Transducer Misalignment in Ultrasound Tongue Imaging

Tamás Gábor Csapó[1,2], Kele Xu[3,4], Andrea Deme[5,2], Tekla Etelka Gráczi[6,2], Alexandra Markó[5,2]

[1]*Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary*
[2]*MTA-ELTE „Lendület" Lingual Articulation Research Group, Budapest, Hungary*
[3]*National Key Lab of Parallel and Distributed Processing, National University of Defense Technology, Changsha, China*
[4]*School of Computer, National University of Defense Technology, Changsha, China*
[5]*Department of Applied Linguistics and Phonetics, Eötvös Loránd University, Budapest, Hungary*
[6]*Research Institute for Linguistics, Budapest, Hungary*
csapot@tmit.bme.hu, kelele.xu@gmail.com, deme.andrea@btk.elte.hu,
graczi.tekla.etelka@nytud.hu, marko.alexandra@btk.elte.hu

Abstract

A long-standing problem for ultrasound tongue imaging is the transducer misalignment during longer data recording sessions. In this paper, we present an initial idea for analyzing such misalignment. The method employs Mean Square Error (MSE) distance to identify the relative displacement between the chin and the transducer. We visualize these measures as a function of the timestamp of the utterances. Experiments are conducted on various ultrasound tongue datasets (UltraSuite, and recordings of Hungarian children and adults). The results suggest that extreme values of MSE indicate corruptions or issues during the data recordings, which can either be caused by transducer misalignment, lack of gel, or missing contact between the skin and the transducer. The methods are language independent and might be applied in phonetic analysis of ultrasound recordings.

Keywords: ultrasound, speech production, speech articulation

1. Introduction

In order to fix head movement during the ultrasound recordings, various solutions have been proposed. Stone and Davis (1995) aimed to provide reliable tongue motion recordings by head immobilization and positioning the transducer in a known relationship to the head, by proposing the HATS system. Palatron by Mielke et al. (2005) is an algorithm to track the palate, thus could be used to align the ultrasound tongue images. Whalen et al. (2005) proposed optical tracking methods for head-correction. The metal headset of Articulate Instruments Ltd. is a popular and well designed solution which was used in a number of studies (to mention a few, articulatory-to-acoustic mapping (Csapó, Grósz, et al. 2017; Csapó, Al-Radhi, et al. 2019), Hungarian child recordings (Markó et al. 2019; Gráczi et al. 2020), and UltraSuite (Eshky et al. 2018)). Hueber et al. (2011) proposed a set of accelerometers to track the position and orientation of the transducer, relative to the head. Recently, a non-metallic system by Derrick et al. (2018) and UltraFit by Spreafico, Pucher, and Matosova (2018) are lightweight headsets to record ultrasound and EMA data.

Despite these substantial efforts, it is still a question whether the use of a headset itself is enough to ensure that the transducer is not moving during the recordings. Even if a transducer fixing system is used, large jaw movements during speech production (or drinking, swallowing) can cause the ultrasound transducer to move, and misalignment or full displacement might occur. Besides, the subjects, having discomfort due to the fixing system, sometimes readjust the headset. This way the recordings from the same session will not be directly comparable, which can be a serious issue during analysis of tongue contours. Although there exist methods for non-speech ultrasound transducer misalignment detection (Narayanan et al. 2014; Bolsterlee, Gandevia, and Herbert 2016), they cannot be directly used in speech production research.

The goal of our study was to initiate discussion of the above problems, and to propose simple methods to semi-automatically detect such ultrasound transducer misalignment or other issues during recording. Csapó and Xu (2020) presented an initial version of this topic, which is further discussed here on partly new data.

2. Methods and procedure

2.1. Ultrasound data

2.1.1. 'Hungarian children' dataset

In Gráczi et al. (2020), two Hungarian children, a girl and a boy read aloud nonsense words in 5 recording sessions within the course of 2 years, recorded using the "Micro" ultrasound system of Articulate Instruments Ltd. at 81.67 fps. The ultrasound transducer was fixed below the subjects' chin by the ultrasound stabilization headset designed for speech recordings (Articulate Instruments Ltd.). The speech signal was recorded with a Beyerdynamic TG H56c tan omnidirectional condenser microphone. The ultrasound data and the audio signals were synchronized using the tools provided by Articulate Instruments Ltd. Before each repetition, swallowing (drinking water with a straw) was also recorded, for getting information about the palate. More details about the recording set-up can be found in Csapó, Grósz, et al. (2017). The raw scanline data of the ultrasound was 64×842 pixels.

For this dataset, we acquired manual tracings for a number of images (Markó et al. 2019; Gráczi et al. 2020). Ultrasound images were extracted at the middle of the target vowels, and tongue contours were manually traced using the APIL's web-based tracer tool (`https://github.com/myedibleenso/apil-web`).

2.1.2. *'Hungarian adults' dataset*

In the 'Hungarian adults' dataset (Csapó, Al-Radhi, et al. 2019), 3 female and 6 male adults were recorded using the "Micro" system (with the same recording methodology as for 'Hungarian children') while reading 200 sentences, for articulatory-to-acoustic mapping experiments. The raw scanline data of the ultrasound was 64×842 pixels.

2.1.3. *UltraSuite*

The publicly available UltraSuite repository (Eshky et al. 2018), contains ultrasound data that was recorded using the "Micro" system, for English children of two groups, of which the UXTD (typically developing) subset was used in the current study. The raw scanline data of the ultrasound was 63×412 pixels.

2.2. Measuring transducer misalignment

In order to quantify the amount of misalignment, we compare all utterances with each other in the order in which they were recorded (Csapó and Xu 2020). First, for a given speaker and given session, we go through all of the ultrasound recordings (utterances), and calculate the pixel by pixel mean image (across time) of each utterance (see **Figure 1**). Next, we compare these mean images: we measure the Mean Square Error (MSE) between the UTI pixels ([0-255] grayscale values). MSE is an error measure, therefore the lower numbers indicate higher similarity across images. For a session with n consecutive utterances, all compared with each other, the result is an $n \times n$ matrix (see **Figure 2**). We assume that if there is misalignment in the ultrasound transducer, than the matrix of measures would show this. The full details of the method, including two more similarity measures were introduced in Csapó and Xu 2020.

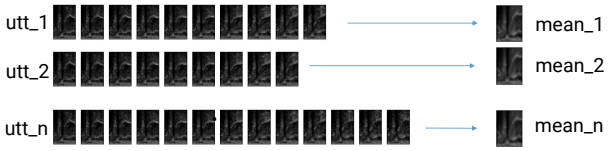

Figure 1: Calculation of pixel-by-pixel means across the image sequences of the utterances.

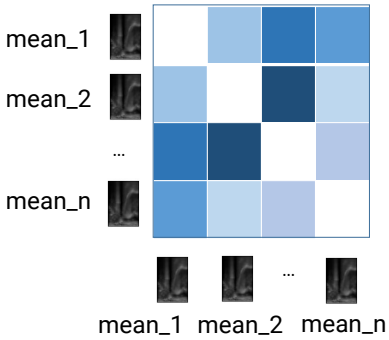

Figure 2: Calculation of the MSE matrix using the mean ultrasound images.

3. Demonstration results

The results are demonstrated in **Figures 3–7**. The figures contain samples (ultrasound images, tongue contours, and MSE measure) from a few speakers hand-selected for visualization. The MSE objective distance is shown for well aligned, misaligned and corrupted ultrasound utterance sequences.

3.1. Results on the 'Hungarian children' dataset

Figure 3 shows the MSE matrix (left) and several manual tracings (right), as a sample when the transducer did not move within the recording session (two repetitions of 81 words). In the MSE figure, all colors are bluish, indicating that MSE across most utterances is relatively small. In terms of tongue contours (**Figure 3** right), the two repetitions are similar; indicating that there was no (or only minimal) misalignment during the session.

Figure 4 shows a sample containing clear misalignment. According to MSE, utterances 1–81 are highly different from utterances 82–162. Meanwhile, differences within both utterances 1–81 and 82–162 are small. This might be because after each repetition (i.e., between utterances 81 and 82), the participant took a small break and was instructed to drink water for recording swallow. Most probably, the headset got displaced during this break. The manually traced tongue contours support this assumption: the second repetition (blue line) is shifted lower and left compared to the first repetition (red dashed line). In Gráczi et al. (2020), which was comparing the tongue contours, we had the same observation, when measuring Nearest Neighbor Distance (NND) (Zharkova, Hewlett, and Hardcastle 2011) between the tongue contours. However, NND is not suit-

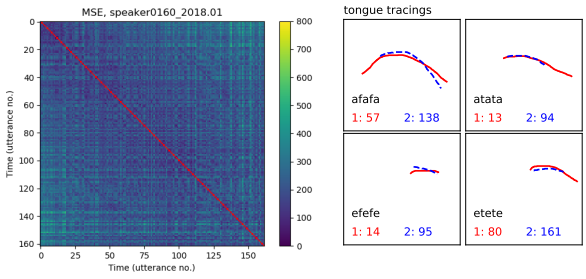

Figure 3: Sample for well aligned data across two repetitions, from the 'Hungarian children' dataset. Repetition 1: utterances 1–81; repetition 2: utterances 82–162. MSE: lower values (blue colors) indicate smaller misalignment. The diagonals contain NaN values. In the tongue tracing subfigure on the right, 1: 57 denotes that the first repetition is utterance no. 57.

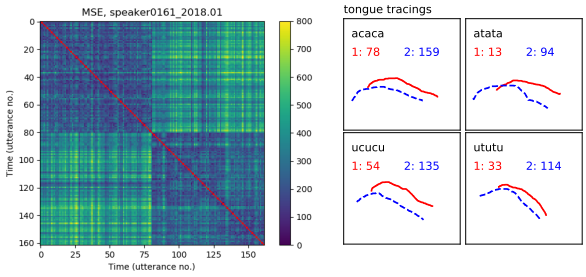

Figure 4: Strong misalignment across two repetitions, from the 'Hungarian children' dataset. Repetition 1: utterances 1–81; repetition 2: utterances 82–162.

able to quantify this type of shift, because it calculates the point-by-point minimal differences, and not the shift of the points.

3.2. Results on the 'Hungarian adults' dataset

In case we do not have manually traced tongue contours, it is more difficult to observe the misalignments on the ultrasound images itself. **Figure 5** shows a sample from an adult speaker, where the MSE matrix (left subfigure) indicates slight misalignment around frames #90–95, but it is barely visible on the mean ultrasound images plotted as a function of time (right subfigure). It is a question whether this small MSE difference is caused by regular tongue motion, or by some systematic movement of the headset compared to the head.

As another interesting example, in the MSE matrix of **Figure 6**, there are two outlier values – probably the headset was readjusted during the session, but after one single utterance, it went back to the original position. The mean ultrasound images (**Figure 6** right) do not show clearly why the outlier MSE value occurred. Without manually traced tongue contours, it is difficult to compare the MSE values with the mean images.

3.3. Results on the 'UltraSuite' dataset

For the English children dataset, **Figure 7** presents another kind of data corruption, for speaker 03F. Between utterances 3–18, and 19–28, the MSE is relatively small (whereas it is higher when comparing these two ranges). Starting from utterance 30, the MSE is extremely small; but in this case, this does not indicate well aligned transducer position. If we check the mean ultrasound images (**Figure 7** right), we can see that the transducer got fully displaced (e.g. there was no more gel between the top of the transducer and the skin), and the tongue movement was not recorded between utterances 30–55. The images in the right subfigure show that in the last utterances (e.g. in 041D), the tongue surface is not visible, most probably because of the missing contact between the transducer and the chin.

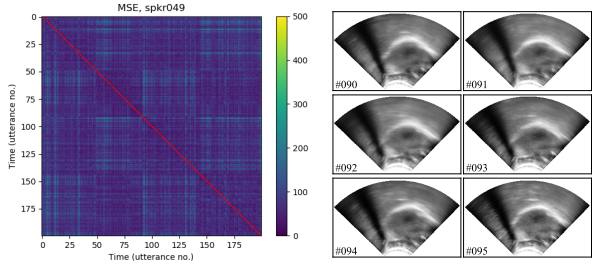

Figure 5: Slight misalignment, from 'Hungarian adults'.

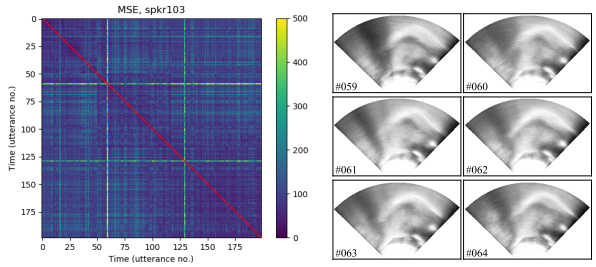

Figure 6: Occasional but strong misalignment, from 'Hungarian adults'.

4. Discussion and conclusion

For fixing ultrasound transducer position during recordings, various approaches can be used (see Sec. 1), but none of these methods are perfect and they cannot guarantee that tongue position or orientation would be the same in a longer recording session. If it is important that the relation between the tongue and the trancducer at time of various repetitions of the same target are at the same position, other methods are suggested to be used besides ultrasound transducer position fixing. An example for this is the measurement of the occlusal plane with a biteplate and rotating / shifting the data to a reference coordinate system (James M Scobbie et al. 2011; James M. Scobbie, Stuart-Smith, and Lawson 2012; Percival et al. 2020). However, it requires significant amount of manual work, and according to our knowledge, until now there have been no methods for auto-rotating within longer ultrasound recording sessions.

We have shown how the MSE misalignment measure indicates various issues in ultrasound recordings of tongue movements: slight, strong, and occasional misaligments due to headset issues, and lack of gel. These can be critical when tongue contours are traced for articulatory investigations. Although we did not attempt to show a direct relationship between the quantified measure and amount of shift in tongue tracings, the results might be useful for phonetic research investigating tongue shapes and positions (Palo, Schaeffler, and J. Scobbie 2020). Unfortunately, the single (mid)sagittal recordings do not allow to track for changes in the rotation or orientation of the transducer, or lateral misalignments – for this, optical tracking or accelerometers are necessary as suggested by Whalen et al. (2005) and Hueber et al. (2011). The methods can easily be applied on other datasets (containing wedge-formatted, non-raw ultrasound data), other languages, and other imaging techniques (e.g. MRI or lip video).

In the future we plan to develop automatic classification methods to warn during analysis of the tongue contours if the ultrasound transducer is clearly misaligned within a recording session; or give confidence intervals related to the reliability. It is also possible that toolkits from the field of medical imaging registration or fusion can be applied for our purposes. Checking the neutral tongue position (e.g. at the beginning or end of utterances) and detecting the change in this reference image could also help for the automatic detection of recording issues. In future work, we also plan to investigate transducer misalignments in a controlled experiment, i.e. how visible are the shifts of transducer on the ultrasound images?

The code implementations and the MSE matrix images of all subjects are accessible at `https://github.com/BME-SmartLab/UTI-misalignment/`.

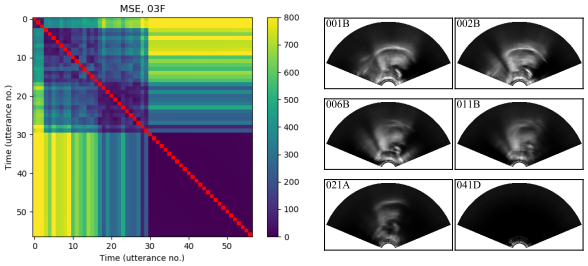

Figure 7: Corrupted data, from 'UltraSuite'.

5. Acknowledgements

The authors acknowledge the kind help of Julianna Jankovics in the manual tongue tracings, and thank the Ultrax2020 project for providing the UltraSuite articulatory database. The research was funded by the National Research, Development and Innovation Office of Hungary (FK 124584 and PD 127915 grants), by the Bolyai János Research Scholarschip of the Hungarian Academy of Sciences, the ÚNKP-20-5 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund, and the Thematic Excellence Program of the Ministry for Innovation and Technology.

6. References

Bolsterlee, Bart, Simon C. Gandevia, and Robert D. Herbert (2016). "Effect of Transducer Orientation on Errors in Ultrasound Image-Based Measurements of Human Medial Gastrocnemius Muscle Fascicle Length and Pennation". In: *PLOS ONE* 11.6. Ed. by Miklos S. Kellermayer, e0157273.

Csapó, Tamás Gábor, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó (2017). "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface". In: *Proc. Interspeech*. Stockholm, Sweden, pp. 3672–3676.

Csapó, Tamás Gábor, Mohammed Salah Al-Radhi, Géza Németh, Gábor Gosztolya, Tamás Grósz, László Tóth, and Alexandra Markó (2019). "Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder". In: *Proc. Interspeech*. Graz, Austria, pp. 894–898. arXiv: 1906.09885.

Csapó, Tamás Gábor and Kele Xu (2020). "Quantification of Transducer Misalignment in Ultrasound Tongue Imaging". In: *Proc. Interspeech*. Shanghai, China, pp. 3735–3739.

Derrick, Donald, Christopher Carignan, Wei-rong Chen, Muawiyath Shujau, and Catherine T. Best (2018). "Three-dimensional printable ultrasound transducer stabilization system". In: *The Journal of the Acoustical Society of America* 144.5, EL392–EL398.

Eshky, Aciel, Manuel Sam Ribeiro, Joanne Cleland, Korin Richmond, Zoe Roxburgh, James M Scobbie, and Alan Wrench (2018). "UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions". In: *Proc. Interspeech*. Hyderabad, India: ISCA, pp. 1888–1892.

Gráczi, Tekla Etelka, Tamás Gábor Csapó, Márton Bartók, Andrea Deme, and Alexandra Markó (2020). "Articulatory and acoustic differentiation of /s/ and /S/ in children's speech: longitudinal case studies". In: *(Dys)fluency in children's speech*. Ed. by Judit Bóna.

Hueber, Thomas, Bruce Denby, G Dreyfus, R Dubois, and P Roussel (2011). *Device for reconstructing speech by ultrasonically probing the vocal apparatus*.

Markó, Alexandra, Tamás Gábor Csapó, Andrea Deme, Tekla Etelka Gráczi, and Márton Bartók (2019). "Gyermekek lingvális artikulációjának variabilitása magánhangzós nyelvkontúrok alapján". In: *Az anyanyelv-elsajátítás folyamata hároméves kor után*. ELTE Eötvös Kiadó, pp. 165–190.

Mielke, Jeff, Adam Baker, Diana Archangeli, and Sumayya Racy (2005). "Palatron: a technique for aligning ultrasound images of the tongue and palate". In: *Coyote Papers* 14, pp. 97–108.

Narayanan, M. M., Narender Singh, Anish Kumar, C. Babu Rao, and T. Jayakumar (2014). "An absolute method for determination of misalignment of an immersion ultrasonic transducer". In: *Ultrasonics* 54.8, pp. 2081–2089.

Palo, Pertti, Sonja Schaeffler, and James Scobbie (2020). "Change Measures for Ultrasound: Relating Pixel Difference on Raw Data to Nearest Neighbour on Splines". In: *UltraFest IX*.

Percival, Maida, Tamás Gábor Csapó, Márton Bartók, Andrea Deme, Tekla Etelka Gráczi, and Alexandra Markó (2020). "Ultrasound imaging of Hungarian geminates". In: *UltraFest IX*.

Scobbie, James M, Eleanor Lawson, Steve Cowen, Joanne Cleland, and Alan A Wrench (2011). "A common co-ordinate system for mid-sagittal articulatory measurement". In: *QMU CASL Working Papers WP-20* June.

Scobbie, James M., Jane Stuart-Smith, and Eleanor Lawson (2012). "Back to front: A socially-stratified ultrasound tongue imaging study of Scottish English /u/". In: *Italian Journal of Linguistics* 24.1, pp. 103–148.

Spreafico, Lorenzo, Michael Pucher, and Anna Matosova (2018). "UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science". In: *Proc. Interspeech*. Hyderabad, India, pp. 1517–1520.

Stone, Maureen and EP Davis (1995). "A head and transducer support system for making ultrasound images of tongue/jaw movement". In: *The Journal of the Acoustical Society of America* 98, pp. 3107–3112.

Whalen, D H, Khalil Iskarous, Mark K Tiede, David J Ostry, Heike Lehnert-Lehouillier, Eric Vatikiotis-Bateson, and Donald S Hailey (2005). "The Haskins optically corrected ultrasound system (HOCUS)". In: *Journal of Speech, Language and Hearing Research* 48.3, pp. 543–553.

Zharkova, Natalia, Nigel Hewlett, and William J Hardcastle (2011). "Coarticulation as an Indicator of Speech Motor Control Development in Children : An Ultrasound Study Acquisition of Coarticulation by Children". In: *Motor Control* 15, pp. 118–140.