A Machine Learning Ensemble to Automatically Classify Tongue Ultrasound Contours Based on Displacement Measurements

Simon Gonzalez

The Australian National University

u1037706@anu.edu.au

Abstract

This paper introduces a Machine Learning Ensemble Model to automatically classify tongue ultrasound contours. It has been trained on displacement measurements in English Coronal Obstruents (/t/, /s/, /tʃ/, /ʃ/), from eight female native speakers of Australian English. The model has an accuracy of 97.6% for the Random Forests and 74.4% for the Decision Tree. The accuracy is higher for fricatives than for stops. Results also show that the most reliable area for classification is from 20% to the 40% of the contour length, which corresponds to the tongue area between the tongue front and the tongue body.

Index Terms: speech ultrasound, tongue contours, Machine Learning, English coronal obstruents, displacement

1. Introduction

The tongue is described as a highly mobile organ, whose shape can be deformed in extremely rapid succession through subtly different movements [1]. One of the main reasons for these changes in shape is to achieve articulatory targets. This complex motions and intricate postural adjustments make it difficult to classify tongue shapes during speech [2]. One of the technologies that has been used for this purpose is tongue ultrasound imaging (henceforth ToUS), which is used to identify contours from images as well as articulatory landmarks [3][4][5][6].

Due to the time-consuming aspect of tracing and extracting contours from ToUS images, several automatic and semi-automatic approaches have been developed for identification [7][8] and error detection [9]. Since the rise of Machine Learning (henceforth ML) in the last two decades, it has been one of the most widely implemented techniques to achieve automatic tasks in ToUS imaging. These techniques include Deep Learning [10], Support Vector Machines [2], Deep Neural Networks [11], Deep Belief Networks [12], and Convolutional Networks [13][14].

These implementations have made invaluable contributions to the tasks of contour identification and extraction, and they have increased the amount of data that can be processed in unprecedented ways. However, having larger datasets demands adequate computational approaches that can be developed for efficient processing and analysis of articulatory data [15]. This analysis area is one where ML has not been implemented, especially in ToUS imaging for linguistic purposes. One relevant question is whether ML models can offer more efficient ways to deal with the complexity of ultrasound data and its articulatory implications. One of the challenges in ToUS analysis is that identifying relevant articulatory sections of the tongue is not always a straight-forward process. For example, when investigating the articulation of alveolar segments in English, the front section of the tongue is a relevant area to observe, but there are other sections of the tongue, e.g. the tongue body, which also shows strong articulatory activity [16]. In this case, the question is not just what happens at the front section, but also at other areas that can be relevant for the gestural description of the segments, information that can be used to automatically classify contours to the right category.

Current analyses of ToUS include SSANOVAs [17][18][19], distances [20][21], velocities [22][23][24], heatmaps [16][25], and Principal Component Analysis [26][27]. These methodologies have been able to address relevant linguistic questions, providing strong analysis frameworks. However, much of the process of analysis still requires a great level of qualitative interpretation on behalf of the researcher. This is an accurate approach, yet not very efficient when large datasets are analysed. Another issue with these approaches is that they generally analyse data based on individual speaker patterns, then generalisations are made across all speakers in the data. What is therefore needed is approaches that can analyse all available data and give all factors (speaker-dependent and speaker-independent factors) the same opportunity to contribute to the classification and description of tongue contours. This is the reason why the implementation of ML can offer an opportunity to examine ToUS data in a more generalisable way.

2. Aim of Paper

The aim of this paper is to develop and implement an ML ensemble approach to analyse and automatically classify mid-sagittal tongue contours of English coronal obstruents, specifically palato-alveolars and alveolars. This is designed to capture spatio-temporal tongue gestures, the dynamics of the tongue contour over time, from previous vowel to maximum constriction. We employ a grid-lines approach, where tongue displacements between articulatory landmarks are captured by multiple lines and create descriptors to classify tongue trajectories across articulatory landmarks. We aim to train an ML ensemble (Random Forests and Decision Trees) using paired data to predict new data. We chose displacement measurements as the baseline for classification. These are calculated as tongue movements between two landmarks, previous vowel and maximum constriction for each of the four segments analysed, two alveolars and two palato-alveolars.

3. Methodology

In this section, we describe the methodology used for the analysis. We first present information on the data (equipment, collection, processing), then on the acoustic and articulatory landmarks, and finally on the measurements and the ML ensemble developed.

3.1. Equipment and Data Formats

We recorded mid-sagittal images of the tongue contours using a portable Sonosite 180 Plus ultrasound machine with a C11/7-4 MHz 11-mm broadband curved array transducer. The transducer was fixed to a stabilisation helmet from Articulate Instruments [24]. The acoustic signal was recorded using a Shure KSM137. The microphone was connected to an M-audio DMP3 preamplifier, and the audio output from the amplifier was sent to the audio input of a camcorder. A Sony DCRTRV103 digital camcorder in NTSC format (30 fps) did a simultaneous recording of both ultrasound and acoustic signals. The data was then downloaded to a computer using the Adobe Premiere Elements software (www.adobe.com). The audio signal was digitised at a sampling frequency of 48 KHz with 16-bit quantisation. Each video was then saved as a sequence of still images in JPG format (29.97 fps). We also saved the audio signal as WAV files. The audio recordings were saved at a sampling frequency of 44.1 KHz with 16-bit quantisation.

3.2. Speakers and Stimuli

The participants were eight adult females, all native speakers of Australian English with no reported hearing or articulatory impairment. Target segments were two alveolars (/t/, /s/) and two palato-alveolars (/tʃ/, /ʃ/). The target segments were elicited in monosyllabic words in onset position: *tack*, *sack*, *Chack*, and *shack*. The vowel /æ/ (the lowest front vowel in Australian English, see [28]) was chosen as the common context vowel because of its articulatory properties, being the most suitable context vowel for an ultrasound study of coronal segments. The target words were placed in a carrier sentence. This isolates relevant articulatory parameters for the segments analysed, in terms of tongue advancement and tongue height. The carrier sentence was *Please, utter X publically*, where *X* represents the target word. This controls the previous vowel /ə/. In Australian English, a non-rhotic variety, the word utter is pronounced [ˈʁrə]. Thus, the target segment occurred after the mid central vowel /ə/ and before the vowel /æ/.

3.3. Data Segmentation

The still images in JPG format were used for tracking the tongue surface lines (contours) using the EdgeTrak software [29]. The audio signal was used for the acoustic analysis of the data. Contours were saved as con files, which is the default format in Edgetrak, and they are coded as numeric Cartesian values for each point of the contour line ([*x,y*] coordinates), and totalling 100 points per contour. For each of the four segments, we selected five repetitions. The format of the data follows the structure as in [16] with a csv file containing the following columns: Speaker, Segment, Repetition, and Frame.

3.4. Landmark Identification

We identified two landmarks: previous vowel (PV) and maximum constriction (MC). The motivation is to examine which parts of the tongue are the ones that activate and make the main articulatory gestures to achieve the constriction. The identification of these landmarks was based on two approaches. PVs were labelled by the authors from spectrogram and waveform display in Praat [30] (further description in Section *3.4.1*) and MCs based on articulatory landmarks using the Ultrasound and Visualisation App (UVA) [16], which carries out both static and dynamic analyses from the numeric output files exported by EdgeTrack. The app has a landmark identification task, in which users visualise contours in context (frames before and after), which helps the coding of landmarks. We used this functionality to assign the corresponding Maximum Constriction (MC) contours for each sequence. Below, we present the acoustic and gestural cues observed when identifying the landmarks.

3.4.1. Acoustic identification of Previous Vowel /ə/

The vowel /ə/ was segmented on the basis of clear vocalic peak pulses in the acoustic signal. The onset was placed at zero crossing point preceding the peak of the first clear vocalic pitch pulse (boundary (a) in Figure 1) after the offset of /t/. The offset of /ə/ was placed at zero crossing point following the peak of the last clear vocalic pitch pulse (boundary (b) in Figure 1) before the onset of the following target consonant. The tongue contour corresponding to the time mid-point of the vowel duration was assigned as the PV for each sequence.

3.4.2. Articulatory identification of Maximum Constriction

MC contours were identified by holistically examining ToUS contours one by one in each sequence. The MC frame was the frame previous to the first downward movement of the tongue from consonant MC to the following vowel /æ/. This was the contour in which the maximum raising/advancement was observed.

3.5. Measurements

The analysis baseline was done on a gridlines approach, which is implemented within UVA and also in other work (c.f. [23][31][32]). With this type of approach, we can carry out ToUS analysis in both advancement and height dimensions, across all the extent of the contours available. The process is represented in the Figure 1. The first step is to create a point of origin from the available contours. Then gridlines are projected from this point to capture sections relevant for articulation. Then intersections are calculated between the gridlines and the contours, which gives the common area of analysis encompassing all data that can be compared across all available contours. In our analysis, we selected 20 gridlines to capture the necessary articulatory activity for both upward and downward movement of the tongue. This is done for each speaker individually.

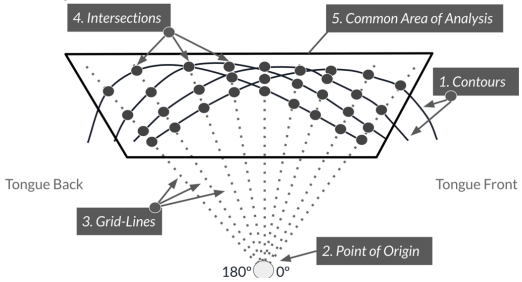

Figure 1: *Gridlines and intersections calculations.*

Since the gridlines are fixed for all contours in a given speaker, then any observed difference can be reliably interpreted as pertaining to articulatory differences.

We then selected the dynamic measurements and worked on the displacement calculations in mm as measurement units. Here, motion is defined as the change of a tongue section within the mouth in respect to time, more specifically, how fast the tongue is moving. Displacement, which is the length of the path travelled by the tongue section from one landmark to another, based on a specific gridline. It is important to point out that this

displacement is a relative measure from one tongue contour to another, i.e. the displacement calculated can only measure the relative movement from point A to point B in a given gridline.

The process is represented in Figure 2, and it first calculates the distances from the origin point to all contour intersections across all gridlines. For each gridline, new distances are calculated from the first landmark (PV) to the second one (MC). The displacement can therefore be positive (e.g. first gridline), zero (e.g. second gridline), or negative (e.g. third gridline).

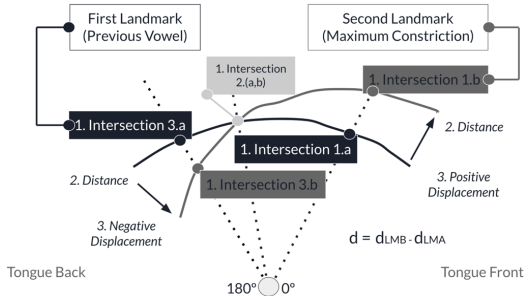

Figure 2: *Displacement calculation between landmarks.*

3.6. Analysis

For the analysis, we used a supervised ML algorithm ensemble in R [34]. An important characteristic of supervised algorithms is that they are trained to identify the category to which an object belongs based on the characteristic of the object itself [33]. We then divided the data into training (70%) and test (30%) subsets.

This approach was chosen to address two main goals. The first one was to identify which of the sections across the tongue contour surfaces were the ones relevant to describe the distinctive gestural behaviours. In order to keep measurement consistency, we grouped the 20 gridlines into five percentage incremental sections from right to left: gridlines 1 to 4 (0%-20%), 5-8 (20%-40%), 9-12 (40%-60%), 13-16 (60%-80%), and 17-20 (80%-100%). Lower gridlines (and percentages) would capture more advanced articulatory gestures at the front section of the tongue, and higher gridlines (and percentages) would capture more retracted gestures at the back sections of the tongue.

To address this research goal, to identify and distinguish relevant articulatory sections, we used a Random Forest algorithm (RF), which combines the output of multiple decision trees to reach a single result. RF is employed for classification, regression, and other activity based on the construction of a multitude of decision trees during the training and generates the class that represents the overall prediction of the single trees [35].

The second goal was to identify the amount of displacement that was relevant when classifying tongue contours. For this, we ran a Decision tree algorithm, which creates a model to predict the value of an outcome variable by learning decision rules from data features [36]. We used this to examine which tongue contour percentages are used to distinguish between the type of segments, or categories (alveolars vs palato-alveolars) and at what stage in the classification process.

4. Results

For both ML algorithms, we ran models where the category to predict was the segment individual type. We added the five percentage sections and the speaker as predictors. The reason to keep speaker as a predictor was to check whether individual speaker articulatory patterns or individual displacements were relevant. The raw displacement values for all speakers are shown in Figure 3 below. These show that most of the distinctive articulatory patterns happen in the first 60%-70% percent of contours. For the 80%, the differences are smaller than the other percentages, and in the 100% all displacements are negative. These indicate that all segments lower the back of the tongue at 100% and the 80% functions as a pivot or anchor in the tongue movement (see [37] for similar results in vowels).

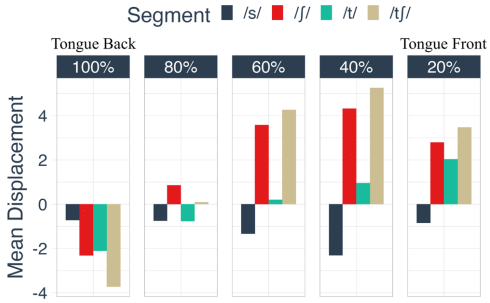

Figure 3: *Raw displacement across segments.*

4.1. Random Forest Results

The RF classification task had an accuracy of 97.6%. The only four wrong classifications came from four /ʃ/ tokens which were classified as /tʃ/. The variable importance results from the RF model are shown in Figure 4. These show that most of the distinctions across all segments take place at the 40%, which corresponds to the area of the tongue between the tongue front (highly active in alveolars) and the tongue body (highly active in palato-alveolars).

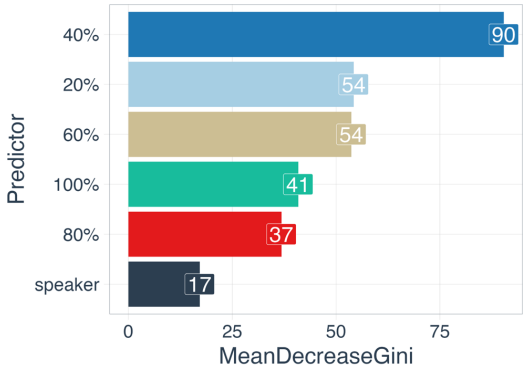

Figure 4: *Variable Importance from the Random Forest Model.*

4.2. Decision Tree Results

Figure 5 shows the Decision Tree classification of the test data from the training data. The figure shows the parameters that are most relevant for the classification of segments. It shows that the 40% is the most relevant parameter, since it is used several times and at top splits to distinguish between categories. This is similar to the RF results, but it adds further details by specifying at what point the distinctions are activated. Results also show the displacement thresholds at each of the nodes. It shows that at Node 1, there is a clear distinction between alveolars and palato-alveolars at the 40% and showing that displacements below 1.2mm are a cutoff point, with alveolars having lower displacements and palato-alveolar higher ones. Nodes 2 and 3 separate /s/ from /t/, with /t/ having higher displacements than /s/ at 40% and 80%. Nodes 4 and 5 separate /tʃ/ from /ʃ/, at both 80% and 100%. At the 80% section, /ʃ/ has lower displacements

that /tʃ/, with the cutoff being 3.1mm. Examining the final classifications of the tree, results show that when distinguishing between manner of articulation at the same place of articulation, stops display higher displacements than the fricative counterparts.

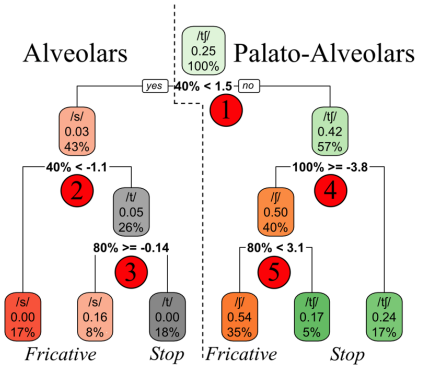

Figure 5: *Decision Tree from the trained model.*

The Decision Tree model had a predictive accuracy of 74.4%. The prediction accuracy is shown on Figure 6, which combines a correlation heatmap plot and a dendrogram. Segment /s/ was the segment with the highest accuracy (90%), with 6% classified as /ʃ/ and 4% as /t/. Segment /ʃ/ had the second highest accuracy (80%), with 12% classified as /tʃ/ and 7% as /s/. /t/ had the lowest accuracy (59%) with 33% classified as /ʃ/ and 9% as /s/. Finally, /tʃ/ had slightly higher accuracy (67%), with 33% classified as /ʃ/. These general results show that fricatives have higher accuracy than stops, 85% and 63%, respectively.

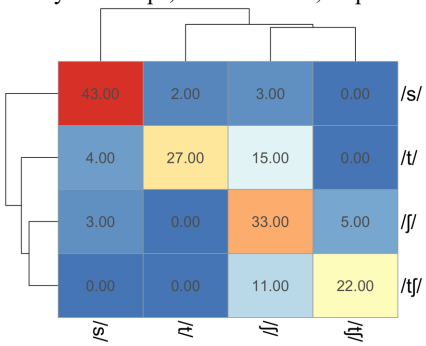

Figure 6: *Accuracy and errors in the Decision Tree output.*

5. Discussion

Random Forest results show the order of importance of the sections that are most relevant to distinguish between phonological categories. Their ranking closely reflects the areas where there was more gestural activity (20%-40%-60%). But the RF shows that the 40% is the most relevant, then the 20% and 60%. These sections are the ones that are driving the main gestural differentiations, whereas the 80% functions as the anchor point for all the segments. The 100% shows that all the segments have a negative displacement, with /tʃ/ having the largest and /s/ the smallest displacement, and /t, ʃ/ with mid-values. In this case, we have evidence of distinctive patterns among segments, but also common patterns observed.

The Decision Tree analysis adds more crucial information on the stage at which these parameters are activated and the approximate length of displacement that drives the distinctions. These show again that the 40% is the most crucial area for gestural behaviour, but the 80% also triggers further distinctions. Results from this model also tells us the amount of displacement, and what direction, that is necessary to identify segment types. Combining all results, it shows that there is a cutoff at around 1.5mm, with alveolars being lower and palato-alveolars larger. In turn, fricatives have lower displacements than the stop counterparts. This higher displacement for stops can be understood as an articulatory requirement. Since these segments achieve complete constriction, it makes sense that they would require larger displacements of the tongue to reach its articulatory target.

In terms of the error for both models, stops show higher rates than fricatives. In addition, between /t/ and /tʃ/, the alveolar segment is the one that triggers the most errors. All errors are not random. For the palato-alveolars, in most of the errors they are confused by their counterpart with different manner of articulation. In the case of the alveolars, they are mainly confused with /ʃ/, and not the manner of articulation counterpart.

6. Limitations and Future Work

There are three main limitations identified in this paper. The first one is that the displacement measurements have been extracted for only female speakers. Further analysis would include male speakers to account for vocal tract size differences. Another limitation is that it has only been tested on voiceless coronal obstruents in English. We aim to test this with other segments, e.g. velars and palatals, as well as voiced vs voiceless classifications. Another limitation is that it was only trained on four consonants in a single vowel context, and it is yet to be shown whether it would perform well on a more diverse set of data. The final limitation is that the context for each token was fully controlled. Further investigation would reveal classification accuracy in cases where not all the tokens share the same context, especially in free speech.

7. Conclusions

In this paper, we have developed a Machine Learning Ensemble that analyses and classifies ToUS contours based on displacement measurements. Its high accuracy makes it a robust model to be implemented on new input data. Since it was trained on eight different speakers, it can be used to predict new data from different speakers. By analysing the variable of importance from the RF and the predictors at each node split in the Decision Tree, the approach allows having a clear understanding on what are the driving parameters when dealing with paired comparisons (e.g. alveolrars vs palato-alveolars (Place of Articulation), or stops vs fricatives (Manner of Articulation)). This model can be used to classify new data, given that it follows a similar format. The code and the model can be accessed through the following GitHub repository: https://github.com/simongonzalez/TongueUltrasoundAndML.

The methodological advancements presented here are relevant to the field of ToUS analysis in which identification and classification of articulatory landmarks can be automated, and thus, maximising the examination of the phonological implications of such differences.

8. Acknowledgements

I want to thank the anonymous reviewers for their comments and suggestions, which have improved this paper in many ways. The errors that remain are entirely my own responsibility.

9. References

[1] Wen, S., "Automatic Tongue Contour Segmentation using Deep Learning", Thesis, University of Ottawa, 2018.

[2] Tang, L., Hamarneh, G. and Bressmann, T., "A Machine Learning Approach to Tongue Motion Analysis in 2D Ultrasound Image Sequences", Machine Learning in Medical Imaging, 151-158, 2011.

[3] Stone, M., "A guide to analysing tongue motion from ultrasound images", Clinical Linguistics & Phonetics, 19(6-7):455-501, 2005.

[4] Gick, B., Campbell, F. and Oh, S., "A cross-linguistic study of articulatory timing in liquids", The Journal of the Acoustical Society of America, 110(5), 2001.

[5] Mielke, J., "An ultrasound study of Canadian French rhotic vowels with polar smoothing spline comparisons", The Journal of the Acoustical Society of America, 137(5), 2015.

[6] Roxburgh, Z., Cleland, J., Scobbie, J.M. and Wood, S.E., "Quantifying changes in ultrasound tongue-shape pre- and post-intervention in speakers with submucous cleft palate: an illustrative case study", Clinical Linguistics & Phonetics, 36:2-3, 146-164, 2022.

[7] Karimi, E., Ménard, L. and Laporte, C., "Fully-automated tongue detection in ultrasound images", Computers in biology and medicine, 111, 103335, 2019.

[8] Roon, K.D., Chen, W.-R., Iwasaki, R., Kang, J., Kim, B., Shejaeya, G., Tiede, M.K. and Whalen, D.H., "Comparison of auto-contouring and hand-contouring of ultrasound images of the tongue surface", Clinical Linguistics & Phonetics, 2022.

[9] Csapó, T.G. and Lulich, S.M., "Error analysis of extracted tongue contours from 2d ultrasound images", INTERSPEECH 2015, September 6-10, Dresden, Germany, 2015.

[10] Mozaffari, M.H. and Lee, W., "Deep Learning for Automatic Tracking of Tongue Surface in Real-time Ultrasound Videos, Landmarks instead of Contours", IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2785-2792, 2020.

[11] Jaumard-Hakoun, A., Xu, K., Roussel-Ragot, P., Dreyfus, G. and Denby, B., "Tongue contour extraction from ultrasound images based on deep neural network", ArXiv, abs/1605.05912, 2016.

[12] Fasel, I.R. and Berry, J., "Deep Belief Networks for Real-Time Extraction of Tongue Contours from Ultrasound During Speech", 20th International Conference on Pattern Recognition, 1493-1496, 2010.

[13] Zhu, J., Styler, W. and Calloway, I.C., "A CNN-based tool for automatic tongue contour tracking in ultrasound images", ArXiv, abs/1907.10210, 2019.

[14] Xu, K., Csap'o, T.G. and Feng, M., "Convolutional Neural Network-Based Age Estimation Using B-Mode Ultrasound Tongue Image", ArXiv, abs/2101.11245, 2021.

[15] Barros, F., Valente, A.R., Albuquerque, L., Silva, S.S., Teixeira, A.J. and Oliveira, C., "Contributions to a Quantitative Unsupervised Processing and Analysis of Tongue in Ultrasound Images", ICIAR 2020: Image Analysis and Recognition, 170-181, 2020.

[16] Gonzalez, S., "Gridlines approach for dynamic analysis in speech ultrasound data: A multimodal app", Journal of the Association for Laboratory Phonology 12(1):16,1-28, 2021.

[17] Davidson, L., "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance", Journal of the Acoustical Society of America, 120(1):407-15, 2006.

[18] Gu, C., "Smoothing spline ANOVA models", New York, NY: Springer, 2002.

[19] Chiu, C., Wei, P.C., Noguchi, M. and Yamane, N., "Sibilant Fricative Merging in Taiwan Mandarin: An Investigation of Tongue Postures using Ultrasound Imaging", Lang Speech. 63(4):877-897, 2020.

[20] Zharkova, N., Hewlett, N. and Hardcastle, W.J., "Coarticulation as an Indicator of Speech Motor Control Development in Children: An Ultrasound Study", Motor Control, 15:118-140, 2011.

[21] Barbier, G., Perrier, P., Payan, Y., Tiede, M.K., Gerber, S., Perkell, J.S. and Ménard, L., "What anticipatory coarticulation in children tells us about speech motor control maturity", PLOS ONE 15, 2020.

[22] Ostry, D.J. Keller, E. and Parush, A., "Similarities in the control of the speech articulators and the limbs: Kinematics of tongue dorsum movement in speech", Journal of Experimental Psychology: Human Perception and Performance. 9, 622, 1983.

[23] Strycharczuk, P. and Scobbie, J. M., "Velocity measures in ultrasound data. Gestural timing of post-vocalic /l/ in English", In Proceedings of the 18th International Congress on Phonetic Sciences, 10 August - 14 August, Glasgow, U.K, 2015.

[24] Wrench, A. and Scobbie, J.M., "Very high frame rate ultrasound tongue imaging", 2011.

[25] Carignan, C., "Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels", The Journal of the Acoustical Society of America, 143(5), 2588, 2018.

[26] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P. and Stone, M., "Eigentongue feature extraction for an ultrasound-based silent speech interface", in IEEE International Conference on Acoustics, Speech and Signal Processing, 1245-1248. Honolulu, HI: Cascadilla Press, 2007.

[27] Carignan, C. and Mielke, J., "Extracting articulatory signals from lingual ultrasound video using principal component analysis", MS, 2014.

[28] Cox, F., "Vowel transcription systems: An Australian perspective", International Journal of Speech-Language Pathology, 10:327-333, 2008.

[29] Li, M., Kambhamettu, C. and Stone, M., "Automatic contour tracking in ultrasound images", Clinical Linguistics & Phonetics, 19(6-7):545-554, 2005.

[30] Boersma, P. & Weenink, D., "Praat: doing phonetics by computer [Computer program]", Version 5.4.07, retrieved 22 March 2015 from http://www.praat.org/, 2005.

[31] Wrench, A., "Articulate Assistant Advanced User Guide", Edinburgh: Articulate Instruments Ltd., 2012.

[32] Liker, M., Zorić, A. V., Zharkova, N. and Gibbon, F. E., "Ultrasound Analysis of Postalveolar and Palatal Affricates in Croatian: A Case of Neutralisation", in S. Calhoun, P. Escudero, M. Tabain and P. Warren [Eds], Proceedings of the 19th International Congress of Phonetic Sciences, 3666-3670, Melbourne, Australia: Australasian Speech Science and Technology Association Inc., 2019.

[33] Micucci, M. and Iula, A., "Recent Advances in Machine Learning Applied to Ultrasound Imaging", Electronics 11(11):1800, 2022.

[34] R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2021.

[35] Breiman, L., "Random forests", Machine Learning, 45:5-32, 2001.

[36] Gareth, J., Witten, D., Hastie, T. and Tibshirani, R., "An introduction to statistical learning: with applications in R", New York:Springer, 2013.

[37] Kim, B., Tiede, M.K. and Whalen, D.H., "Evidence for pivots in tongue movement for dipthongs", in S. Calhoun, P. Escudero, M. Tabain and P. Warren [Eds], Proceedings of the 19th International Congress of Phonetic Sciences, 3666-3670, Melbourne, Australia: Australasian Speech Science and Technology Association Inc., 2019.