# Error analysis of extracted tongue contours from 2D ultrasound images

*Tamás Gábor Csapó*[1], *Steven M. Lulich*[2]

[1]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
[2]Department of Speech and Hearing Sciences, Indiana University, Bloomington, IN, USA
csapot@tmit.bme.hu, slulich@indiana.edu

## Abstract

The goal of this study was to characterize errors involved in obtaining midsagittal tongue contours from two-dimensional ultrasound image sequences. Toward that end, two basic experiments were conducted. First, manual tongue contours were obtained from 1,145 tongue ultrasound images recorded from four speakers during production of the sentence *'I owe you a yoyo'*, and the uncertainty associated with the contours was quantified. Second, tongue contours from the same images were obtained using the EdgeTrak, TongueTrack, and AutoTrace algorithms, and these were compared quantitatively with the manual tongue contours. Three basic error types associated with the tongue contours are identified, indicating areas in need of improvement in future algorithmic developments. Depending on the speaker, RMS errors for the algorithmically obtained contours ranged from 1.76 to 7.11 mm, and the standard deviation of manual contours ranged from 0.97 to 2.07 mm.

**Index Terms**: ultrasound, tongue contour, automatic tracking

## 1. Introduction

Phonetic research has employed 2D ultrasound for a number of years for investigating tongue movements during speech [1, 2]. The typical result of 2D ultrasound recordings is a series of gray-scale images in which the tongue surface contour has a greater brightness than the surrounding tissue and air (for a guide to tongue ultrasound imaging and processing, see [2]). Extracting tongue contours from these images is critical for later analyses, including comparison of tongue shapes, measuring parameters related to tongue curvature, addressing phonological questions related to articulation, and so on. Although manual tracing of an image can be as fast as 2 seconds, for a continuous image sequence at typical ultrasound frame rates of 30–100 fps, manual tracing is not a practical option [3].

While clear applications for tongue ultrasound exist in linguistics, speech science, speech and swallowing therapy, and orthodontics [4, 5, 6, 7, 8], studies of the kinds and magnitudes of errors associated with manual and automatic tracing are few. Previous studies of variability in manual tongue contours traced by pairs of experts yielded maximum errors between $0.49$ and $0.7$ $mm$ in [1, 9], and mean absolute errors between $0.73$ and $2.04$ $mm$ in [3, 10, 11]. Previous studies of errors in tongue contour tracings generated by computer algorithms yielded mean absolute errors between $0.54$ $mm$ and $1.06$ $mm$ for the EdgeTrak program [10], and between 2 and $4$ $mm$ for the TongueTrack program [11]. For the AutoTrace program, the mean absolute error was reported as $5.656$ $pixels$ [3], which is $1.67$ $mm$ if a conversion factor of $0.295$ mm per pixel is assumed, as in [11].

This paper expands the error analyses of the previous studies to 1) examine and quantify the variability of tongue contours traced manually by multiple individuals, and 2) characterize and quantify the major errors associated with tongue contour tracings obtained automatically from EdgeTrak, TongueTrack, and AutoTrace.

## 2. General Methods

Two female and two male adult subjects (denoted F1, F2, M1 and M2) with normal speaking abilities were recorded producing the sentence *'I owe you a yoyo'* twice, using a Philips EpiQ-7G ultrasound system with an xMatrix 6–1 MHz transducer. The recordings were made in a soundproof booth in the Speech Production Laboratory at Indiana University. The ultrasound recordings were performed in accordance with guidelines published in [2]: 1) the ultrasound transducer was held in the subject's hand and slightly pressed against the chin, 2) a midsagittal orientation was maintained with the shadows of the jaw and the hyoid bones visible at opposite sides of the scan wedge, 3) the midline was continuously examined by the experimenter. The image frame rates were between $42 - 44$ $fps$ for each speaker. The ultrasound data were recorded in DICOM format with $800x600$ resolution and they were converted to JPG images using Image-J [12].

There was a combined total of $1,145$ ultrasound tongue images ($389$, $275$, $241$ and $240$ for speakers F1, F2, M1, and M2, respectively). The image sequences were split into two halves because some automatic tongue contour tracing programs can not load entire sequences of this length. There were therefore $8$ ultrasound image sequences (two for each speaker) analyzed in this study. Each sequence contained one complete utterance of the phrase *'I owe you a yoyo'*. The speaking rate varied from $2.53$ to $4.65$ syllables per second (the mean was $3.74$ syllables per second). It is well known that ultrasound image quality is highly variable and dependent on a number of factors [2]. For our corpus, speaker F1 was in general the best, followed in order by F2, M1, and M2.

## 3. Experiment 1: Manual tracing

### 3.1. Methods

An Apache webserver was used to present ultrasound images on a secure website developed by the Speech Production Laboratory at Indiana University. Authorized tracers could therefore access the images via a web browser. Seven tracers participated in this experiment. Two of the tracers were the authors, and the remaining five tracers were undergraduate students in the Speech and Hearing Sciences Department at Indiana Univer-

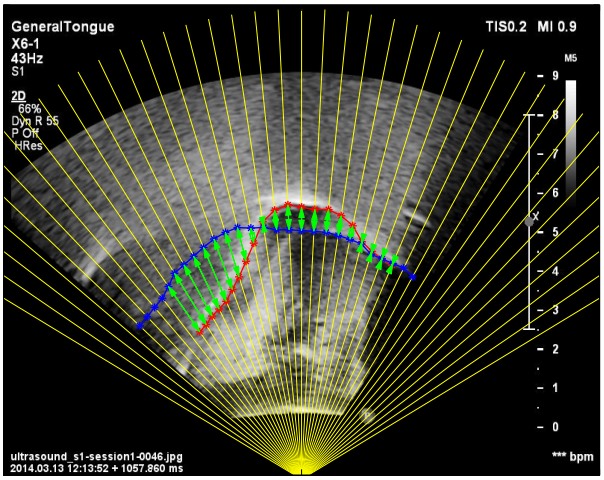


Figure 1: Sampled tongue tracings for speaker F1. The 41 radial lines are shown in yellow. The red curve is a manual tracing, the blue curve is an automatic tracing with significant error. Green arrows show the distances between the points in the two curves.

sity. The student tracers were given training with feedback before starting to trace the images for this study. Each of the seven individuals traced each of the $1,145$ images by using a mouse to click on the image in their web browser, dragging the mouse along the visible tongue contour from the left (posterior) side of the image to the right (anterior) side. The x-y coordinates of the traced image pixels were then saved to an SQL database for further anlaysis. The cumulative time required for each individual to trace the $1,145$ images ranged from approximately $3$ to $5$ hours.

A radial coordinate system was defined in order to compare tracings and to quantify variability. The origin was located at the point of intersection between the straight lines defining the sides of the ultrasound wedge. A total of 41 radial lines was defined, spanning $-60$ to $60$ degrees (relative to vertical) in steps of $3$ degrees. All tongue contours were up-sampled by linear interpolation, and then down-sampled to $41$ points falling along the $41$ radial lines. Figure 1 shows an example of the final sampled tongue contours for one manual tracing and one automatic tracing.

In order to quantify the uncertainty associated with the manual tracings, the mean and unbiased standard deviation of values along each radial line were obtained for each frame. Furthermore, the grand mean and standard deviation were calculated across all radial lines and across all frames for each of the 8 image sequences. It frequently occurred that the extent of the seven tracings toward the left and right edges was not uniform, with some tracings extending further than others. Mean values were obtained for all radial lines with at least one tracing value, and standard deviations were obtained for all radials lines with at least two tracing values.

### 3.2. Results

The distribution of standard deviations across radial lines and frames for each of the image sequences was skewed and roughly log-normal (data not shown). Therefore, the grand mean and grand standard deviation were calculated from the log-transformed distributions and then transformed back to millimeter units. The grand mean and standard deviation of the unbiased radial line standard deviations are given separately for each speaker in the top line of Table 1. The overall unbiased standard deviation was $1.33\ mm$, and ranged from $0.95\ mm$ for speaker F1 to $2.11\ mm$ for speaker M2. These standard deviations for seven tracers with varying levels of experience are very similar to those reported previously for pairs of expert tracers.

Table 1: Grand mean and standard deviations of unbiased standard deviations (manual) and RMSEs (automatic) of tongue contour tracings (in mm).

| tracer | F1 | F2 | M1 | M2 | avg |
|---|---|---|---|---|---|
| Manual | 0.95 (0.29) | 1.09 (0.32) | 1.17 (0.31) | 2.11 (0.32) | 1.33 (0.31) |
| AutoTrace3.5 | 1.15 (0.35) | 1.93 (0.31) | 1.78 (0.29) | 2.19 (0.28) | 1.76 (0.31) |
| AutoTrace3 | 5.85 (0.33) | 7.06 (0.43) | 5.59 (0.32) | 9.94 (0.28) | 7.11 (0.34) |
| EdgeTrak | 1.95 (0.45) | 3.46 (0.37) | 1.89 (0.41) | 5.15 (0.40) | 3.11 (0.41) |
| TongueTrack | 1.96 (0.53) | 3.15 (0.37) | 2.76 (0.38) | 3.60 (0.37) | 2.87 (0.41) |
| Baseline | 3.59 (0.40) | 4.32 (0.33) | 4.50 (0.33) | 4.01 (0.37) | 4.11 (0.36) |

## 4. Experiment 2: Automatic tracing

### 4.1. Methods

A number of semi-automatic and automatic solutions have been proposed for tracing tongue contours from ultrasound images. This study makes use of the three automatic tracing programs that are freely available: **EdgeTrak**, which uses a snakes-based algorithm [13, 14, 10]; **TongueTrack**, which uses a machine learning approach in combination with a higher-order Markov Random Field energy minimization frame [15, 16, 11]; and **AutoTrace**, which uses deep belief networks (DBNs) that rely on prior tongue contour tracings for training [3, 17, 18, 19, 20].

Because the focus of this experiment is on characterizing the kinds and magnitudes of errors associated with automatic tracing algorithms in general, and since optimization of program parameters is likely to be highly dependent on the image data set, the default 'out of the box' parameters were adopted for each of these programs. The results therefore do not necessarily represent optimal performances by any of these three programs. The procedures for obtaining tracings of all $1,145$ images from each of the three programs are described below.

In this study, we compare the results of automatic tongue contour tracking using AutoTrace, EdgeTrak and TongueTrack. In the following, we describe the details that were used for each of the three automatic contour tracking programs.

For **AutoTrace**, the ultrasound JPG images were resized to $720x480$ pixels and shifted so that they fit within the internally defined wedge and radial coordinate system of the program. The Region of Interest (RoI) was set manually. AutoTrace requires manual tracings for training the DBNs, and its performance was tested twice using different sets of images for training and testing. In both cases, the mean of the 7 manual tracings for each frame in the training set was used.

In the first test, denoted 'AutoTrace3.5', the training set consisted of images from 7 of the 8 sequences (hence, 3.5 of the 4 speakers' recordings) and the test set consisted of images from the remaining sequence (half of the remaining speaker's

recordings). This test was repeated 8 times so that all 8 image sequences were in the test set one time. Because each sequence contained the phrase *'I owe you a yoyo'*, the data in the test set is always closely matched with the data from the same speaker in the training set.

In the second test, denoted 'AutoTrace3', the training set consisted of images from 6 of the sequences (hence, 3 of the 4 speakers' recordings) and the test set consisted of images from the remaining two sequences (from the remaining speaker). This test was repeated 4 times so that both sequences from each speaker were tested. In this case the training data and the test data are mismatched, although the utterance was the same for each sequence.

For **EdgeTrak**, the original $800x600$ pixel JPG images were used. For each sequence, the program was initialized by providing the mean manual tracing of the first image, and the RoI was manually determined for each sequence.

For **TongueTrack**, anisotropic and despeckle filters were applied to the original $800x600$ pixel JPG images in accordance with the TongueTrack manual, and the part of each image falling within the RoI defined for EdgeTrak was extracted and saved in an MHD format using the Medical Image Processing Toolbox [21]. Automatic tracings were obtained from these MHD images. For each sequence, the program was initialized by providing the mean manual tracing of the first image.

### 4.2. Analyses

Errors in the automatic tracings were determined relative to the mean of the 7 manual tracings for each frame. This required the automatic tracings to be sampled in the same way as the manual tracings, in order to conform to the 41 radial lines described above. Radial lines for which either the mean manual tracing or the automatic tracing was not defined were excluded from the analysis. An example is given in Fig. 1, in which only errors that could be defined are shown as green double arrows. Errors were quantified by calculating the absolute error (AE) along each radial line, and the root mean square error (RMSE) was calculated across the log-transformed AEs and then transformed back to millimeter units.

### 4.3. Results

Table 1 gives the mean RMSE and the standard deviation of the RMSEs calculated across all images for each speaker. These data may be compared with the manual tracing results as well as with a baseline result, which was determined by guessing that the tongue contours in all images of a sequence were identical to the mean manual tracing of the first image of the sequence. Inspection of the data reveals that AutoTrace3.5, EdgeTrak, and TongueTrack have mean RMSE values smaller than the mean RMSE of the baseline. In contrast, AutoTrace3, which used mismatched training and test data, has larger mean RMSEs than the baseline. This pattern holds for each speaker except speaker M2, for which EdgeTrak also has a larger mean RMSE than the baseline.

In all cases, the unbiased standard deviations of the manual tracings are smaller than the RMSEs of the automatic tracings, with the highly matched training of AutoTrace3.5 consistently producing the smallest RMSEs. The 'out of the box' performances of EdgeTrak and TongueTrack were generally equally good on average.

In order to obtain a more detailed picture of the types and magnitude of errors produced by the automatic tracing programs, the AEs were examined qualitatively in a series of graphs plotting AEs as a function of frame and radial line. The radial lines were indexed from 1 to 41, with index 1 corresponding to the most posterior radial line. Figure 2 presents these data, along with the data for the baseline comparison. Also included are the AE data for one of the tracers (author SML). The magnitude of the AE is indicated by the color scale, with hotter colors corresponding to larger AEs.

In general, the results shown in Figure 2 reflect the findings from Table 1: the errors of AutoTrace3.5, EdgeTrak, and TongueTrack are smaller than or similar to the baseline tracing, whereas AutoTrace3 has higher errors in general. Errors associated with the manual tracer are generally small and uniform. Across speakers and algorithms, three basic types of errors are apparent:

**Error type** 1**.** The periodic pattern in the baseline tracking data is caused by tongue movement. For example, for speaker F1, the tongue is in neutral position for roughly the first 30 frames before the speaker begins to talk, and therefore the errors are small. After that, when the *'I owe you a yoyo'* sentence begins at around frame 30, the errors increase and show a periodic pattern as a function of the tongue movement between back and front vowel configurations. At around frame 145, when the sentence was finished, the errors are small again because the tongue has returned to its original neutral position. The same periodic error pattern is reflected in the tracings of AutoTrace3 and TongueTrack. In most cases where EdgeTrak has large errors, it is in the same periodic manner, although this occurs less frequently than for the other programs. An unexpected finding is that the pattern of 'out of the box' errors for TongueTrack is essentially identical with the errors from the baseline, even though the mean RMSEs for TongueTrack are smaller than those of the baseline. It appears that the default parameters of TongueTrack do not allow sufficient freedom for the tongue contour tracing to change its shape from one image to the next, at least for the relatively high frame rates used in this study.

**Error type** 2**.** In a few cases, sides of the tongue are consistently traced poorly while other parts are tracked well. This type of error occurs most notably with AutoTrace3 for speaker M2, for which angle indices smaller than 20 have large errors throughout the sequence.

**Error type** 3**.** Tracings frequently include only a relatively small region of the tongue (i.e. the posterior or anterior part of the tongue is not traced at all). This is typical of AutoTrace3.5 and AutoTrace3. For example, for speaker F2, almost all the points for angle indices smaller than 11 and larger than 29 are undefined for AutoTrace3. These kinds of errors are not reflected in the RMSE statistics given in Table 1 since AEs were not defined for errors of this kind (but see [22]).

## 5. Discussion

This study investigated errors associated with manual and automatic tongue contour tracings. With regard to manual tracings, previous studies involving pairs of experts found errors ranging from $0.49$ to $2.04$ $mm$, while the current study involving seven individuals with varying levels of training and experience found mean errors (unbiased standard deviations) ranging from $0.95$ to $2.11$ $mm$. These mean error measures were themselves distributed roughly log-normally, with standard deviation between $0.29$ and $0.32$ $mm$ after transforming back into linear units. After substracting the standard deviations from the means, the smallest standard deviation thus obtained is $0.95 - 0.29 = 0.66$ $mm$ for speaker F1, which is comparable

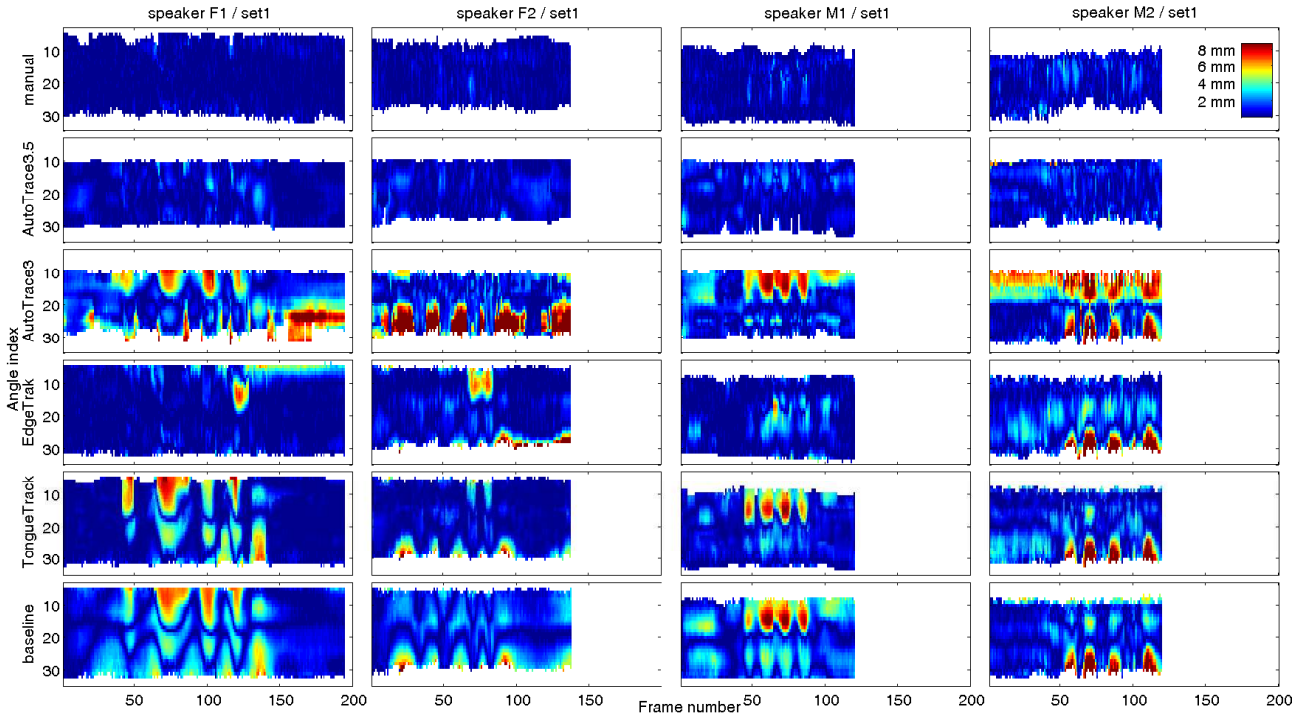


Figure 2: Error maps of tongue contour tracings. The color scale represents Absolute Error.

to the best previous results reported by [1, 9]. Likewise, after adding the standard deviations to the means, the largest standard deviation thus obtained is $2.11 + 0.32 = 2.43\ mm$ for speaker M2, which is comparable to the errors reported by [3, 10, 11]. Thus it appears that the degree of variability in manual tracings is likely more sensitive to image quality than to the expertise of the tracers, although expertise may also have an effect. Since the unbiased standard deviations were small, the mean tracings were accepted as the 'gold standard' for further investigation of errors associated with automatic tracings.

RMSEs for automatic tracings were consistently larger than the standard deviations of manual tracings. This indicates that, on average, manual tracers achieve greater agreement than automatic tracers are currently able to achieve, at least when default 'out of the box' parameter settings are used. Nonetheless, automatic tracings frequently returned very good results. The best performance was obtained by AutoTrace3.5, for which training and test sets were highly matched. The worst performance was obtained by AutoTrace3, for which training and test sets were mismatched with regard to speaker, but matched with regard to linguistic content. A baseline test in which each tongue contour was guessed to be identical to the first mean manual tracing of the sequence resulted in consistently smaller errors than AutoTrace3, but typically larger errors than EdgeTrak and TongueTrack. The fact that both EdgeTrak and TongueTrack sometimes returned errors approaching the magnitude of the baseline errors indicates these programs are not especially accurate for some data sets, especially those with lower image quality. The mean RMSEs reported here for EdgeTrak, TongueTrack, and AutoTrace3.5 (ranging from $1.15$ to $5.15\ mm$) are similar to the mean absolute errors previously reported by [10, 11, 3], which ranged from $0.54$ to $4\ mm$.

Three basic kinds of errors were identified. These included 1) periodic errors associated with a program's inability to trace quickly moving tongue contours, exemplified by AutoTrace3, TongueTrack, and the baseline test; 2) noisy errors which appear to be random and are associated with either poor image quality or instances in which the tongue contour, once 'lost' by the algorithm, is not immediately recovered, as exemplified by AutoTrace3; and 3) errors of omission in which parts of the tongue are simply not traced at all, exemplified by AutoTrace3.5 and AutoTrace3. On the whole, AutoTrace3.5 yielded the best results, with relatively homogeneous error distributions and small RMSEs, but it was also most strongly supported by highly matched training data.

## 6. Conclusions

The error analyses conducted in this study reveal that expertise is likely to have a secondary influence on manual tracing accuracy, while image quality has a primary influence. Nonetheless, manual tracings typically are in good agreement and close to the mean, which may be considered the 'gold standard'. The analyses also show that automatic tracings can achieve very good accuracy under appropriate conditions, but that errors falling into three major categories prevent automatic tracings from achieving accuracy rates as high as manual tracings, at least when default 'out of the box' parameter settings are used.

Our results might be useful for articulatory-acoustic investigations, e.g. for extending articulatory text-to-speech systems.

## 7. Acknowledgements

The first author was supported by a Fulbright scholarship and by the travel grant of the Hungarian Academy of Engineering.

We thank the 5 students who helped with the manual tongue contour tracing. We also thank the authors of the automatic contour tracing programs for their help and suggestions.

# 8. References


[1] M. Stone, B. Sonies, T. Shawker, G. Weiss, and L. Nadel, "Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system," *Journal of Phonetics*, vol. 11, pp. 207–218, 1983.

[2] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical Linguistics & Phonetics*, vol. 19, no. 6-7, pp. 455–501, Jan. 2005.

[3] J.-H. Sung, J. Berry, M. Cooper, G. Hahn-Powell, and D. Archangeli, "Testing AutoTrace: A Machine-learning Approach to Automated Tongue Contour Data Extraction," in *Ultrafest VI*, Edinburgh, UK, 2013, pp. 9–10.

[4] T. Bressmann, C.-L. Heng, and J. C. Irish, "Applications of 2D and 3D ultrasound imaging in speech-language pathology," *Journal of Speech-Language Pathology and Audiology*, vol. 29, no. 4, pp. 158–168, 2005.

[5] O. Rastadmehr, T. Bressmann, R. Smyth, and J. C. Irish, "Increased midsagittal tongue velocity as indication of articulatory compensation in patients with lateral partial glossectomies," *Head & Neck*, vol. 30, no. 6, pp. 718–726, Jun. 2008.

[6] T. Hueber, E.-l. Benaroya, B. Denby, and G. Chollet, "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 593–596.

[7] L. Ménard, J. Aubin, M. Thibeault, and G. Richard, "Measuring tongue shapes and positions with ultrasound imaging: a validation experiment using an articulatory model," *Folia Phoniatrica et Logopaedica*, vol. 64, no. 2, pp. 64–72, Jan. 2012.

[8] N. Zharkova, "A normative-speaker validation study of two indices developed to quantify tongue dorsum activity from midsagittal tongue shapes," *Clinical Linguistics & Phonetics*, vol. 27, no. 6-7, pp. 484–96, Jul. 2013.

[9] M. Stone, T. H. Shawker, T. L. Talbot, and a. H. Rich, "Cross-sectional tongue shape during the production of vowels." *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1586–1596, Apr. 1988.

[10] M. Li, C. Kambhamettu, and M. Stone, "Automatic contour tracking in ultrasound images," *Clinical Linguistics & Phonetics*, vol. 19, no. 6-7, pp. 545–554, Jan. 2005.

[11] L. Tang, T. Bressmann, and G. Hamarneh, "Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves," *Medical Image Analysis*, vol. 16, no. 8, pp. 1503–1520, Dec. 2012.

[12] National Institutes of Health USA, "Image-J [Computer program], Version 1.46a," 2014. [Online]. Available: http://imagej.nih.gov/ij

[13] Y. S. Akgul, C. Kambhamettu, and M. Stone, "Automatic extraction and tracking of the tongue contours," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 1035–1045, Oct. 1999.

[14] Y. Akgul, C. Kambhamettu, and M. Stone, "A task-specific contour tracker for ultrasound," in *Proc. IEEE MMBIA-2000*. Hilton Head Island, SC, USA: IEEE Comput. Soc, 2000, pp. 135–141.

[15] L. Tang and G. Hamarneh, "Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization," in *IEEE workshop on Mathematical Methods for Biomedical Image Analysis (IEEE MMBIA) in conjunction with the IEEE Conference on Computer Vision and Pattern Recognition (IEEE CVPR)*, San Francisco, CA, USA, 2010, pp. 154–161.

[16] L. Tang, G. Hamarneh, and T. Bressmann, "A Machine Learning Approach to Tongue Motion Analysis in 2D Ultrasound Image Sequences," in *MICCAI MLMI*, vol. 7009, Toronto, Canada, 2011, pp. 151–158.

[17] I. Fasel and J. Berry, "Deep Belief Networks for Real-Time Extraction of Tongue Contours from Ultrasound During Speech," in *Proc. ICPR*, Istanbul, Turkey, Aug. 2010, pp. 1493–1496.

[18] J. Berry and I. Fasel, "Dynamics of tongue gestures extracted automatically from ultrasound," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 557–560.

[19] J. Berry, I. Fasel, L. Fadiga, and D. Archangeli, "Training Deep Nets with Imbalanced and Unlabeled Data," in *Proc. Interspeech*, Portland, OR, USA, 2012, pp. 1756–1759.

[20] G. V. Hahn-powell, D. Archangeli, J. Berry, and I. Fasel, "AutoTrace: An automatic system for tracing tongue contours," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, p. 2104, Oct. 2014.

[21] "Medical Image Processing Toolbox [Computer Program], Version 1.0," 2014. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/41594-medical-image-processing-toolbox

[22] T. G. Csapó and D. Csopor, "Ultrahangos nyelvkontúr követés automatikusan: a mély neuronhálókon alapuló AutoTrace eljárás vizsgálata [Automatic tongue contour tracking: investigation of the deep neural network based AutoTrace method] (in Hungarian)," *Beszédkutatás 2015 [Speech Research 2015]*, pp. 177–187, 2015.