MTracker - a CNN-based tool for automatic tracking of tongue contours

Jian Zhu [1], *Will Styler* [2], *Ian Calloway* [1]

[1] Department of Linguistics, University of Michigan, USA
[2] Department of Linguistics, University of California San Diego, USA

lingjzhu@umich.edu, wstyler@ucsd.com, iccallow@umich.edu

Background: In linguistic and clinical phonetics, extracting tongue contours is usually the first step in analyzing ultrasound images, but this process is time-consuming. In the past decade, various methods for semi-automatic or automatic tongue contour extraction have been proposed to facilitate the analysis of ultrasound data, notably the Active Contour (Snake) based methods (e.g. M. Li et al., 2005), graph based methods (Tang& Hamarneh, 2010), and neural network based methods (e.g.Jaumard-Hakoun et al., 2016).

Dense U-Net is a network architecture that can complete segmentation tasks at the pixel level (e.g. X. Li et al., 2018). In this paper, we describe an implementation of Dense U-Net to identify ultrasound tongue surface contours. This model has been packaged as MTracker, an open-source, publicly available software tool for the automatic extraction and manual correction of tongue surface contours. We also report the performance of this tool across several data sets.

Methods: The standard DenseNet-121 architecture (Huang et al., 2017) was adopted for the downsampling and upsampling paths of this model. The loss function used for this model was a weighted sum of two measures: 1) Dice Similarity Coefficient (Milletari et al., 2016), which penalizes a mismatch between the predicted white pixels (representing the tongue region) and the white edge in the predicted contour, and 2) cross-entropy loss. Foreach input image, the output is a heatmap of the same dimension as the input, with intensity corresponding to the likelihood that the pixel is part of the tongue. A 50% threshold is applied to filter out unlikely predictions. A skeletonization algorithm is then used to reduce the white edge to a single pixel wide representation. It is then interpolated and smoothed using 'UnivariateSpline' in the SciPy Package with default settings. The resulting output is a 100-point Cartesian coordinate representation of the predicted tongue shape. Training data were 35,160 human-annotated ultrasound frames from 11 American English speakers producing vowel and vowel-lateral syllable nuclei in C2lC and C2C pairs (e.g. 'bulk' and 'buck'). The data were split into training, validation and test sets through random partitioning, each consisting of 45%, 5% and 50% of the total data respectively. This model was trained only on the training dataset. Three additional data sets were used to test the model: an ultrasound recording of two American English speakers reading 'The North Wind and the Sun' (NS), the Ultrax data set (Eshky et al., 2018),and the UltraSpeech data set (Fabre et al., 2017). All training and test images were scaled to 128×128 pixels.

Test set	NS	Ultrax	Ultraspeech
3.79	5.71	5.60	5.72

Table 1: *Mean (and standard deviation) of MSD (in pixels) for different test datasets*

Results and discussion: The metric for evaluation of error from human annotation was Mean Sum of Distance (MSD), which permits the comparison of two curves without requiring point-wise alignment (M. Li et al., 2005). As observed in Table1, the model achieved the lowest MSD when tracking images generated by the same speaker and machine as the training set, but the respective MSDs for the other three data sets were comparable to one another, despite differences in machine and speaker background. The average speed for this model is approximately 10-20 frames per second on a consumer-grade laptop without a GPU. The contour extraction performed by our tool can potentially facilitate the time-consuming manual annotations in phonetic and clinical research. This tool, data and its corresponding paper has been made freely available online at: https://github.com/lingjzhu/mtracker.github.io.

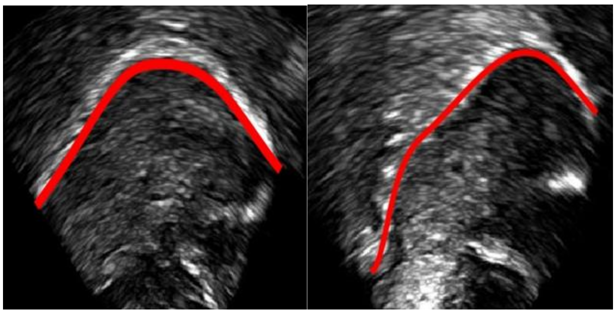

Figure 1: *Sample predictions given by Dense U-Net*

References

Eshky, A., Ribeiro, M. S., Cleland, J., Richmond, K., Roxburgh,Z., Scobbie, J. M., & Wrench, A. A. (2018). Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In Interspeech 2018: Proceedings of the 19thannual conference of the international speech communication association (ISCA), 2-6 September 2018, Hyderabad, India.

Fabre, D., Hueber, T., Girin, L., Alameda-Pineda, X., & Badin,P. (2017). Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. Speech Communication,93, 63–75.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q.(2017). Densely connected convolutional networks. In Cvpr(Vol. 1, p. 3).

Jaumard-Hakoun, A., Xu, K., Roussel-Ragot, P., Dreyfus, G.,& Denby, B.(2016, May). Tongue contour extraction from ultrasound images based on deep neural network.arXiv:1605.05912 [cs]. (arXiv: 1605.05912)

Li, M., Kambhamettu, C., & Stone, M.(2005, January). Automatic contour tracking in ultrasound images. Clinical Linguistics & Phonetics,19(6-7), 545–554.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., & Heng, P.-A.(2018). H-dense unet: Hybrid densely connected Unet for liver and tumor segmentation from CT volumes. IEEE Transactions on Medical Imaging.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 3D vision (3DV), 2016 Fourth International Conference on 3D Vision (pp. 565–571).

Tang, L., & Hamarneh, G. (2010, June). Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 154–161).

Keywords:
- Methodological research
- Ultrasound Tongue tracking
- Deep learning