowel position, with English tongue splines almost overlapping for both the positions compared to Malayalam tongue postures for these positions.

Although cross-linguistic phonological influence is not evident in these figures, the deviations visible between conditions possibly indicate the reliance on auditory feedback, which is critical in altering the speech motor-control system as speech sounds are acoustically defined [18].

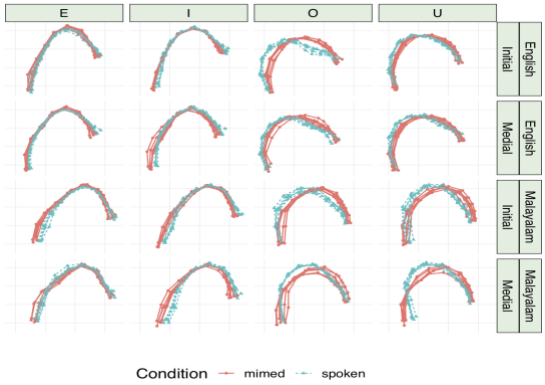

Figure 2: Plot displaying tongue shape variation across mimed and spoken speech conditions in both the languages (for every repetition of the words for each vowel)

3.2. Principal component analysis (PCA)

The obtained principal components scores are depicted in Figure 3 using R [17]. In the figure, each point represents a tongue shape for a single vowel repetition, in spoken speech condition (dots) and mimed speech condition (crosses). Figure 3 illustrates the tongue shapes that correspond to the extremes of the PC axes. The first principal component (PC1) accounted for 65.96% of tongue shape variation (tongue advancement differences), with negative values indicating front vowels and the positive values representing back vowels. The second principal component (PC2) accounted for 18.27% of tongue shape variation (tongue height differences), with positive values showing increased tongue height (elevation) and the negative values displaying reduced tongue height (depression).

Using PCA on shape data permits visualisation of the shape changes observed in the ultrasound data. Noticeable distinctions across vowels are apparent in Figure 3, based on speaking conditions (dots versus crosses) and languages. In comparison to the back vowels, the front vowels exhibit less articulatory variability between languages in both spoken and mimed conditions. On the other hand, back vowels for Malayalam in both the conditions are produced with increased tongue elevation, whereas for English, they are produced with reduced tongue height.

Many speakers of British English and American English use a high central rounded vowel (/ʉ:/) [22]. However, being a native Malayalam speaker, the participant in this study used a high back rounded vowel (/u:/) in Malayalam words, which is seen to be used in English words as well, suggesting the possibility of cross-linguistic phonological influence across languages. Furthermore, lack of auditory feedback might have contributed to the production of a somewhat high central rounded vowel in the mimed speech condition.

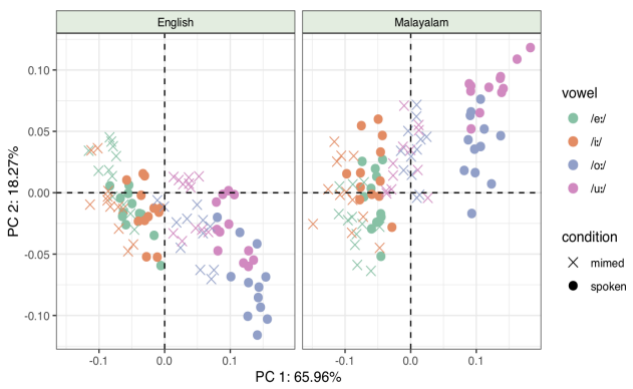

Figure 3: Combined PCA of shapes for four vowels across languages and conditions

A benefit of using GPA-normalised shape data is that differences in shape can be visualised directly, giving physical meaning to the axes in Figure 3.

Figure 4 presents deformation grids for mean, minimum and maximum scores of PC1 and PC2 for both the languages and conditions, illustrating how tongue shape changes throughout the study. From the Procrustes mean shape of the tongue towards the negative PC1 scores, the deformation grid is associated with more protrusion of the anterior part of the tongue and more gap at the posterior part of the tongue (possibly related to the front vowels /e:/ and /i:/). Towards the positive PC1 scores, the deformation grid corresponds to a retraction of the tongue body with less gap at the posterior part of the tongue (possibly related to the back vowels /u:/ and /o:/). In other words, PC1 indicated that the most shape variation in the dataset could be attributed to differences in tongue advancement.

Similarly, from the Procrustes mean shape of the tongue towards the negative PC2 scores, the deformation grid is associated with marginal lowering of the tongue or low tongue height (possibly relating to the low vowels /e:/ and /o:/). Towards the positive PC2 scores, the deformation grid corresponds to unnoticeably minimal raising of the tongue or high tongue height (possibly relating to the back vowels /u:/ and /i:/). Consequently, it could be concluded that PC2 indicated differences in tongue height (second most variation in the data set).

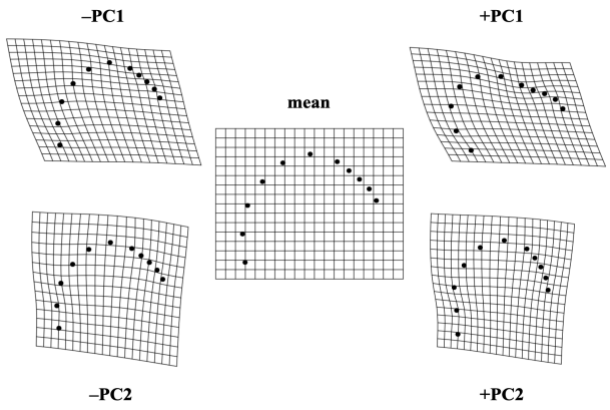

Figure 4: Deformation grids of mean, minimum and maximum scores for PC1 and PC2

Procrustes ANOVA revealed significant differences in the tongue position across conditions (spoken speech condition and mimed speech condition) (p=0.001) and across languages (Indian English and Malayalam) (p=0.001). Additionally, the study revealed a significant difference in tongue shape for different conditions across languages (p=0.018), suggesting that the effect of condition on tongue shape may be language-specific.

These results are indicative of within-speaker articulatory variation across languages and conditions. Although they do not directly indicate phonological transfer across languages, variation in tongue kinematics across independent variables may be suggestive of the influence of auditory feedback. Additionally, this study contributes to a better understanding of mimed speech in a bilingual speaker whose language variations are not well studied and provides information about within-speaker articulatory variation. This could contribute to future development of ultrasound-based SSI, aiming to give voice to bilingual alaryngeal speakers.

4. CONCLUSION

This paper reports the case study of a 25-year-old female Malayalam-English bilingual speaker. The study focused on estimating any phonological transfer of vowels across these languages and two speaking conditions, with mimed speech reflecting the alaryngeal speech. This paper presents a novel combination of state-of-the-art techniques for data collection (ultrasound tongue imaging) and data analysis (pose estimation using DeepLabCut™ and geometric morphometric analysis), with GMM being a new method for ultrasound derived data. The results revealed significant differences in the tongue kinematics across languages and conditions, without direct indication of phonological transfer. However, the understanding of within-speaker variation might help in transferring this knowledge in the development of an ultrasound-based Silent Speech Interface as a voice restoration method for bilingual laryngectomy patients.

7. REFERENCES

[1] Adams, D., Rohlf, F., Slice, D. 2004. Geometric morphometrics: Ten years of progress following the ‘revolution’. *Italian Journal of Zoology*, 71(1), 5-16.

[2] Articulate Instruments Ltd. 2022. Articulate Assistant Advanced (Version 219.08) [Computer software]. Edingurgh: Articulate Instruments Ltd. Retrieved April 01, 2022 from http://www.articulateinstruments.com/downloads/

[3] Cleland, J., Wrench, A., Scobbie, J., Semple. S. 2011. Comparing articulatory images: An MRI/Ultrasound tongue image database. *Proceedings of the 9th International Seminar on Speech Production*, 163-170.

[4] Crevier-Buchman, L., Gendrot, C., Denby, B., Pillot-Loiseau, C., Roussel, P., Colazo-Simon, A., Dreyfus, G. 2011. Articulatory strategies for lip and tongue movements in silent versus vocalized speech. *Proceedings of ICPhS XVII*, pp. 1-4.

[5] Csapó, T. G. 2020. Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal

tract. *Proceedings of INTERSPEECH 2020,* 2722-2726.
[6] Csapó, T. G., Tóth, L., Gosztolya, G., Marko, A. 2021. Speech synthesis from text and ultrasound tongue image-based articulatory input. *11th ISCA Speech Synthesis Workshop (SSW 11),* 26-28.
[7] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., Brumberg, J. S. 2010. Silent speech interfaces. *Speech Communication*, 52(2010), 270-287.
[8] Di Paolo, M., Yaeger-Dror, M., Wassink, A. 2010. Analyzing vowels. In: Di Paolo, M., Yaeger-Dror, M. (eds), *Sociophonetics: A student's guide.* Routledge,
[9] Dromey, C., Black, K. 2017. Effects of laryngeal activity on articulation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12), 2272-2280.
[10] Flege, J. E., Bohn O. 2021. The revised Speech Learning Model. In: Wayland, R. (ed.), *Speech Language Speech Learning, Theoretical and Empirical Progress*. Cambridge University Press.
[11] Grosjean, F. 1997. The bilingual individual. *Interpreting*, 2(1-2), 163-187.
[12] Gully, A. 2021. Quantifying vocal tract shape variation and its acoustic impact: A geometric morphometric approach. *Proceedings of INTERSPEECH 2021*, 3999-4003.
[13] Jarvis, S., Pavlenko, A. 2008. *Crosslinguistic influence in language and cognition.* Taylor & Francis.
[14] Krishnamurti, B. 2003. *The Dravidian Languages.* Cambridge University Press.
[15] Polly, P. 2012. *Procrustes, PCA, and 3D coordinates.* Geometric morphometrics module lecture notes. Indiana University: Department of Earth & Atmospheric Sciences.
[16] Porras, D., Sepúlveda-Sepúlveda, A., Csapó, T. G. 2019. DNN-based acoustic-to-articulatory inversion using ultrasound tongue imaging. *International Joint Conference on Neural Networks.*
[17] R Core Team. 2021. R (Version 4.1.2) [Computer software]. Vienna: Bell Laboratories. Retrieved November 1, 2022 from https://www.r-project.org/
[18] Simmonds, A. J., Wise, R. J. S., Leech, R. 2011. Two tongues, one brain: imaging bilingual speech production. *Frontiers in psychology*, 2(166), 1-13.
[19] Spreafico, L., Pucher, M., Matosova, A. 2018. UltraFit: A speaker-friendly headset for ultrasound recordings in speech science. *INTERSPEECH* 2018, 1517-1520.
[20] Teplansky, K., Tsang, B., Wang, J. 2019. Tongue and lip motion patterns in voiced, whispered, and silent vowel production. *Proceedings of ICPhS XIX,* 2831-2835.
[21] The Mathis Group, The Mathis Lab. 2022. DEEPLABCUT (Version 2.2.2) [Computer software]. Switzerland: The Mathis Group, Swiss Federal Institute of Technology Lausanne. Retrieved July 2022 from http://www.mackenziemathislab.org/deeplabcut
[22] Wells, J. 1986. *Accents of English 3: Beyond the British Isles.* Cambridge University Press.
[23] Wrench, A., Balch-Tomes, J. 2022. Beyond the edge: Markerless pose estimation of speech articulator from ultrasound and camera images using DeepLabCut. *Sensors*, 22(3), 1133-1162.