Testing AutoTrace: A Machine-learning Approach to Automated Tongue Contour Data Extraction

Jae-Hyun Sung[1], **Jeff Berry**[2], **Marissa Cooper**[1], **Gustave Hahn-Powell**[1], and **Diana Archangeli**[3]

[1]*Department of Linguistics, University of Arizona, USA*
[2]*InsideSales.com*
[3]*Linguistics, University of Hong Kong, HK*

Articulatory imaging is important for analyzing the rules of speech and can be utilized for many purposes such as analyzing sounds in different dialects, learning a second language, and speech therapy (Archangeli and Mielke (2005), Adler-Bock et al. (2007), Gick et al. (2008), Scobbie et al. (2008)). When analyzing phonological data, researchers can implement various experimental methods concerning articulation - EMMA, MRI, Palatography, and ultrasound. Although ultrasound is inexpensive, non-toxic, and portable, it does have a significant drawback. After the data is collected, someone must then trace the tongue surface contours, which creates a bottleneck for analyzing the results. Several different approaches to this problem have been proposed (Li et al. (2005), Fasel and Berry (2010), Tang et al. (2012)), with promising results. In this paper, we analyze the performance of the Deep Neural Network approach of Fasel and Berry (2010) at scale, and announce an open source project to further develop this method, named AutoTrace.

AutoTrace automates the process of extracting the data from ultrasound images, greatly reducing the amount of time necessary for tracing images, as shown in Berry et al. (2012) on a small data set. This paper reports on our tests of the efficacy of AutoTrace on a much larger data set consisting of approximately 40,600 ultrasound images taken from Harvard sentences read by 12 American English speakers. This ensured a wide variety of tongue shapes, due both to different speakers and to different types of sounds. For training data sets, we selected a combination of most and least diverse images based on their deviation from pixel averages, using the heuristic proposed by Berry (2012).

AutoTrace used training data sets of different sizes to learn networks. Each network was tested against the same set of 100 randomly selected images. These traces were hand-corrected by human expert tracers, and each network was retrained. The automatically traced contours were compared to traces made by human experts to gauge how well the program performed.

Four separate tests are considered:

a. Most diverse images only vs. most + least diverse images: the combination most & least gave better results.

b. n images vs. n images (ranging from 250 to 1056 images) taken at intervals of y: the larger the training set, the better the results.

c. No retraining vs. retraining: retraining produced a "41% decrease in the number of images needing hand correction" (Berry, 2012, p. 50).

d. n images from the most diverse set vs. r randomly selected images from the whole ($n = r$): the n most diverse images performed better.

Currently, the comparison of two human expert tracers shows a pixel-by-pixel average difference of 2.467 pixels per image. Comparison of machine vs. human tracers shows a pixel-by-pixel average difference of 5.656 pixels per image. We are currently examining the types of errors AutoTrace is making to see whether we can improve the training technology for better results.

References

Adler-Bock, M., Bernhardt, B. M., Gick, B., and Bacsfalvi, P. 2007. The use of ultrasound in remediation of English /r/ in two adolescents. *American Journal of Speech-Language Pathology*, 16(2):128–139.

Archangeli, D. and Mielke, J. 2005. Ultrasound research in linguistics. In *34th Annual Meeting of The Linguistic Association of the Southwest*.

Berry, J. 2012. *Machine learning methods for articulatory data*. PhD thesis, The University of Arizona.

Berry, J., Fasel, I., Fadiga, L., and Archangeli, D. 2012. Training deep nets with imbalanced and unlabeled data. In *Interspeech*.

Fasel, I. and Berry, J. 2010. Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 1493–1496.

Gick, B., Bernhardt, B. M., Bacsfalvi, P., and Wilson, I. 2008. Ultrasound imaging applications in second language acquisition. In Edwards, J. G. H. and Zampini, M. L., editors, *Phonology and Second Language Acquisition*, pages 309–322. John Benjamins.

Li, M., Kambhamettu, C., and Stone, M. 2005. Automatic contours tracking in ultrasound images. *Clinical Linguistics and Phonetics*, 19:545–554.

Scobbie, J. M., Stuart-Smith, J., and Lawson, E. 2008. Looking variation and change in the mouth: developing the sociolinguistic potential of ultrasound tongue imaging. Queen Margaret University.

Tang, L., Bressmann, T., and Hamarneh, G. 2012. Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves. *Medical Image Analysis*, 16:1503–1520.