Transducer Misalignment in Ultrasound Tongue Imaging

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

[1]Budapest University of Technology and Economics [2]MTA-ELTE ,,Lendület" Lingual Articulation Research Group [3]National University of Defense Technology, Changsha, China [4]Eötvös Loránd University [5]Research Institute for Linguistics, Budapest, Hungary csapot@tmit.bme.hu

Introduction In order to fix head movement during the Ultrasound Tongue Imaging (UTI), various solutions have been proposed, e.g. the metal headset of Articulate Instruments Ltd. [1, 2], UltraFit [3], and others listed in [4]. Despite these substantial efforts, it is 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 can cause the ultrasound transducer to move, and misalignment or full displacement might occur. This way the recordings from the same session will not be directly comparable, which can be a serious issue during analysis of tongue contours.

Methods and procedure Here we analyze several datasets containing ultrasound tongue images, and show that transducer misalignment or other issues during recording can often happen, but can be automatically detected. *'Hungarian children' dataset [1]:* two 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 Instrum. Ltd. Manual tracings were acquired for a number of images (in the middle of the target word). *'Hungarian adults' dataset [5]:* 3 female and 6 male adults recorded using the "Micro" system while reading 200 sentences. *UltraSuite [2]:* ultrasound data recorded using the "Micro" system, for English children of two groups, of which the UXTD (typically developing) subset was used.

In order to quantify the amount of misalignment, we comprare all utterances with each other in the order in which they were recorded [4]. 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. 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 will be an $n \times n$ matrix. We assume that if there is misalignment in the ultrasound transducer, than the matrix of measures will show this. The full details of the method, including two more similarity measures are introduced in [4].

Demonstration results The results are demonstrated in Figures 1–5. The figures contain samples (ultrasound images, tongue contours, and MSE measure) from a few speakers hand-selected for visualization.

'Hungarian children': Fig. 1 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 (Fig. 1 right), the two repetitions are similar; indicating that there was no (or only minimal) misalignment during the session. Fig. 2 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).

'Hungarian adults': In case we do not have manually traced tongue contours, it is more difficult to observe the misalignments on the ultrasound images itself. Fig. 3 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). As another interesting example, in the MSE matrix of Fig. 4, there are two outstanding 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 (Fig. 4 right) do not show clearly why the outstanding MSE value occurred. Without manually traced tongue contours, it is difficult to compare the MSE values with the mean images.

UltraSuite: Fig. 5 presents another kind of data corruption, for speaker 03F. 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 (Fig. 5 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 skin.

Discussion and conclusions We have show hown 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. 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).

The code implementations are accessible at `https://github.com/BME-SmartLab/UTI-misalignment/`.
The first author was funded by the NRDIO of Hungary (FK 124584 and PD 127915 grants).

References

[1] T. E. Gráczi *et al.*, "Articulatory and acoustic differentiation of /s/ and /S/ in children's speech: longitudinal case studies," in *(Dys)fluency in children's speech*, J. Bóna, Ed., 2020.

[2] A. Eshky *et al.*, "UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions," in *Interspeech*, 2018, pp. 1888–1892.

[3] L. Spreafico *et al.*, "UltraFit: A Speaker-friendly Headset for Ultrasound Recordings in Speech Science," in *Interspeech*, 2018, pp. 1517–1520.

[4] T. G. Csapó and K. Xu, "Quantification of Transducer Misalignment in Ultrasound Tongue Imaging," in *subm. Interspeech*, 2020.

[5] T. G. Csapó *et al.*, "Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder," in *Interspeech*, 2019, pp. 894–898.

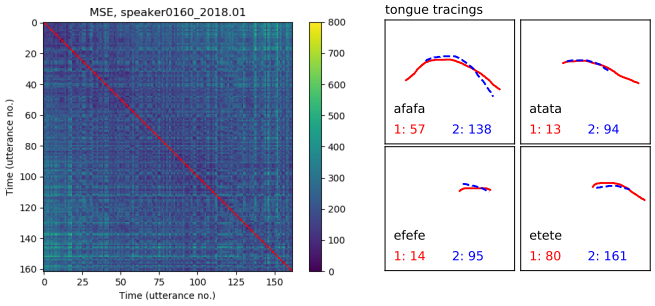

Figure 1: Sample for well aligned data across two repetitions, from the 'Hungarian children' dataset. Repetition 1: utterances 1–81; repetition 2: utterances 82–162. MSE: lower values (blue colors) indicate smaller misalignment. The diagonals contain NaN values. In the tongue tracing figure (last column),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

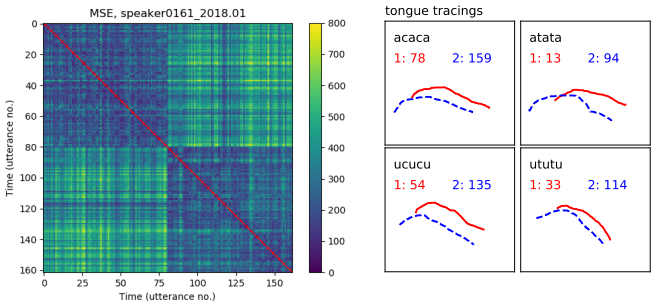

Figure 2: Strong misalignment across two repetitions, from the 'Hungarian children' dataset. Repetition 1: utterances 1–81; repetition 2: utterances 82–162.

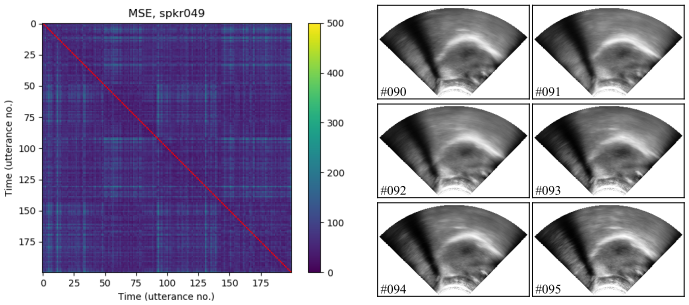

Figure 3: Slight misalignment, from 'Hungarian adults'.

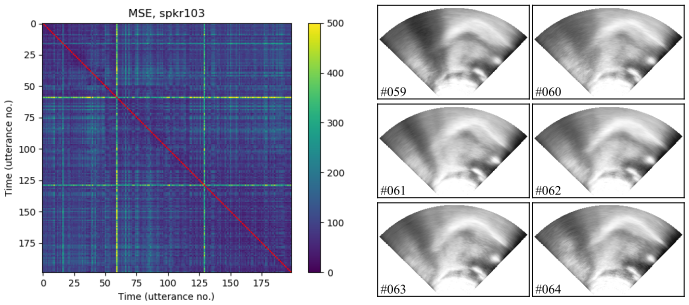

Figure 4: Occasional but strong misalignment, from 'Hungarian adults'.

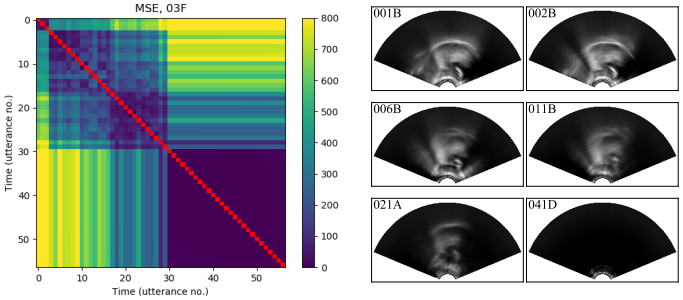

Figure 5: Corrupted data, from 'UltraSuite'.