Interaction of phonetic materials and simulated probe position on the mean square error metric of tongue ultrasound probe alignment – a methodological case study

Pertti Palo[1], Steven M. Lulich[2], Daniel Aalto[1]

[1] *University of Alberta*
[2] *Indiana University*

Abstract

In tongue ultrasound imaging shifts in probe placement can cause problems in data interpretation if they go undetected. We analyse a promising metric – the mean squared error (MSE) of the mean ultrasound images – as a metric of probe stability. The metric's performance is evaluated against systematically varied speech materials (fronted articulation versus backed articulation) and probe displacement. The speech materials consist of 54 different /$C_1VC_2VC_3V$/ utterances in random order produced by one native speaker of Finnish and recorded with a Micro ultrasound setup using Articulate Assistant Advanced. In the fronted condition the vowel is /i/ and consonants are varied systematically among /n,s,t/. In the backed condition the vowel is /o/ and the consonants are varied among /h,k,ŋ/. The probe displacement is both simulated and produced intentionally in the real world. For the latter the 54 utterances were repeated in a second block in a different random order. The differences between the results of the two displacement methods indicate that this dual approach merits further study. The results also indicate that varying speech materials may overshadow probe displacement which leads to a tentative recommendation of comparing like with like in speech materials when using MSE to detect probe movement.

1. Introduction

When recording any kind of speech data, acquiring data that is comparable with recordings from other speakers let alone the same speaker, is of utmost importance. Otherwise, it is difficult to make reliable inferences based on the data. To state this more formally, we have two desirable qualities: intra-speaker comparability and inter-speaker comparability. In recording articulatory data, the first quality is satisfied

by recording the same anatomical area throughout the session (or keeping the sensors or pellets in place for point tracking methods). Satisfying the second quality involves tackling the problem of anatomical normalisation. This can be much more difficult as in general it requires a method of anatomical normalisation.

When using Ultrasound Tongue Imaging (UTI) to acquire articulatory speech data from a given speaker, the intra-speaker comparability condition stipulates that we want to capture the same region of the tongue for all recordings. This means that we need some method of first setting the system up to record the correct area and of consequently assuring that this remains true across the recording session. Satisfying the second quality on a general level is not easy with tongue ultrasound as it does not readily provide data on dimensions of the vocal apparatus beyond the tongue. However, many methods circumvent this problem by quantifying the geometrical qualities of articulatory positions instead (Ménard et al., 2012; Stolar & Gick, 2013; Zharkova et al., 2015; Dawson et al., 2016). However, rotating out of the mid-sagittal plane can cause problems with even these methods as the central groove of the tongue may appear as an extra feature in the images that could lead to false conclusions about the actual articulatory position. This makes it important to be able to minimise or track probe movement during recording and to check for it in already recorded data.

In satisfying the intra-speaker comparability condition, there have been various efforts to either reduce variability in collecting tongue ultrasound data or to reduce variability by postprocessing. The first kind have mainly taken the form of stabilising the ultrasound probe in different ways, while the second kind have consisted of tracking the probe's position in relation to the speakers head, and using a biteplate and/or anatomical measurements to orient the images and resulting extracted tongue surfaces.

The most basic approach to probe stabilisation is for either an experimenter or the participant to hold it in their hand. This has the obvious drawbacks that the inherent unsteady nature of this method potentially causes frequent significant probe misalignment, and that documenting the imaging position is challenging. It is, however, often the only method that will work with very young children as they may not be able to tolerate wearing a relatively heavy headset for the duration of the experiment (Zharkova et al., 2017).

It can also be beneficial in a clinical treatment setting because it gives more flexibility in

what is actually imaged (Adler et al., 2007). The reliability of data from hand stabilisation can be improved with stability measures like laser pointers attached to the probe and the speaker's head (Gick, 2002). And if employing analysis methods that do not rely on an unmoving frame of reference such as those from for example Dawson et al. (2016) and others mentioned above the data will be readily analysable as long as the imaging plane of the probe has been relatively stable.

The probe can also be stabilised by attaching it to something very immobile and this can be combined either with asking the speaker to hold still with their chin resting on the probe or by immobilising the speaker's head as well (Stone & Davis, 1995). While this solution is simple, it is hardly portable. These days, stabilising probe position is perhaps most usually done with helmets and headsets which can be either rigid (Articulate Instruments Ltd, 2008; Spreafico et al., 2018) or elastic (Derrick et al., 2018). These have the advantage of being fairly portable, but still allow the speaker's head to move in relation to the probe.

Another way of tackling the problem is to relate the image data to anatomic markers. This can take the form of locating anatomical features in the images after recording and can involve using special calibration recordings to capture them. The former utilises usually the tendon of the genioglossus, but can also include the palate (Stone, 2005; Wrench & Balch-Tomes, 2022; Aalto et al., 2024). Perhaps the most used calibration recording is obtaining a tongue depressor or biteplate trace (Stone & Davis, 1995). These give a common reference line – the occlusal plane of the speaker – that can be used to rotate the images to the same orientation across sessions and speakers. This comes with the obvious caveat that if the participant has malocclusion, the occlusal plane is not well-defined – which might be a potential problem even when there is no malocclusion because the teeth do **not** generally form a perfect plane. Furthermore, a biteplate does not guarantee that probe positioning stays the same in a given session nor that the probe is correctly positioned in neither translational nor rotational sense.

To guarantee or at least record the probe position in relation to the head, we need to employ some form of tracking. This can be done with optical point tracking methods (Whalen et al., 2005), tracking aids attached to the speakers head and the ultrasound probe combined with image processing (Mielke et al., 2005), accelerometers (Hueber, 2013), or electromagnetic articulography (Kirkham et al., 2023). All of these methods

require extra equipment and potentially heavy post-processing to detect head/probe movement and/or to make corrections.

An alternative light-weight method was first developed by Csapó & Xu (2020); Csapó et al. (2020). It involves using Mean Squared Error (MSE) of mean ultrasound images as a metric to identify potential probe movement. The method is computationally relatively light and does not require any extra equipment and does not require any special setup in the data recording. On top of this the MSE analysis results are easy to interpret as they provide a holistic view to the data from a given recording session in one glance.

The lack of special requirements on the data makes the MSE method very useful as it can be applied to any existing tongue ultrasound data. Given the ease of interpreting the results, it can also be used for quick assessment of data quality right after recording, allowing for prompt re-recording in case misalignment is detected.

In this study we provide a first step towards evaluating the MSE method's performance by using controlled data as well as simulated data. In the data we varied the speech materials to assess their effect on the metric, and we also intentionally changed the probe position during the recording session. Furthermore, we simulated probe rotation via a bootstrap method.

2. Materials and Methods

2.1. Audio and ultrasound recordings

Synchronised audio and ultrasound were recorded with Articulate Assistant Advanced (Articulate Instruments Ltd, 2024) using a Micro ultrasound system from Articulate Instruments Ltd. Audio was sampled at 22.050 kHz and ultrasound was acquired at 63.5 fps with Field of View of $\approx 102.7°$, 64 scanlines and depth of 120 mm / 898 pixels.

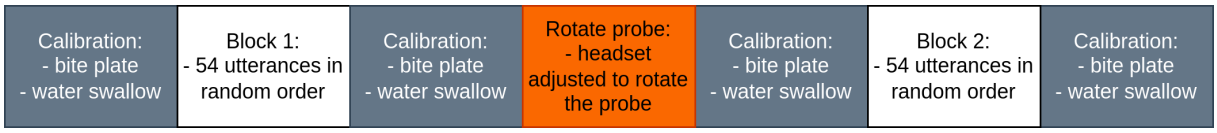

Figure 1: Block design of the experiment.

Figure 1 illustrates the over all design of the experiment. The speech materials consisted of /$C_1VC_2VC_3V$/ utterances produced by one male native speaker of Finnish. The vowels and consonants were chosen to elicit fronted and backed articulations.

Overall we chose to keep the vowel constant within each condition and vary the consonant by choosing a nasal, fricative and a plosive that have target constrictions corresponding to the condition. In the fronted condition the vowel was the fronted high corner vowel /i/ and consonants (C_1, C_2, andC_3) were varied among /n,s,t/ which all have frontal articulatory targets. In the backed condition the vowel was the back round vowel /o/ and the consonants were varied among /h,k,ŋ/. While the latter two have velar targets, in Finnish, /h/ has allophonic variation and can be produced as a pharyngeal fricative (Suomi et al., 2008). The participant was instructed to aim for the pharyngeal allophone of /h/. In each condition, the vowel remained constant, and the consonant was systematically varied to produce all combinations (including e.g., in the fronted condition /ninini/ and /tinisi/ among others). Combining the vowels and consonants resulted in $2 * 3 * 3 * 3 = 54$ unique /$C_1VC_2VC_3V$/ utterances.

The utterances were repeated in two independently randomised blocks, each consisting of the full set of 54 utterances. Between the blocks the headset was adjusted with the aim to rotate the probe by approximately 5-10°. The first block was used as the basis for the simulated rotation described below. Both blocks contained some speech errors including false starts, disfluencies, and repeated productions of the target utterance. These were not removed from the data but rather included to see how they affected the results. Calibration tasks were recorded before and after each block. These included a bite plate recording with a wooden tongue suppressor and a water swallow.

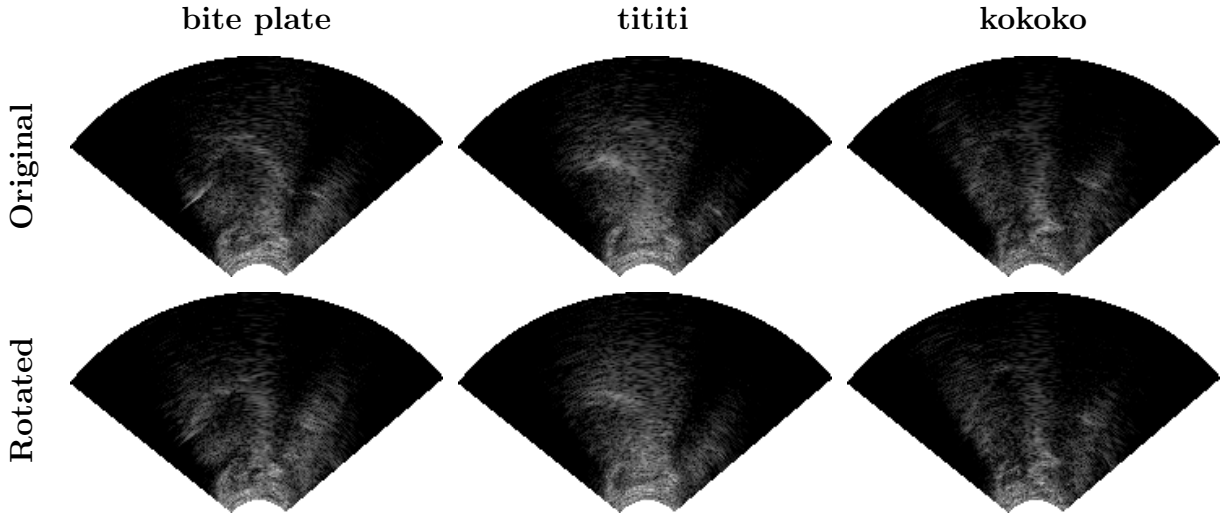

Figure 2: Representative tongue profiles.

Figure 2 shows tongue profiles from 6 different conditions. The images in the first row

are from the first block and images in the second row are from the second block. In all of the images in this study the tongue tip is to the left and pharynx is on the right. The first column shows frames captured with the tongue depressor, while the tongue depressor does not appear to change position between these two images, the hyoid bone shadow does do so. See discussion for a further comment on this. The other two columns give examples of fronted – here a /tititi/ utterance – and backed – here a /kokoko/ utterance – productions. These frames are from the middle vowel of each utterance.

2.2. Probe alignment measurement with MSE

The probe alignment measurement with Mean Squared Error (MSE) is based on raw ultrasound frames. Raw frames are the uninterpolated or probe return frames illustrated in Figure 3. In our data each raw ultrasound frame has 64×898 pixels.

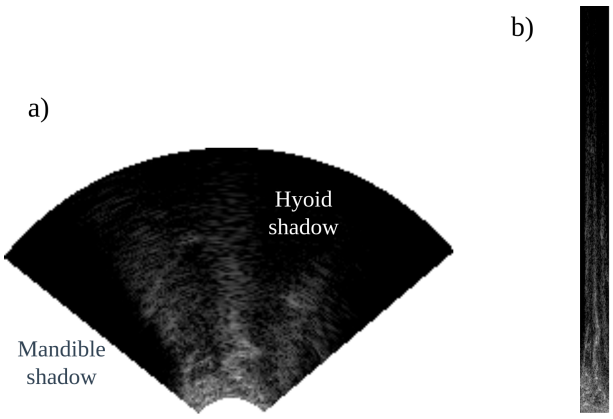

Figure 3: a) Interpolated (human-readable) ultrasound frame with mandible and hyoid shadows indicated, and b) the corresponding raw frame (probe return data).

In the following we will treat each raw ultrasound frame in a recording as a vector of l pixels indexed with k, use j to index the m consecutive images in a recording, and i to index the n consecutive recordings. So, the k^{th} individual pixel in the j^{th} frame of the i^{th} recording is $im(i, j, k)$. To calculate the MSE we will first calculate the average frame $\overline{im}$ of each recording by averaging each individual pixel over the recording:

$$\overline{im(i,k)} = \frac{1}{m} \sum_{j=1}^{m} im(i,j,k) \tag{1}$$

Figure 4 shows the mean images for the same recordings that were sampled in Figure 2. Interestingly, in Figure 4 rather than rotated, the tongue contour seems to be slightly lower in all of the images for the second block than for the first block.

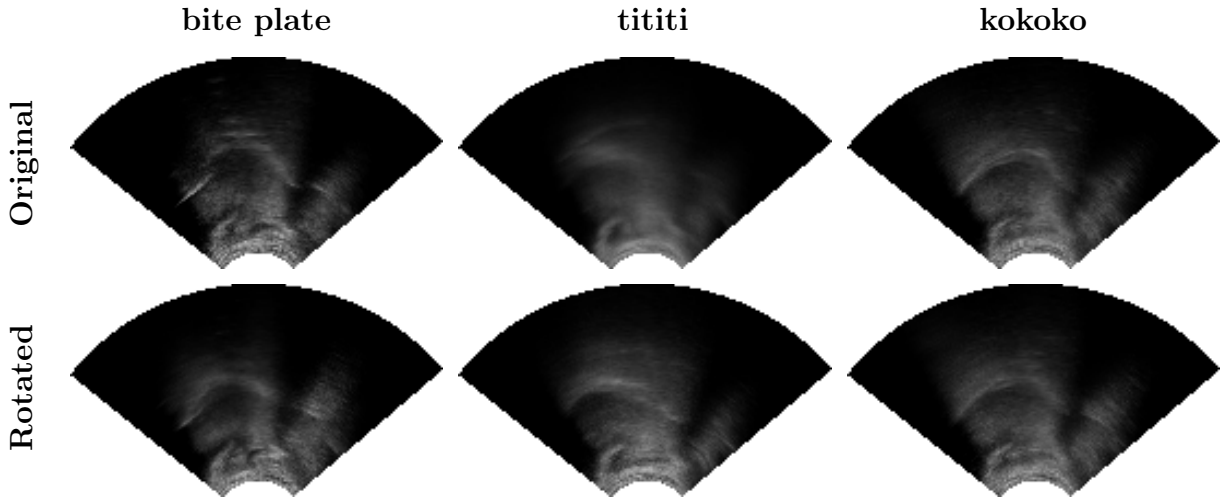

Figure 4: Examples of mean tongue profiles.

After this we calculate the MSE between each pair of average images (s, t) as follows:

$$MSE(s,t) = \frac{1}{l}\sum_{k=1}^{l}\left(\overline{im(s,k)} - \overline{im(t,k)}\right)^2. \quad (2)$$

The results of the MSE calculation can be represented as a $n \times n$ distance matrix where each individual element gives a measure of likeness between the recordings corresponding to the row and the column of the element.

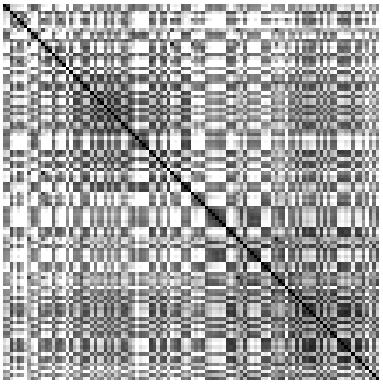

Figure 5: MSE matrix calculated on our whole (unsorted) data set. The dark line on the diagonal is the result of comparing each recording also to itself.

Figure 5 illustrates a distance matrix produced with MSE on the whole (unsorted) data after removing the calibration recordings (biteplate and water swallow). In it both rows and columns correspond to the individual recordings and each pixel is produced by calculating the MSE between the mean images of the recordings corresponding to the pixel's row and column. The closer to white a pixel is the more different the recordings are while darker pixels mark more similar recordings. The recordings appear in order of recording. The pixel in row 1, column 1 is comparison of the first recording to itself and

the dark pixels on the diagonal are comparisons of recordings to themselves. Importantly the left top quarter contains within-block comparisons of recordings in block one and right bottom quarter is comparisons within block two. Bottom left, and top right quarters are mirrored as they both are cross-block comparisons between recordings of blocks one and two.

2.3. Simulating probe rotation

We also simulated rotating the probe under a speaker's chin by producing artificially rotated data from the first (unadjusted) block of recordings. First we selected a subsector of each mean image to use in the analysis. This is done by excluding scanlines from either the front or the back of the image producing a continuous sector which is part of the original data as shown in Figure 6. We then compare two differently selected sectors over all the baseline recordings in our data. This simulates the probe having been turned by the amount that the sectors differ within the data.

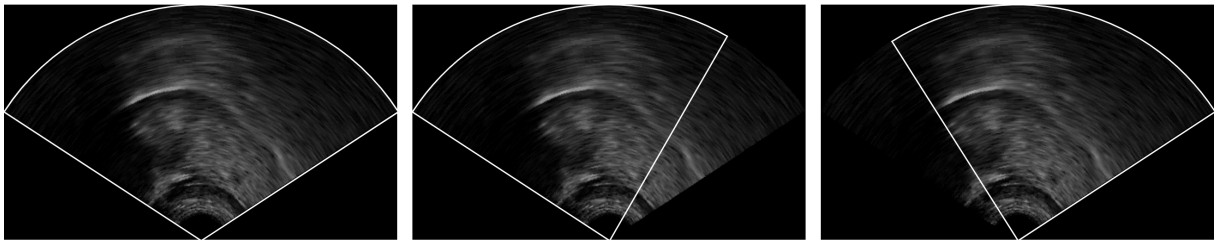

Figure 6: Simulating probe rotation by selecting sectors from the data. Panel on the left illustrates the original scanned sector (tongue tip on the left), middle panel shows a selection on the extreme left and panel on the right on as far right as possible. Intermediate selections can be made easily by moving the selection by one scanline at a time.

2.4. Code availability

The analysis was implemented in Python as part of the PATKIT software package (Palo et al., 2025) and stimulus lists were constructed in R (R Core Team, 2013) with the Randomise AAA Stimulus List scripts (Palo, 2024).

3. Results

Figure 5 shows the MSE matrix for the whole data without any sorting with the calibration recordings (biteplate and water swallow) removed. While there does appear to be some structure in it, it is very difficult to see a difference between blocks 1 and 2 in that image.

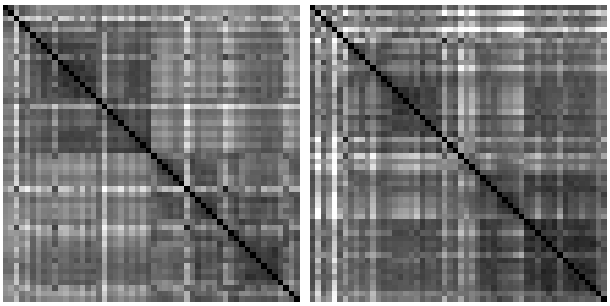

Figure 7: MSE matrix calculated on the fronted speech samples (left) and backed speech samples (right).

3.1. Varying speech materials and rotating the probe physically

Running the analysis with only the fronted or backed utterances, the situation becomes much clearer. In Figure 7 the effect of probe rotation can be seen fairly clearly in the fronted samples and less clearly in the backed samples. It appears as the two darker quadrants bisected by the diagonal and two lighter quadrants off the diagonal. This means that within block/rotation condition the recordings appear to be more similar.

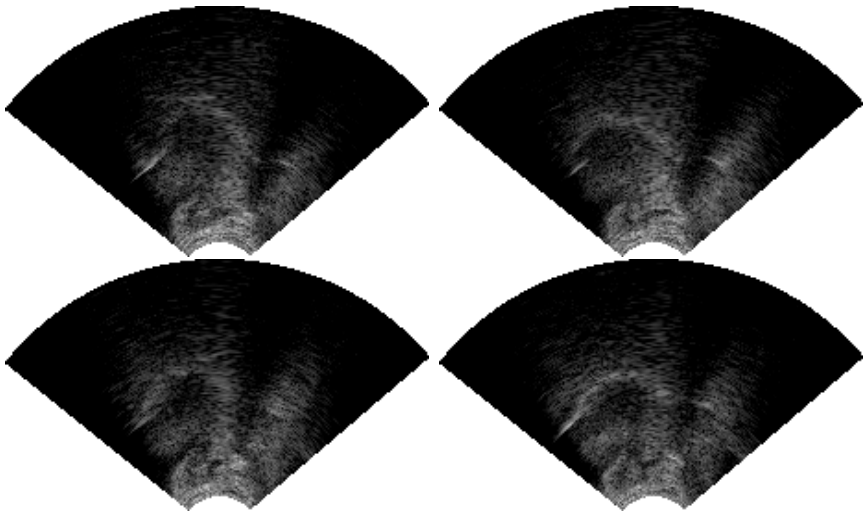

Figure 8: Bite plate frames. Top row left: block one begin, top right: block one end, bottom left block: two begin, and bottom right: block two end.

Interestingly, while the probe attachment was turned approximately 5-10° between the blocks, the biteplate traces are very much in the same direction, as seen in Figure 8. Instead, it seems that the angle of the hyoid shadow changes, but that is unfortunately more difficult to measure. And as mentioned already above in connection to Figure 4, instead of rotating it seems that the tongue contour is lowering from Block 1 to Block 2.

3.2. Simulation

The simulated data was produced from the first (unadjusted) block of original recordings. To simulate two probe positions the data was sampled as explained above in

Section 2.3. Table 1 lists the step lengths and corresponding rotation angles in radians and degrees used. For each step length condition we get two probe positions – position 1 and 2 in the following figures – rotated by the angle in Table 1. Figures 9-11 display the results from simulating probe rotation. In the Figures the recordings have been sorted within each simulated block (positions 1 and 2). Each position shows first the frontal – containing /i/ and then the backed articulations – containing /o/, and are sorted alphabetically within those sub-blocks. The squares under each step length heading contain cross-comparisons of each recording within that step length condition. No comparisons were made between non-matching step-lengths as this would mean comparing images of different size which is not a defined operation for the MSE metric.

In Figure 9 the first two vowel rows correspond to the first simulated rotation position – simulated Block 1 – and the other two to the second simulated position – simulated Block 2. The columns repeat this same pattern nesting within each step length the simulated positions (i.e., rotations), and within the positions the front and back articulatory conditions marked here by vowel qualities /i/ and /o/, respectively.

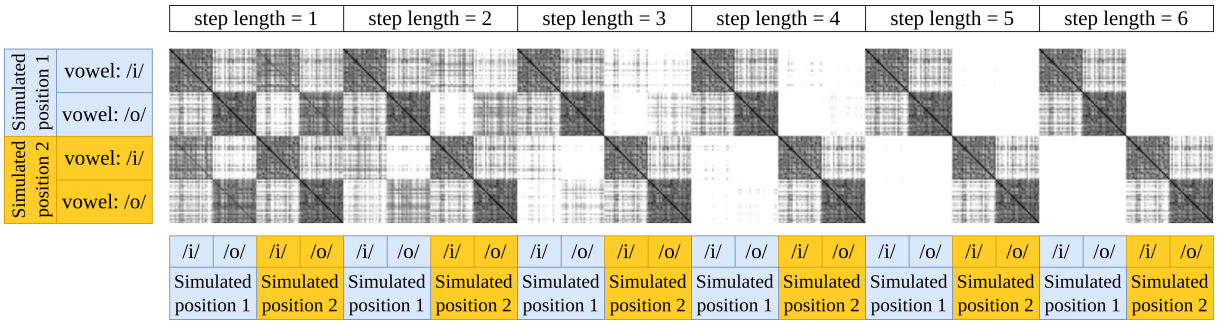

Figure 9: Effects of simulated probe rotation.

Looking at the columns where rotation step length = 1, we can see a black line on the diagonal. These are formed by pixels which correspond to recordings being compared to themselves resulting in MSE = 0. After that the darkest part of the figure are the areas where the articulation condition (/i/ or /o/) matches in the same rotation. These are the dark squares that surround the black pixels on the diagonal. Next darkest are comparisons of matching articulation types but differing rotation – for example third

Table 1: Rotation angles used in simulating probe rotation.

Step length	1	2	3	4	5	6
Radians	0.028	0.056	0.084	0.112	0.14	0.168
Degrees	1.60	3.21	4.81	6.42	8.02	9.63

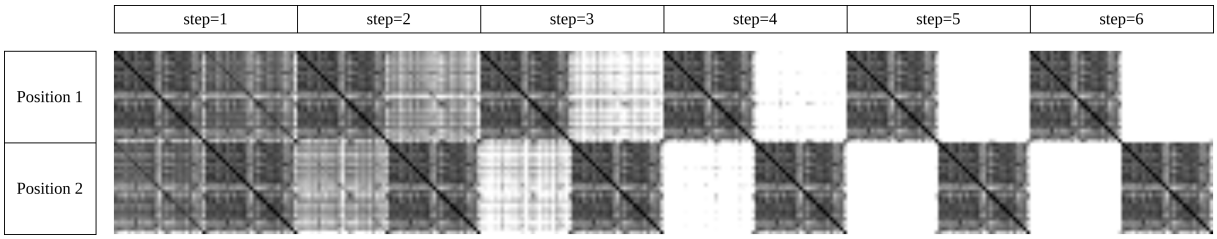

Figure 10: Simulated probe rotation for only the fronted utterances, vowel /i/ in Figure 9.

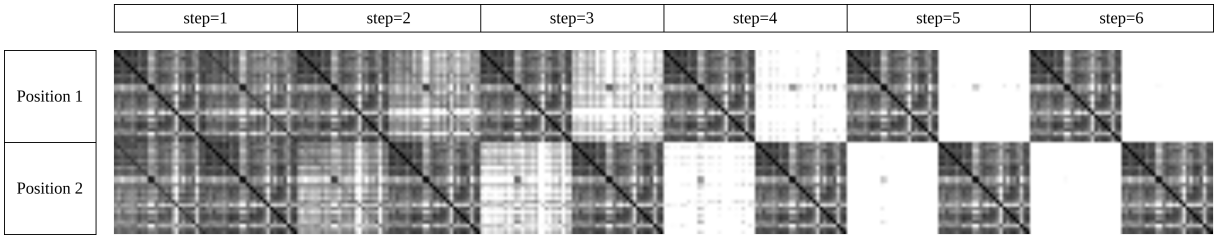

Figure 11: Simulated probe rotation for only the backed utterances, vowel /o/ in Figure 9.

quarter 'row' (Simulated position 2, vowel /i/) counting from the top in the first quarter 'column' (Simulated position 2, vowel /i/) – and only after that we get the same rotation but differing articulation type. In step = 2 the order is less clear and from step = 3 onwards the darkest squares are found in matching rotation with steps 5 and 6 having effectively totally white squares for mismatching rotation.

This gradient effect of increasing simulated rotation can be more easily observed in Figures 10 and 11. In these figures the MSE results are plotted only for fronted and backed utterances respectively. As contrast changes with subsetting the data, the effect of individual recordings is much clearer compared with Figure 9.

In Figure 10 the most prominent light lines are utterances /sisini/ and /sisisi/ in the middle of the block, and /tititi/ at the end of the block. It is unclear what sets the other two apart but /tititi/ has large non-speech movement after the acoustic utterance.

In Figure 11 the prominent utterances are /kohoho/ and /kohoko/ (in the middle of the block) which both have a similar pattern of non-speech movement to /tititi/, and (in the bottom half of the block) /ŋohoho/ – which does not have anything immediately different about the articulation, and /ŋoŋoho/ – which has a repeated utterance. It should be noted that all of the repeated utterances, false starts or too early starts were in the backed utterances in the first block, but only /ŋoŋoho/ shows any prominence in Figure 11.

4. Discussion

Both Figure 5 and the simulated rotation results show that probe alignment changes may be totally overshadowed by variation in speech materials. This means that there is reason to prefer comparing like articulations with like articulations when using MSE to evaluate probe movement. Simple means of mitigating this effect include sorting the data and plotting only part of the data at a time.

It should be noted that while /h/ was included in the speech materials in the hope of eliciting the pharyngeal fricative allophonic variant – and thus a backed tongue configuration – from the speaker, the fact that there is allophonic variation in /h/ resulted in the speaker mainly producing glottal variants. A way to avoid this problem in the future is to recruit speakers whose phonological system does not include this sort of variation, but rather specifically has a phonemic pharyngeal fricative or more generally does not include wide variation in realisations of the target sounds.

The attempt at physical rotation of the probe proved trickier than we expected. While changing probe position resulted in apparent change in tongue position, it was not the expected simple rotational change. This might be due to speaker adaptation and/or just complex mechanical interaction between the headset, probe and the speaker's anatomy. We originally intended to quantify the angle of probe adjustment based on the biteplate traces, but this proved to not be possible due to the biteplate traces not really changing.

It is easy to think that this is less of a problem in regular studies where all that matters is that the probe stays in a stable position with the desired structures in view. However, in the current context as well as more widely there are two points to consider. First, adjusting the probe holder in a given way does not necessarily have the expected result, which may affect not just how a participant makes contact with the probe but also how they articulate. Second, for purposes of simulating probe movement physical experiments involving a headset may have extra complexity in the form of speaker adaptation and physical interactions of the measurement setup and the speaker. In this sense, physically turning the probe and simulating it can be very different. This does not invalidate either approach. Rather it makes the point for using both methods for studying probe alignment evaluation methods. Physical experiments are an essential tool in ensuring that findings are grounded in reality and simulation experiments are an essential tool in checking how different manipulations affect the results in principle.

It should also be noted that the current study is limited in its use of regular 2D ultrasound. A very interesting expansion would be using 3D/4D ultrasound (Lulich et al., 2018) as the basis of the simulation. Taking 2D slices of 3D data would allow us to also examine the effects of the probe rotating position outside the sagittal plane.

Finally, we did not examine the effect of how much speech there is in a sample. Some of the bright lines visible in Figures 10 and 11, might be mitigated by limiting the analysis to just the speech. Then again depending on what we are trying to find in the data, an opposite selection – excluding the speech part from analysis – might be preferable.

5. Conclusion

We conclude that comparing like with like is the safest approach in trying to detect probe alignment issues. Furthermore, this type of evaluation method shows promise but to get closer to ground truth we need 3D/4D data as basis of the simulations and more participants.

Mean squared error as introduced for evaluating ultrasound probe alignment (Csapó & Xu, 2020), is a relatively simple but very informative tool. It is sensitive to articulatory position, probe orientation and other factors. Its power lies in combining all of these into a relative measure of similarity which can be used to flag recordings and parts of recording sessions for more fine-grained scrutiny. And as such, the method itself definitely merits further study.

Acknowledgements

This study was inspired by the work of our late colleague and friend Dr. Tamás Gábor Csapó. Pertti Palo and Steven M. Lulich remember with warmth the conversations they had with him over the years and the collaborations we were fortunate to work on with him. He is dearly missed.

Pertti Palo and Daniel Aalto work at the University of Alberta which is located on Treaty 6 territory, a traditional gathering place for diverse Indigenous peoples including the Cree, Blackfoot, Métis, Nakota Sioux, Iroquois, Dene, Ojibway/Saulteaux/Anishinaabe, Inuit, and many others whose histories, languages, and cultures continue to influence our vibrant community.

The authors wish to thank the anonymous reviewer and the editors for constructive comments that improved the quality of the article as well as Ms Hilary Warner-Evans for

proofreading the article.

References

Aalto, E. M. A., Yoshida, M., Ménard, L., Cardoso, W., & Laporte, C. (2024). Effects of an ultrasound biofeedback session on maximal tongue movements. In *Proceedings of ISSP 2024*, (pp. 63–66)., Autrans, France.

Adler, B. M., Bernhardt, B. M., Gick, B., & Bacsfalvi, P. (2007). The Use of Ultrasound in Remediation of North American English /r/ in 2 Adolescents. *American Journal of Speech-Language Pathology*, *16*(2), 128–139.

Articulate Instruments Ltd (2008). Ultrasound Stabilisation Headset Users Manual: Revision 1.4. Manual.

Articulate Instruments Ltd (2024). Articulate Assistant Advanced User Guide: Version 221.4.2. Software.

Csapó, T. G. & Xu, K. (2020). Quantification of Transducer Misalignment in Ultrasound Tongue Imaging. In *Interspeech 2020*, (pp. 3735–3739). ISCA.

Csapó, T. G., Xu, K., Deme, A., Gráczi, T. E., & Markó, A. (2020). Transducer Misalignment in Ultrasound Tongue Imaging. In *Proceedings of the 12th International Seminar on Speech Production (ISSP 2020)*, (pp. 166–169)., Online / New Haven, CT.

Dawson, K. M., Tiede, M. K., & Whalen, D. H. (2016). Methods for quantifying tongue shape and complexity using ultrasound imaging. *Clinical linguistics & phonetics*, *30*(3-5), 328–344.

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*(5), EL392.

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, *32*(2), 113–121.

Hueber, T. (2013). Ultraspeech-tools Acquisition, processing and visualization of ultrasound speech data for phonetics and speech therapy.

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

Lulich, S. M., Berkson, K. H., & de Jong, K. (2018). Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion. *Journal of Phonetics*, *71*, 410–424.

Ménard, L., Aubin, J., Thibeault, M., & Richard, G. (2012). Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model. *Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP)*, *64*(2), 64–72.

Mielke, J., Baker, A., Archangeli, D., & Racy, S. (2005). Palatron: A technique for aligning ultrasound images of the tongue and palate. *Coyote Papers*, *14*, 96–107.

Palo, P. (2024). Randomise AAA stimulus list [R and Matlab software package]. Available in a public software repository, accessed 31 October 2024. https://github.com/giuthas-speech-research-tools/randomise_AAA_stimulus_list/.

Palo, P., Moisik, S. R., & Faytak, M. (2025). PATKIT: Phonetic Analysis ToolKIT [Python software package]. Available in a public software repository, accessed 8 February 2025. https://github.com/giuthas/patkit.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Spreafico, L., Pucher, M., & Matosova, A. (2018). UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science. In *Proc. Interspeech 2018*, (pp. 1517–1520).

Stolar, S. & Gick, B. (2013). An Index for Quantifying Tongue Curvature. *Canadian Acoustics*, *41*(1), 11–15.

Stone, M. (2005). A Guide to Analyzing Tongue Motion from Ultrasound Images. *Clinical Linguistics and Phonetics*, *19*(6–7), 455–502.

Stone, M. & Davis, E. P. (1995). A head and transducer support system for making ultrasound images of tongue/jaw movement. *The Journal of the Acoustical Society of America*, *98*(6), 3107–3112.

Suomi, K., Toivanen, J., & Ylitalo, R. (2008). *Finnish Sound Structure – Phonetics, Phonology, Phonotactics and Prosody.* STUDIA HUMANIORA OULUENSIA. University of Oulu.

Whalen, D., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-Lehouillier, H., Vatikiotis-Bateson, E., & Hailey, D. S. (2005). The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech, Language, and Hearing Research*, *48*, 543–553.

Wrench, A. & Balch-Tomes, J. (2022). Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut. *Sensors*, *22*(3), 1133.

Zharkova, N., Gibbon, F. E., & Hardcastle, W. J. (2015). Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. *Clinical Linguistics & Phonetics*, *29*(4), 249–265.

Zharkova, N., Gibbon, F. E., & Lee, A. (2017). Using ultrasound tongue imaging to identify covert contrasts in children's speech. *Clinical Linguistics & Phonetics*, *31*(1), 21–34.