Guided Learning of Pronunciation by Visualizing Tongue Articulation in Ultrasound Image Sequences

M. Hamed Mozaffari, Shenyong Guan, Shuangyue Wen, Nan Wang, Won-Sook Lee
School of Electrical Engineering and Computer Science (EECS) University of Ottawa Ottawa, Canada
{mmoza102, sguan044, swen046, nwang085, wslee}@uottawa. ca

***Abstract*—Ultrasound has been used as one of the primary technologies utilized widely for clinical diagnosis due to its affordability, non-invasive characteristic, portability, and its fast performance in acquisition. Recently, it started to be used as a visual feedback method for tongue articulation, thanks to its capacity of real-time visualization and video capture of underlying structures inside the mouth. When an Ultrasound transducer is placed along the mid-line under a chin, it shows the tongue motion in sagittal view while speaking. As it is still quite difficult to understand the structure in ultrasound images, we proposed a guided learning system for pronunciation by visualizing tongue articulation in Ultrasound image sequences. Video image registration technique has been employed to project sagittal section of tongue back to the corresponding position on the subject head. The proposed system targets speech therapy and foreign language pronunciation lessons. Two main technology components are (i) Ultrasound tongue image segmentation and tracking (ii) registration of Ultrasound image sequences on video of a subject during the speech. Our experiments on Chinese English learners revealed that the proposed system is capable of providing the beneficial improvement on English pronunciation.**

***Index Terms*—ultrasound tongue imaging, real-time video visualization, linguistic guided learning, video composition, Ultrasound image segmentation and registration**

I. Introduction

One natural form of communication is speech. Wrong articulation in pronunciation causes misunderstanding. To pronounce a complete word or sentence, a complicated combination of phonemes needs to be articulated correctly. Fig. 1 illustrates the approximate position of the regions which tongue can reach when speaking vowels at the same time the vertical and horizontal coordinates of tongue differs during speech. For the case of consonants, tongue shape is even more complicated than that of vowels.

Traditionally, a language learner learns to pronounce a new word just by hearing. New researches have revealed that visual feedback techniques, such as ultrasound imaging, can help people to learn new languages with higher efficiency [1]. Ultrasound is one of the most prevailing imaging modality in many clinical applications, thanks to real-time system performance, its affordable cost, non-invasive characteristic, and its portability.

When considering tongue articulation and pathology, Ultrasound has been used for tongue visualization [3], tongue dorsum tracking [2], and many educational investigations [4]. Moreover, these technologies have assisted speech therapist for the rehabilitation of many speech disorders. Survey on possible application of utilizing ultrasound for language learning can be found in [5], [6].

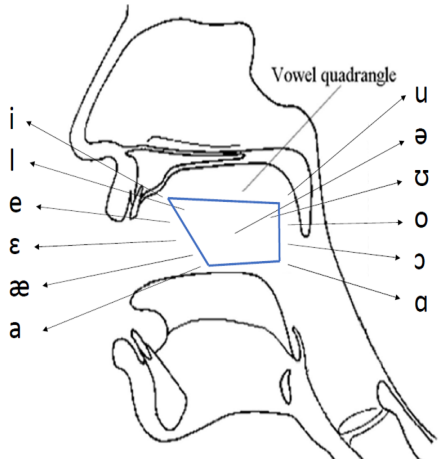

Fig. 1. The approximate position of tongue when producing vowel (figure reproduced in credit of Raymond Hickey [8]). Sounds varies depending on tongue's position and shape which are not visible from outside

Our study aims to address this pronunciation learning issue by providing visual feedback of tongue surface in Ultrasound images. In this process, visualization of both tongue movements of learner as well as a native speaker can be brought to the learner. Then, the learner can see the differences of two tongue shape, thus this visual feedback helps him/her to modify the tongue movement and at the same time speed up the process of learning.

We proposed and deployed a video registration method for real-time visualization of tongue surface, captured from Ultrasound video frames, and the other from an RGB video of a speaker during speech. The tongue dorsum is segmented in the sagittal view from Ultrasound video and the Region of Interest (ROI) of tongue is overlaid on profile of a subject.

We have done many experiments for a case of a language learner. The learner uses this system to see her/his tongue in real-time as a visual feedback as well as to compare it with tongue of a native speaker which was captured in advance. The experimental results revealed the strength of this method in teaching and learning of some English words which Chinese people typically have difficulty to pronounce correctly.

978-1-5386-4618-2/18/$31.00 ©2018 IEEE

II. LITERATURE REVIEW

A. Visual feedback based methods

Current research using ultrasound imaging as feedback for studying tongue movements, mainly is related to help people who aim to study a second language [21], [22], as well as part of a therapy in treating speech-related diseases [19], [20]. The importance of visual feedback is significant for human to regulate their behavior [23], [24], which is also known as motor learning principle [20]. With proper visual feedback, the human subject can learn the moving patterns and postures with higher efficiency, which could be a painstaking process that took up a lot of time without such feedback. It also grants the teacher a direct way of evaluating students [23], enables a better understanding of performance of the subject.

B. Speech therapy, phonetics study, and speech training

Ultrasound technology has been utilized in the last three decades for speech therapy and developed for treatment of English lingual stops, vowels, and sibilants and severe hearing impairment, residual speech impairment, or accented speech [1].

For phonetic and speech training, a sagittal view of ultrasound imaging is widely adopted, as it displays relative back, height, and the slope of various regions of the tongue [19]. A human subject makes a speech consisting of specific phonemes. Then, an ultrasound video showing tongue movement corresponding to those phonemes is displayed on a screen. With such scheme, it reported an improved performance for treating speech sound disorder [19], [20], and also aiding speech production and perception [21], [22].

C. Tongue database

Seeing Speech [16] is a collaborated project at six Scottish Universities, which recorded ultrasound tongue imaging (UTI) videos, MRI videos of all the vowels and consonants, even some other symbols like voiceless labial-velar fricative or approximation. A set of corresponding animations showing articulation of phonemes is also made based on the information from UTI and MRI. This is one of the most complete database regarding the topic that can be found online.

III. GUIDED LEANING SYSTEM

In this guided learning system, we proposed a platform for a language learner getting visual feedback of his/her tongue ultrasound images while being able to check a native speakers example.

A. Predefined words with difficulty by non-native speakers

We selected two kinds of words: pairs and single words, by conducting a survey among a group of non-native students about the words which they have difficulties to pronounce. The result was shown in the Table I.

TABLE I
A SET OF WORDS WITH DIFFICULTY FOR CHINESE LEARNERS

Pairs	Korea Career	Stuff Staff	Tone Tune	Word World	Girl Grow	Virtual Control	Pool Pull
Single word	Little	Studio	Sugar				

Fig. 2. An RGB video of a speaker reading predefined words. A RGB video and an ultrasound imaging video are simultaneously recorded

B. Capturing Ultrasound imaging and a video

We provided a system to record pronunciation of a speaker reading some predefined words. We recorded using both ultrasound machine and a video camera (we call it as RGB video) as shown in 2. The Fig. 2 shows the profile view of the speakers head in an RGB video and the transducer along the mid-line under the chin captures the sagittal view of his tongue being displayed on the Ultrasound machine monitor.

C. Two video imaging registration to overlay Ultrasound video on top of a RGB video

Our ultrasound system is from Ultrasonix Tablet Co. and it has a general purpose linear transducer. Our presets of ultrasound device were defined by a tongue expert in advance for tongue image acquisition with focal length setting only for tongue surface area (around 7 to 9 cm) with frame rate of 25fps and frequency of 6.6MHz. Our RGB video camera is on mode of 1080p at 30 fps.

The speaker reads each word in a specific time order with a time gap between each consecutive words, meanwhile two videos are captured in real-time, one from the Ultrasound machine, the other from the camera where both of them are connected to one computer. Data recording is synced between the camera and the Ultrasound imaging system using the time (with resolution of second) on each frame.

In order to realize real-time video composition registration, we overlaid ultrasound video on the top of the video of speaker by finding correct transform parameters. We asked the native

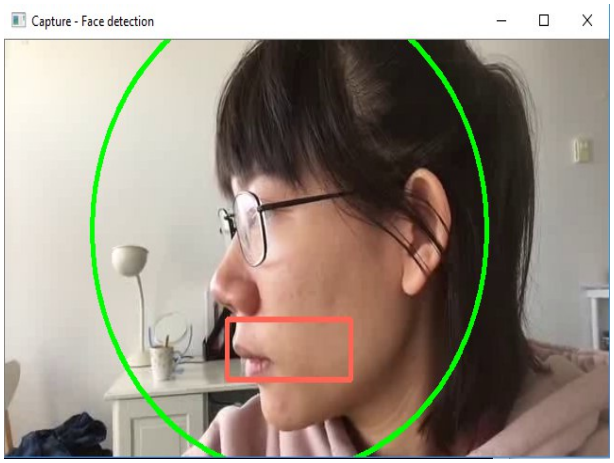

Fig. 3. The profile of a speakers face where two detected areas are shown in green and red for face and mouth respectively. They are detected and being tracked. The speakers facial feature are used for registration with Ultrasound imaging

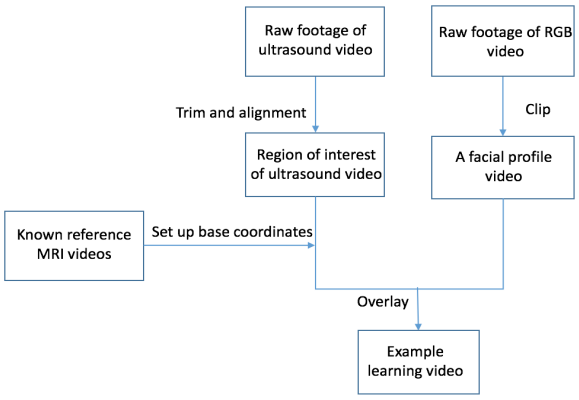

Fig. 4. An RGB video of a speaker reading predefined words. Meanwhile, the video and the ultrasound data are simultaneously recorded

speaker to pronounce the R sound firstly, then we checked the corresponding Ultrasound video to set up the base coordinates. Our designed steps are as follows:

- Face detection, tracking, and recognition using Haar cascades method for a learners video - a video capturing tool starts streaming the video from the camera and starts face detection. When a mouth is found, then his/her tongue area is automatically defined. Then the ultrasound video stream is automatically overlaid on the chin according to the tracking box as shown in Fig. 3. We use OpenCV library [9], [10] and Haar cascades method for video process.
- Finding the region of interest in ultrasound video stream and cropping the region.
- Superimposing the segmented area on the video stream.
- Visualization of pre-processed video of a native speakers data in parallel for the learner.

The flowchart is shown in Fig. 4. Our region of interest from ultrasound video stream is shown in Fig. 5 and the profile of the speakers head and ROI in RGB video data is illustrated in Fig. 3.

As the final display for learning system, the native speakers video will be shown on the monitor in parallel beside the real-time stream of the learners video as shown on the right and left respectively in Fig. 6. For the sake of better illustration, we use a big size monitor on the computer that enables subject learner to see her/his tongue on video with ultrasound overlaid image as well as the correct pronunciation of the words from the native speaker. In this procedure, the learner reads each word and she does not only hear ground truth sound from a native speaker but also see her tongue movement with the native subject tongue movement.
We used some color filters to enhance the Ultrasound data, to a certain extent that the tongue surface could be better seen by using after effect with Premiere software.

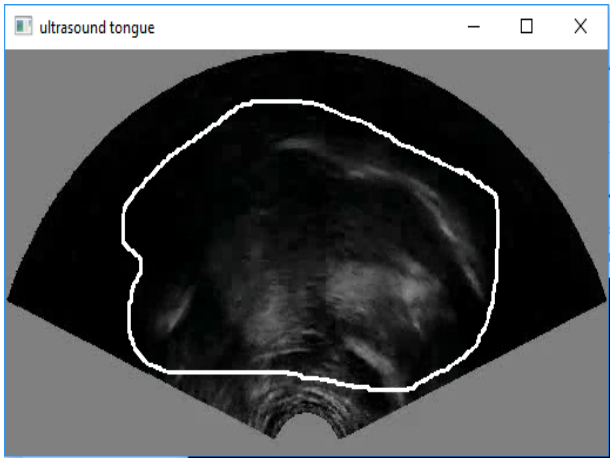

Fig. 5. Raw image showing sagittal view of tongue dorsum from ultrasound device and region of interest

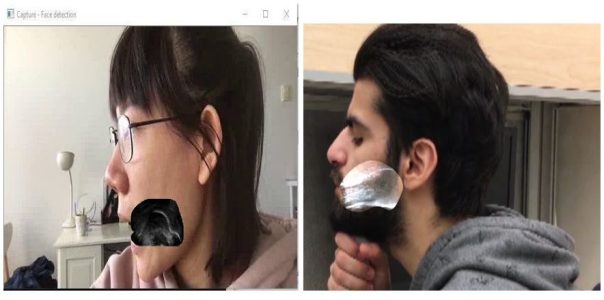

Fig. 6. Registration result between ultrasound imaging and an RGB imaging. The orientation, scale and position are correctly obtained to match two different sources of a speaker. The white area is the ultrasound tongue video with after effect. We show two sets of registered video, one by a learner (left) and the other by a native speaker (right).

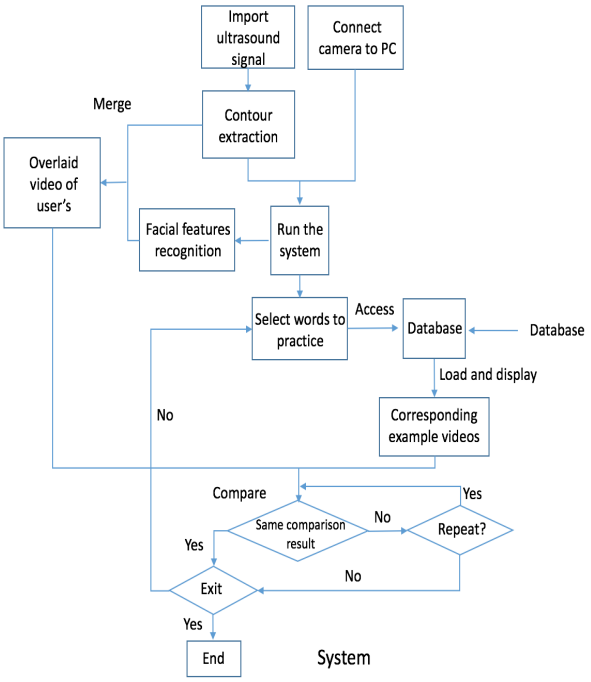

Fig. 7. Schematic of our proposed learning system.

D. Offline and online data capturing

We created two phases of data collection and capture. Phase I) offline data recording for an English native speaker and Phase II) online data recording for non-native English speakers for learning.

The first phase is pre-processed offline recording to prepare a database of a guiding pronunciation. We captured reading of predefined words by a native speaker in a list for both Ultrasound imaging method and RGB video. Then, after registration of two videos in one space, the database is made where each video is tagged with a word or a pair of word.

In the second phase, a non-native speaker selects any word in the list which she/he wants to practice. The system is designed to recognize which word she/he is saying and searches the database to show a corresponding word video from the native speaker. By comparing the tongue movement of the example learning video and the real-time video, the learner can see and mimic differences between two videos. Furthermore, she/he can learn how to pronounce the word correctly by adjusting her/his own tongue movement. After completing the above steps, the learner can either repeat learning the same word or choose another word to practice. The schematic of our proposed learning system is shown in Fig. 7.

IV. Tongue Contour Extraction and Tracking

Localization and interpretation of tongue gestures in noisy Ultrasound images is a challenging task for non-expert users. Therefore, illustrating a curve on top of a tongue dorsum instead of using pure Ultrasound images can significantly improve learning rate.

Active contour models can be considered as the most prevailing technique for tongue contour extraction and tracking during the last decades. In this method, one curve should be drawn for commencing. Then, the initialized curve updates its location toward the tongue as the brightest area in the image. In spite of effective, robust, and successful results of active contour model in many studies, due to its slow performance, dependency to initialization step, and many computational calculations for each frame such as gradient information, it cannot be used in fully automatic real-time applications.

In the recent years, machine learning techniques, especially deep neural networks [13] have been making major advances in solving many problems that have not been solved for many years. Convolutional Neural Networks (CNNs) are the most recent model of deep learning models which presents more robust and efficient than their previous counterparts.

In this study, we applied a modified version of a recent popular convolutional network for biomedical applications, U-net [15], for the problem of tongue segmentation. We proposed a simplified version of U-net (called as sU-net) through decreasing network architecture size in terms of the number of layers (14 convolutional layers in total instead of 23 layers by omitting one layer in each step of encoding-decoding process) and the number of filters (size of image sequences in our database is much less than images in original U-net).

Fig. 8 shows some snapshots of the input Ultrasound video frames and output of the tongue dorsum tracking resulted by our sU-net system. We applied our method on a standard database downloaded from [16] whereas two experts annotated data using our customized designed annotation software. Using Adam optimization algorithm on dice-coefficient loss function [17], we trained sU-net on the database (20/80 split ratio for testing and training). From the results, we can clearly assert that segmentation and extraction outcomes are near to human expert results. In order to validate the results, we calculated mean squared distance (MSD) [18] as a measure of error between ground truth and projected results by sU-net. The MSD value in terms of pixels is 1.43 pixels for sU-net, with a conversion rate of 1 px = 0.638 mm, and the average is 0.91 mm, which gives superior performance compared to the state of the art methods [25].

V. Conclusion

Our guided learning of pronunciation system uses Ultrasound imaging technology to make the movement of the tongue in the mouth visible. To increase the level of understanding and interpretation of the Ultrasound images, we used registration method with two video inputs, one from an Ultrasound machinethe other from a camera. After finding facial features as well as Ultrasound imaging tongue localization and segmentation, we superimposed an Ultrasound tongue on a profile of a speaker with correct scaling, position and orientation. The overlaid video shows that a language learner can understand tongue movement process of an English native speaker as well as check his/her own tongue comparing with

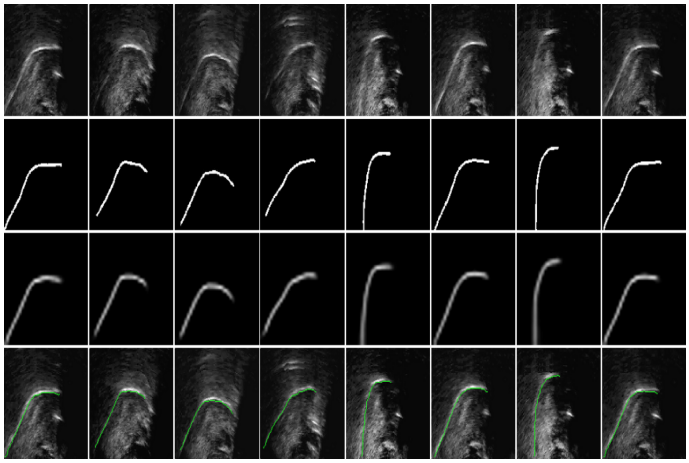

Fig. 8. Results of Ultrasound video segmentation. First row: Some randomly selected raw Ultrasound frames (cropped to be 128 × 128 in predefined position), Second row: annotation of tongue dorsum for each corresponding frame to the first row (128 × 128) used for validation, Third row: Predicted tongue dorsum segmentation (34 × 34) resulted from testing, Fourth row: Scaled contours in green, extracted from the third row images, are superimposed on top of the original frames of the first row.

the native speaker performance. The visualization of tongue gives the learner a direct and natural feedback, enables them to make the pointed adjustment, grants them an efficient and accurate language learning experience.

From our experimental studies, we understood that having this real-time system can help people learn difficult words easier and by practice, one learner can mimic those words similar to a native speaker. This study provides only a preliminary experiment for this problem and clearly with many difficulties still in the way. As for future studies, this system can be improved by testing more data from different Ultrasound devices. Many new ideas for images segmentation can be applied for this problem [30] and improve the final result even more by using heuristic optimization techniques [29]. Using a tongue contour segmentation, a small Ultrasound transducer which can connect to the computer through the USB port, and 3D visualization of tongue model would impact the speech industry.

REFERENCES

[1] Bernhardt, Barbara, et al. "Ultrasound in speech therapy with adolescents and adults." Clinical Linguistics & Phonetics 19.6-7 (2005): 605-617.

[2] Akgul, Yusuf Sinan, Chandra Kambhamettu, and Maureen Stone. "Automatic extraction and tracking of the tongue contours." IEEE Transactions on Medical Imaging 18.10 (1999): 1035-1045.

[3] Xu, Kele, et al. "Contour-based 3D tongue motion visualization using ultrasound image sequences." Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[4] Ihnatsenka, Barys, and Andr Pierre Boezaart. "Ultrasound: Basic understanding and learning the language." International journal of shoulder surgery 4.3 (2010): 55.

[5] McLaughlin, Barry, and Michael Harrington. "Second-language acquisition." Annual review of applied linguistics 10 (1989): 122-134.

[6] Abel, Jennifer, et al. "Ultrasound-enhanced multimodal approaches to pronunciation teaching and learning." Canadian Acoustics 43.3 (2015).

[7] Ilie, Mihai Daniel, Cristian Negrescu, and Dumitru Stanomir. "An efficient parametric model for real-time 3D tongue skeletal animation." Communications (COMM), 2012 9th International Conference on. IEEE, 2012.

[8] R. T. Sataloff, The Human Voice How the voice works was largely unknown until are now improving the care and treatment of the voice, vol. 267, no. December, pp. 108115, 1992.

[9] Bradski, G., and A. Kaehler. "Projection and 3D vision." Learning OpenCV (2008): 405.

[10] Li, Jianguo, and Yimin Zhang. "Learning surf cascade for fast and accurate object detection." Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013.

[11] Mahendru, Harish Chander. "Quick review of human speech production mechanism." International Journal of Engineering Research and Development 9.10 (2014): 48-54.

[12] Kass, Michael, Andrew Witkin, and Demetri Terzopoulos. "Snakes: Active contour models." Proc. 1st Int. Conf. on Computer Vision. Vol. 259. 1987.

[13] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.

[14] Jaumard-Hakoun, Aurore, et al. "Tongue contour extraction from ultrasound images based on deep neural network." arXiv preprint arXiv:1605.05912 (2016).

[15] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[16] Lawson, Eleanor, et al. "Seeing Speech: an articulatory web resource for the study of phonetics [website]." (2015).

[17] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer, Boston, MA, 2005. 321-352.

[18] Fasel, Ian, and Jeff Berry. "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech." Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010.

[19] Bernhardt, May B., et al. "Ultrasound as visual feedback in speech habilitation: Exploring consultative use in rural British Columbia, Canada." Clinical Linguistics & Phonetics 22.2 (2008): 149-162.

[20] Preston, Jonathan L., et al. "Ultrasound visual feedback treatment and practice variability for residual speech sound errors." Journal of Speech, Language, and Hearing Research 57.6 (2014): 2102-2115.

[21] Wilson, Ian, et al. "Ultrasound technology and second language acquisition research." Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006). 2006.

[22] Abel, Jennifer, et al. "Ultrasound-enhanced multimodal approaches to pronunciation teaching and learning." Canadian Acoustics 43.3 (2015).

[23] Mora, Javier, Won-Sook Lee, and Gilles Comeau. "3D visual feedback in learning of piano posture." International Conference on Technologies for E-Learning and Digital Entertainment. Springer, Berlin, Heidelberg, 2007.

[24] Mora, Javier, et al. "Assisted piano pedagogy through 3d visualization of piano playing." Haptic Audio Visual Environments and their Applications, 2006. HAVE 2006. IEEE International Workshop on. IEEE, 2006.

[25] Xu, Kele, et al. "Contour-based 3D tongue motion visualization using ultrasound image sequences." Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[26] Mozaffari, Mohammad Hamed, and WonSook Lee. "3D Ultrasound image segmentation: A Survey." arXiv preprint arXiv:1611.09811 (2016).

[27] Mozaffari, Mohammad Hamed, and Won-Sook Lee. "Freehand 3-D Ultrasound Imaging: A Systematic Review." Ultrasound in Medicine and Biology 43.10 (2017): 2099-2124.

[28] Akgul, Yusuf Sinan, Chandra Kambhamettu, and Maureen Stone. "Extraction and tracking of the tongue surface from ultrasound image sequences. " Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on. IEEE, 1998.

[29] Mozaffari, Mohammad Hamed, Hamed Abdy, and Seyed Hamid Zahiri. "IPO: an inclined planes system optimization algorithm." Computing and Informatics 35.1 (2016): 222-240.

[30] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2018): 834-848.