Real-time Automatic Tongue Contour Tracking in Ultrasound Video for Guided Pronunciation Training

M. Hamed Mozaffari, Shuangyue Wen, Nan Wang and WonSook Lee
School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario, Canada

Keywords: Image Processing with Deep Learning, Ultrasound for Second Language Training, Ultrasound Video Tongue Contour Extraction and Tracking, Convolutional Neural Network, Augmented Reality for Pronunciation Training.

Abstract: Ultrasound technology is safe, relatively affordable, and capable of real-time performance. Recently, it has been employed to visualize tongue function for second language education, where visual feedback of tongue motion complements conventional audio feedback. It requires expertise for non-expert users to recognize tongue shape in noisy and low-contrast ultrasound images. To alleviate this problem, tongue dorsum can be tracked and visualized automatically. However, the rapidity and complexity of tongue gestures as well as ultrasound low-quality images have made it a challenging task for real-time applications. The progress of deep convolutional neural networks has been successfully exploited in various computer vision applications such that it provides a promising alternative for real-time automatic tongue contour tracking in ultrasound video. In this paper, a guided language training system is proposed which benefits from our automatic segmentation approach to highlight tongue contour region on ultrasound images and superimposing them on face profile of a language learner for better tongue localization. Assessments of the system revealed its flexibility and efficiency for training pronunciation of difficult words via tongue function visualization. Moreover, our tongue tracking technique demonstrates that it exceeds other methods in terms of performance and accuracy.

1 INTRODUCTION

Communicative performance, self-confidence, and social interaction of a speaker during a speech depends on many elements. Correct pronunciation and articulation of phonemes, words, and sentences are two of those factors. Importance and challenging part of this communication skill is even more obvious for many language learners especially when they cannot pronounce words difficult and rapid tongue movements.

During articulation of words and sentences, tongue gestures are of great interest as an aid in Second language (L2) pronunciation learning and rehabilitation (Chen et al., 2018). When at rest, the tongue displays an unremarkable gross morphology, but dynamically it is a highly mobile, deformable, and precise organ, with rapid movements especially over its tip. Automatic tracking of tongue movements in recording ultrasound video of a long speech is considered to be very difficult due to the speckle noise in each frame, acoustic artifacts such as shadowing and mirroring, and low signal-to-noise ratio (Stone, 2005). Interpretation of ultrasound tongue frames might become even harder when tongue movements are fast, and the ultrasound frame rate is relatively low. This often results in missing parts in the observed contour (Loosvelt et al., 2014).

Recent studies have revealed that visual feedback techniques such as ultrasound imaging can assist individuals to acquire new language pronunciation skills with higher efficiency (Bernhardt et al., 2005; Wilson et al., 2006; Gick et al., 2008; Abel et al., 2015). Ultrasound systems are fast, safe, portable, relatively inexpensive, and capable of real-time imaging. These capabilities allow researchers to indirectly study subtle and swift movements of the tongue during speech production in different applications (Denby et al., 2010; Preston et al., 2014; Geddes and Sakalidis, 2016).

Mid-sagittal view of ultrasound imaging is usually adopted for illustration of tongue, as it displays relative back, height, and the slope of the tongue (Bernhardt et al., 2008). However,

Mozaffari, M., Wen, S., Wang, N. and Lee, W.
Real-time Automatic Tongue Contour Tracking in Ultrasound Video for Guided Pronunciation Training.
DOI: 10.5220/0007523503020309
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 302-309
ISBN: 978-989-758-354-4

localization and interpretation of tongue gestures in ultrasound images is not an easy task for non-expert users (see Figure 1, for the first time, one cannot recognize exact location of the tongue in ultrasound images). Therefore, highlighting tongue dorsum on ultrasound data, in form of a curved contour which is usually defined under the brightest and longest continues region (Zharkova, 2013; Lee et al., 2015), can significantly assist language learners to recognize the tongue shapes.

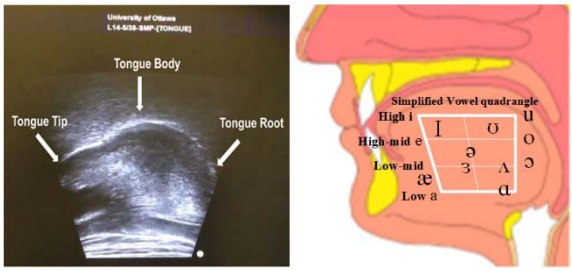

Figure 1: The approximate position of the tongue when producing a vowel. Sounds vary depending on tongue's position and shape which are not visible from outside. Tongue surface can be seen as a bright region on ultrasound image on the left.

So far, few language training systems have been implemented utilizing ultrasound imaging (Hoopingarner, 2005; Gick et al., 2008). Language learners should first be involved in a pre-training stage which helps them to comprehend the location of the tongue in ultrasound videos data.

Our study aim is to provide a system to facilitate second language pronunciation training by depicting tongue function in real-time and recorded videos. A face tracking method is utilized to find optimum position of the tongue on face profile then they are registered to create overplayed videos. Our language training system attempts to alleviate the problem of tongue localization in ultrasound data for non-expert users utilizing our tongue tracking approach that visualizes contour of the tongue, superimposed on ultrasound video frames.

2 LITERATURE REVIEW

A variety of techniques have been tested for tongue contour tracking in ultrasound images such as active contour models (Ghrenassia et al., 2014; Laporte and Ménard, 2015; Xu et al., 2016), graph-based technique (Tang and Hamarneh, 2010), machine learning-based methods (Tang et al., 2012; Fabre et al., 2015), and many more (Laporte and Ménard, 2018). Manual labeling is essential for at least initialization in those researches (Laporte and Ménard, 2018) such that tongue tracking in real-time is impossible also with using famous software packages like EdgeTrak.

Up to now, research on deep learning methods has stirred a great deal of attention, and It shows that deep learning algorithms, particularly convolutional neural network (CNN) (Xu et al., 2017), are powerful enough for solving many problems in pattern recognition and data mining, including ultrasound tongue contour tracking (Laporte and Ménard, 2018). In similar studies (Fasel and Berry, 2010; Berry and Fasel, 2011; Berry, 2012; Csapo and Lulich, 2015; Jaumard-Hakoun et al., 2016), tongue contour was extracted automatically using deep belief networks (DBNs) and deep auto-encoder (Ji et al., 2017).

The accuracy of deep learning methods is highly related to the size of training dataset and the complexity of the deep network model. Hence, there is always a trade-off between the number of training samples, which is a big issue in many applications such as in medicine (Ronneberger et al., 2015; Litjens et al., 2017), and the number of parameters in the network, which it requires more computing and memory units. Results of deep learning techniques like U-net for medical image segmentation illustrate acceptable accuracy. However, due to the deep architecture with many layers, computational resources should be highly powerful in training and testing stages. For real-time applications such as ultrasound tongue contour tracking this performance, the issue is also more fundamental.

In this paper, by inspiration from a famous architecture, so-called "fully convolutional network" (Badrinarayanan et al., 2015; Long et al., 2015; Ronneberger et al., 2015), we propose a new simpler architecture for real-time video tongue contour tracking for using in our designed language training system. The efficiency of our language training system was tested by conducting experiments to teach pronunciation of some difficult words.

3 SYSTEM ARCHITECTURE AND METHODOLOGY

Main modules of our language pronunciation training system are illustrated in Figure 2. In both modules, the ultrasound data, contain tongue contour information extracted by our proposed tracking

method, are overlaid automatically on a video recorded from face profile. In general, two superimposed videos are played on a computer's screen for a language learner, one from real-time recording of him/his tongue and the other from an instructor's tongue during speech.

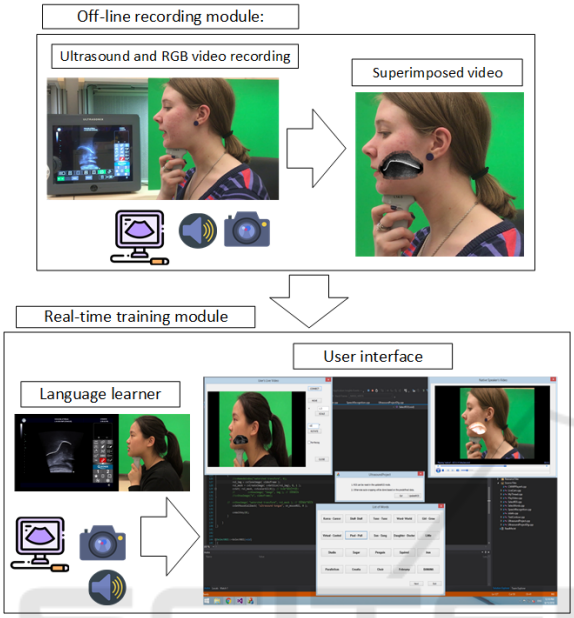

Figure 2: Schematic of our language training system. Language learner can see both off-line and real-time data on a computer screen. The off-line video is played with a small delay and language learner imitate that word. The learner comprehends the differences between two videos and tries to duplicate instructor's video.

3.1 Tongue Tracking and Extraction Module

Tongue contour tracking in ultrasound video is a unique problem such that according to our experiments, the diversity of tongue data is restricted to the flexibility and deformation states of the tongue muscle. We proposed a deep learning approach, employing CNN layers, capable to address the real-time performance issue due to its small architecture and training dataset (see Table 1).

In SegNet architecture (Badrinarayanan et al., 2015), two consecutive convolutional layers are for having a better receptive field, however, these extra layers in each stage of encoding and decoding apply a huge number of parameters to the network in inference stage. We found that omitting one convolutional layer in each stage of encoding and decoding, improves architecture performance in terms of speed though its accuracy is still comparable with the original model. In our architecture, we decreased many repeating convolutional and deconvolutional layers of SegNet as well as many activation functions by try and error experiments. We also added concatenation strategy from the U-net model (Ronneberger et al., 2015) for increasing accuracy of the predictions.

Table 1: Our deep learning architecture (Conv: convolution, Concate: concatenation).

Network Architecture	
Layers	Specification (Down-sampling)
Conv 1	32 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Pool 1	Max-pooling with stride: 2×2
Conv 2	64 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Pool 2	Max-pooling with stride: 2×2
Conv 3	128 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Pool 3	Max-pooling with stride: 2×2
Conv 4	256 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Pool 4	Max-pooling with stride: 2×2
Conv 5	512 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Layers	Specification (Up-sampling)
Un-pool 1	Max-un-pooling repeats the rows and columns by size 2
Concate 1	Concatenate conv 4 outputs and un-pool 1
Conv 6	256 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Un-pool 2	Max-un-pooling repeats the rows and columns by size 2
Concate 2	Concatenate conv 3 outputs and un-pool 2
Conv 7	128 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Un-pool 3	Max-un-pooling repeats the rows and columns by size 2
Concate 3	Concatenate conv 2 outputs and un-pool 3
Conv 8	64 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Un-pool 4	Max-un-pooling repeats the rows and columns by size 2
Concate 4	Concatenate conv 1 outputs and un-pool 4
Conv 9	32 (filters: 3×3, stride: 1×1, activation: ReLU, padding: “Valid”)
Conv 10	1 (filters: 1×1, stride: 1×1, activation: Sigmoid, padding: “Valid”)

The popular tongue tracking software, EdgeTrak, is capable of providing tongue contours for a limited number of frames with the requirement of manual initialization near to the tongue region and tuning it during the extraction process. In contrast, our proposed model is significantly fast, needs a small dataset for training, and it can automatically delineate tongue contours in real-time from long

ultrasound videos without any manual initialization or tuning.

As can be seen from Table 1, our proposed network consists of repeated 3×3 convolutions with no zero padding, each followed by a rectified linear unit (ReLU). A 2×2 max pooling operation with a stride of 2 is applied in each down-sampling step. In decoding path, from the lowest contracted layer to the output feature map, each layer consists of an up-sampling of the feature map followed by a 2×2 convolution that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the corresponding contracting path, and a 3×3 convolutions followed by a ReLU. Our model has only 10 convolutional layers (in contrast to 23 layers in U-net and 26 in SegNet). For the sake of fair comparison, we did not add data augmentation, a powerful strategy for increasing the size of training data, and batch normalization layers, a method for increasing the performance and stability.

The result of the proposed method is a segmented region on the ultrasound image, so we need to extract contours for comparison with other tongue contour tracking techniques. First, we binarize the segmented tongue contour image with a threshold value and then invoke the skeleton extraction method to create thin curves out of the binary image (see Figure 3).

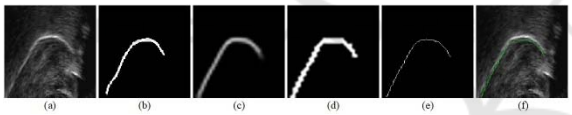

Figure 3: samples of images in segmentation and extraction process: (a) one original image (b) true label or mask (c) predicted area (d) converted binary image (e) contour skeleton of the white region of (d), (f) superimposed image of original and extracted contour (green curve).

3.2 Ultrasound Pronunciation Training System Module

In order to test the potential utility of our language training system in second language, a preliminary investigation was conducted with two Chinese students as participants. Three English native speakers were considered as instructors and trained to read predefined words. As Figure 2 shows, recording procedures is a two-fold process. Off-line recording module (see Figure 4) which a native instructor sits in front of the system who reads a list of predefined difficult words. In our experiment, list of difficult words (see Table 2) was produced by conducting a survey using questionnaires, among 20 Chinese students in University of Ottawa, asking about the words which they have more difficulties to pronounce in English.

Figure 4: An instructor is pronouncing a list of words and her face and voice are recorded by the camera as well as ultrasound video. Manually, cropped ultrasound video and the tongue contour extracted (green line) from that video are superimposed with the video from the camera.

In the second module, similarly, one student keeps an ultrasound probe under his/her chin and reads the list of predefined words to learn their pronunciations (see Figure 5). Ultrasound video of the tongue, RGB video from speaker's face profile, and speech sound are captured in both modules. During the training process, the learner observes instructor video for each word and try to imitate the instructor's tongue in ultrasound video which is shown at the same time on screen with a sort of delay. While the learner is reading the words, tongue contour is superimposed and illustrated automatically in real-time on ultrasound image sequences. For overlaying videos, Haar feature detection algorithm is applied to the RGB video to find the area around the lips of the speaker. Then, segmented ultrasound image sequences are registered on the face profile.

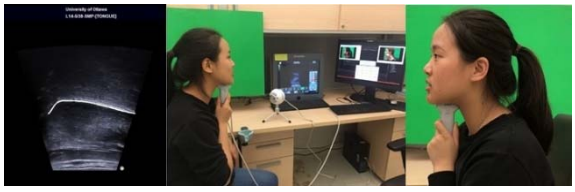

Figure 5: A learner is pronouncing a list of words after hearing the voice and watching the video from the instructor. The white color line is extracted from ultrasound data automatically.

Table 2: A set of English words and pairs with difficulty to pronounce for the Chinese language learner.

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

4 EXPERIMENTAL RESULTS

Due to the lack of similar language training system in the literature, we qualitatively compared our system with similar ideas (Hueber, 2013; Ouni, 2014; Abel et al., 2015). From our experimental results, our language training system is highly user-friendly due to the illustration of tongue contour instead of just depicting raw ultrasound image sequences. Real-time and automatic performance of our system allows linguistics to focus on training process instead of manual editing and interpretation of ultrasound images.

Our tracking method can recognize tongue dorsum in different positions and orientations due to translation invariant characteristic of CNN's. In order to register ultrasound video on face profile video automatically, a face detection algorithm from OpenCV library (Haar cascade) is employed which can find the approximate place of the tongue on the face. Thus, students could use the system with more flexibility than previous studies without using any fixtures for fixing their head and ultrasound probe position. Our system does not require any initialization steps, and due to the simplicity of our deep network architecture, GPU facility is not necessary.

We assessed the proposed deep learning model on a database comprises of 2190 images captured by our ultrasound machine (Tablet Ultrasonix with L14-5 linear probe) and 6631 images from the internet (Lawson et al., 2015). Each image has a size of (128 ×128), each of which comes with the same size corresponding annotated mask created by our annotation software package. The output segmented image has size (34×34). Output results are illustrated in the last row of Figure 6 where we randomly selected some frames for the sake of presentation.

The database was divided into a training set and a validation set with an 80/20 percentage split ratio, respectively. We deployed our method by using the Keras library (Chollet and others, 2015) and TensorFlow backend (Abadi et al., 2016). Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions was used with its default parameters (Kingma and Ba, 2014), in which a learning rate of 0.001, β_1 of 0.9, and β_2 of 0.999, as well as schedule decay of 0.05 after each epoch, were adopted to achieve a better convergence. Default parameter values from relevant publications (Badrinarayanan et al., 2015; Long et al., 2015; Ronneberger et al., 2015) have been used for other parameters such as the number of encoding and decoding layers, number of iterations, and batch sizes.

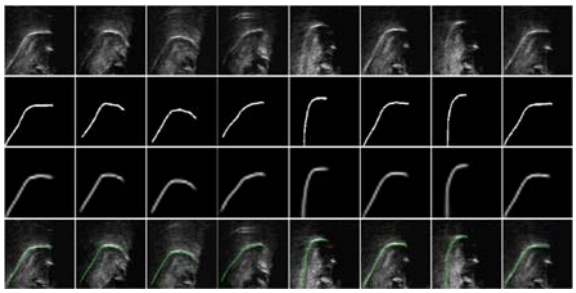

Figure 6: Results of applying the proposed image segmentation and extraction method on video data. First row: Some randomly selected raw ultrasound frames (128×128), Second row: Corresponding masks for each frame in the first row (128×128), Third row: Predicted masks (34×34), Fourth row: Extracted contours using skeleton method on a resized version of third raw images (128×128) and then moved to lower part of the contour region to superimpose with the original frame.

We used binary cross-entropy as the loss function. To evaluate the performance of the proposed method, we also calculated the dice coefficient. Prediction and the labeled data are compared in terms of MSD as defined in (Jaumard-Hakoun et al., 2016). Our experimental results are compared with the results of deep belief network (DBN) (Jaumard-Hakoun et al., 2016) which is the only previous method known to use deep learning in tongue image analysis. Table 3 shows the comparison for the Dice-coefficient validation error related to the number of epochs. In our system, the error declines to around 0.2 when the number of epochs is around 100 and then levels out. The DBN error remains around 0.4 even though the number of epochs increases. Results of our proposed model are shown in Figure 6 such that, for the sake of illustration, frames were selected randomly during the training stage. Images in the third row are predicted maps from our proposed model which are like the true labels in the second row.

Our proposed method outperformed other methods regarding MSD criteria as shown in Table 4. The MSD value in terms of pixels was 1.43 pixels for our proposed method, with a conversion of 1 px = 0.638 mm, giving an average MSD of 0.91 mm,

while the DBN model achieved 1.0 mm (1 px = 0.295 mm). For the active contour tracking method mentioned in (Li et al., 2005), the average MSD is 1.05 mm. It is important to mention that the two human experts participating in active contour tracking experiment produced two different annotation results having an average MSD of 0.73 mm (Li et al., 2005), which may thus be reasonably considered the ultimate minimum MSD of training based automated methods.

We ran our model on a Windows PC with a 7-Core CPU at 3.4GHz and 16GB of memory. In the testing stage, our approach provided the segmentation results of 175 frames in 2.343 seconds, which equals 74.7 fps. The testing performance goes down to 29.8 fps when we add tongue contour extraction. Previous publications have not discussed speed, possibly due to the nature of semi-automatic or manual work. Alternatively, they may have not achieved a high speed.

Table 3: Comparison of our method and previous work in Dice-coefficient validation error. Our method out-performs DBN when the number of epochs reaches 50 and levels off at 100 epochs at around half the error of the DBN system.

Number of epochs	5	50	100	250
Proposed method	0.446	0.243	0.212	0.212
DBN (Jaumard-Hakoun et al., 2016)	0.41	0.38	N/A	0.4

Table 4: The comparison of our model with others in term of average MSD. Our proposed model shows better accuracy than other state-of-the-art methods on tongue dorsum extraction.

Methods	**Ours**	**DBN**	**Active Contour**
Average MSD (mm)	0.91	1.0	1.05

5 CONCLUSION

While the potential benefits of articulatory feedback using ultrasound data for language pronunciation have long been acknowledged, until recently, with the advancement of deep learning techniques, real-time applications are feasible for easier guided language training. The pilot experiment of language learning system using ultrasound imaging outlined in the present paper shows that it is so much promising to add different facilities to ultrasound video in order to enhance the learning process.

In automatic tongue tracking section of our system, a trained deep convolutional neural network architecture has been applied to extract and illustrate tongue contours on real-time ultrasound videos. Automatic registration of ultrasound and RGB videos using Haar feature extraction algorithm, added flexibility to the language training system by omitting fixtures for the head and probe. The choice of network structure was selected and modified in the course of several experiments, decreasing the number of network layers and removing unnecessary steps for this application to speed up the process while keeping the high accuracy.

Extraction of tongue contours from delineated frames was successfully exploited using the skeleton technique. The proposed system does not need initialization or re-initialization (unlike EdgeTrak) for each frame and it is an end-to-end approach. The experimental results displayed the accuracy and speed of the proposed method for real-time automatic tongue tracking and we can assert that convolutional neural networks are superior to their counterparts such as deep belief networks. This kind of system shows its potential abilities for solving other ultrasound problems and being applied to other organs.

REFERENCES

Abadi, M. *et al.* (2016) 'TensorFlow: A System for Large-Scale Machine Learning.', in *OSDI*, pp. 265–283.

Abel, J. *et al.* (2015) 'Ultrasound-enhanced multimodal approaches to pronunciation teaching and learning.', *Canadian Acoustics*, 43(3).

Badrinarayanan, V., Kendall, A. and Cipolla, R. (2015) 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation', pp. 1–14. doi: 10.1109/TPAMI.2016.2644615.

Bernhardt, B. *et al.* (2005) 'Ultrasound in speech therapy with adolescents and adults', *Clinical Linguistics and Phonetics*, 19(6–7), pp. 605–617. doi: 10.1080/02699200500114028.

Bernhardt, M. B. *et al.* (2008) 'Ultrasound as visual feedback in speech habilitation: Exploring consultative use in rural British Columbia, Canada', *Clinical Linguistics & Phonetics*. Taylor & Francis, 22(2), pp. 149–162. doi: 10.1080/02699200701801225.

Berry, J. and Fasel, I. (2011) 'Dynamics of tongue gestures extracted automatically from ultrasound', in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 557–560.

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

Chen, S. *et al.* (2018) 'Direct, Near Real Time Animation of a 3D Tongue Model Using Non-Invasive Ultrasound Images', in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4994–4998. doi: 10.1109/ICASSP.2018.8462096.

Chollet, F. and others (2015) 'Keras: Deep learning library for theano and tensorflow', *URL: https://keras. io/k*, 7, p. 8.

Csapo, T. G. and Lulich, S. M. (2015) 'Error analysis of extracted tongue contours from 2D ultrasound images', *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015–Janua, pp. 2157–2161. doi: 10.1111/j.0956-7976.2004.00672.x.

Denby, B. *et al.* (2010) 'Silent speech interfaces', *Speech Communication*. Elsevier, 52(4), pp. 270–287.

Fabre, D. *et al.* (2015) 'Tongue tracking in ultrasound images using eigentongue decomposition and artificial neural networks', *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015–Janua(2), pp. 2410–2414.

Fasel, I. and Berry, J. (2010) 'Deep belief networks for real-time extraction of tongue contours from ultrasound during speech', in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 1493–1496. doi: 10.1109/ICPR.2010.369.

Geddes, D. T. and Sakalidis, V. S. (2016) 'Ultrasound imaging of breastfeeding—a window to the inside: Methodology, normal appearances, and application', *Journal of Human Lactation*. SAGE Publications Sage CA: Los Angeles, CA, 32(2), pp. 340–349.

Ghrenassia, S., Ménard, L. and Laporte, C. (2014) 'Interactive segmentation of tongue contours in ultrasound video sequences using quality maps', in *Medical Imaging 2014: Image Processing*, p. 903440.

Gick, B., Bernhardt, B., *et al.* (2008) 'Ultrasound imaging applications in second language acquisition', *Phonology and second language acquisition*. John Benjamins Amsterdam, 36, pp. 315–328.

Gick, B., Bernhardt, B. M., *et al.* (2008) 'Ultrasound imging applications in second language acquisition', *Phonology and Second Language Acquisition*, (June), pp. 309–322. doi: 10.1684/abc.2012.0768.

Hoopingarner, D. (2005) 'SECOND LANGUAGE SPEECH PERCEPTION AND PRODUCTION: ACQUISITION OF PHONOLOGICAL CONTRASTS IN JAPANESE', *Studies in Second Language Acquisition*. Cambridge University Press, 27(3), p. 494.

Hueber, T. (2013) 'Ultraspeech-player: intuitive visualization of ultrasound articulatory data for speech therapy and pronunciation training.', in *INTERSPEECH*, pp. 752–753.

Jaumard-Hakoun, A., Xu, K., Leboullenger, C., *et al.* (2016) 'An articulatory-based singing voice synthesis using tongue and lips imaging', in *ISCA Interspeech 2016*, pp. 1467–1471.

Jaumard-Hakoun, A., Xu, K., Roussel-Ragot, P., *et al.* (2016) 'Tongue contour extraction from ultrasound images based on deep neural network', *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*. Available at: http://arxiv.org/abs/1605.05912.

Ji, Y. *et al.* (2017) 'Updating the silent speech challenge benchmark with deep learning', *arXiv preprint arXiv:1709.06818*.

Kingma, D. P. and Ba, J. (2014) 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*.

Laporte, C. and Ménard, L. (2015) 'Robust tongue tracking in ultrasound images: a multi-hypothesis approach', in *Sixteenth Annual Conference of the International Speech Communication Association*.

Laporte, C. and Ménard, L. (2018) 'Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech', *Medical image analysis*. Elsevier, 44, pp. 98–114. doi: 10.1016/j.media.2017.12.003.

Lawson, E. *et al.* (2015) 'Seeing Speech: an articulatory web resource for the study of phonetics [website]'. University of Glasgow.

Lee, S. A. S., Wrench, A. and Sancibrian, S. (2015) 'How To Get Started With Ultrasound Technology for Treatment of Speech Sound Disorders', *SIG 5 Perspectives on Speech Science and Orofacial Disorders*. ASHA, 25(2), pp. 66–80.

Li, M., Kambhamettu, C. and Stone, M. (2005) 'Automatic contour tracking in ultrasound images', *Clinical Linguistics and Phonetics*, 19(6–7), pp. 545–554. doi: 10.1080/02699200500113616.

Litjens, G. *et al.* (2017) 'A survey on deep learning in medical image analysis', *Medical Image Analysis*, 42(1995), pp. 60–88. doi: 10.1016/j.media.2017.07.005.

Long, J., Shelhamer, E. and Darrell, T. (2015) 'Fully convolutional networks for semantic segmentation', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Loosvelt, M., Villard, P.-F. and Berger, M.-O. (2014) 'Using a biomechanical model for tongue tracking in ultrasound images', *Biomedical Simulation*, 8789, pp. 67–75. Available at: http://link.springer.com/chapter/10.1007/978-3-319-12057-7_8.

Ouni, S. (2014) 'Tongue control and its implication in pronunciation training', *Computer Assisted Language Learning*. Taylor & Francis, 27(5), pp. 439–453.

Preston, J. L. *et al.* (2014) 'Ultrasound visual feedback treatment and practice variability for residual speech sound errors', *Journal of Speech, Language, and Hearing Research*. ASHA, 57(6), pp. 2102–2115.

Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-net: Convolutional networks for biomedical image segmentation', in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241.

Stone, M. (2005) 'A guide to analysing tongue motion from ultrasound images', *Clinical Linguistics and Phonetics*, 19(6–7), pp. 455–501. doi: 10.1080/02699200500113558.

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. Elsevier B.V., 16(8), pp. 1503–1520. doi: 10.1016/j.media.2012.07.001.

Tang, L. and Hamarneh, G. (2010) 'Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization', *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pp. 154–161. doi: 10.1109/CVPRW.2010.5543597.

Wilson, I. *et al.* (2006) 'Ultrasound technology and second language acquisition research', in *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006)*, pp. 148–152. doi: doi.org/10.1002/hed.20772.

Xu, K. *et al.* (2016) 'Robust contour tracking in ultrasound tongue image sequences', *Clinical linguistics & phonetics*. Taylor & Francis, 30(3–5), pp. 313–327. doi: 10.3109/02699206.2015.1110714.

Xu, K. *et al.* (2017) 'Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images', *The Journal of the Acoustical Society of America*. ASA, 141(6), pp. EL531--EL537.

Zharkova, N. (2013) 'Using ultrasound to quantify tongue shape and movement characteristics', *The cleft palate-craniofacial journal*. SAGE Publications Sage CA: Los Angeles, CA, 50(1), pp. 76–81.