Tongue Contour Tracking Techniques in Ultrasound Images

Subjects: Engineering, Biomedical | Health Care Sciences & Services

Contributor: Khalid Al-hammuri , Fayez Gebali , Ilamparithi Thirumarai Chelvan , Awos Kanan

Lingual ultrasound imaging is essential in linguistic research and speech recognition. It has been used widely in different applications as visual feedback to enhance language learning for non-native speakers, study speech-related disorders and remediation, articulation research and analysis, swallowing study, tongue 3D modelling, and silent speech interface. Tongue tracking using machine learning-based techniques is superior to traditional techniques, considering the performance and algorithm generalization ability. Meanwhile, traditional techniques are helpful for implementing interactive image segmentation to extract valuable features during training and postprocessing.

tongue contour tracking | computer vision | lingual ultrasound

1. Introduction

Researchers consider the mean sum of distances (MSD) as the primary evaluation criterion for the quantitative analysis of tongue segmentation. MSD is the standard measure of tongue segmentation in research as it considers the variation of tongue length, and it is adopted widely in tongue segmentation publications.

Studying tongue movement during speech is essential to the understanding of human articulation. Different approaches are used to study speech; some rely on a single sensor [1][2][3][4][5], and others use hybrid techniques [6][7][8]. Due to medical imaging modalities advancement and impressive capabilities, linguistic researchers are relying on the medical ultrasound system to capture tongue motion during speech [9]. Ultrasound imaging is considered the most efficient methodology in terms of safety and portability. However, magnetic resonance imaging (MRI) has a better resolution, and it can provide more information about the soft tissues [10], vocal tract, and craniofacial structure [11][12]. MRI is used for real-time image acquisition [10][13][14] to visualize the vocal tract either in 2D or 3D orientation [15][16] and enhance the speech analysis. However, MRI is huge in size and very expensive compared to ultrasound. It requires a special arrangement and a long scanning time, making it impractical for most of the day-to-day uses of speech analysis to limit its application for particular research or clinical studies.

2. Traditional Image Analysis Techniques for Tongue Contour Tracking

Tongue tracking by ultrasound was addressed in early research by the cited works [17][18]. However, the process was manual and required a cautious user attention while handling the ultrasound transducer. To enhance the

transducer guidance, metal pellets were used as a strong reflector to identify few landmarks on the tongue surface. The landmarks were used as a reference to monitor tongue movement during swallowing by comparing the pellets placed on the tongue anterior and posterior segments to the hyoid bone reference at different stages of movement.

There are two main traditional methodologies used to segment the tongue: active contour model (snake algorithm)and shape consistency and graph-based tongue tracking models.

2.1. Active-Contour-Based Methodologies (Snake Algorithm)

To automate tongue contour tracking, many researchers have relied on the snake algorithm [19][20] as the base algorithm for most of the traditional techniques in tongue contour tracking. The snake algorithm is an active contour and energy-based method that adapts to get closer and closer to the object until reaching a certain threshold or energy constraints to fit the object boundary. The snake algorithm has been used widely in vision tasks such as the detection of lines, objects and subjective contours, and motion tracking. In the case of lingual ultrasound, the snake algorithm can be useful for interactively segmenting a tongue contour by applying certain user-imposed constraint forces to localize the tongue features of interest. Examples of the first attempts to use active contours for tongue tracking tasks were provided by [21][22][23], which were made by the same authors and improved consequently.

An adaptive snake algorithm was introduced by [21]. The authors collected 2D ultrasound images and used a head and transducer support system to stabilize the ultrasound transducer. In the first frame, a human expert selected a few candidates of the contour points to generate the initial tongue contour to initiate the snake algorithm. For the following frames, the researchers proposed an adaptive model that estimated an optimized contour that matched the tongue contour edges on each frame. Finally, the algorithm implemented a postprocessing technique to enhance and refine the extracted contours.

The cited work in [22] followed the same process as the work in [21] and extended the work using different constraints to test it in speech and swallowing applications. The authors in [22] showed an improvement in the model performance by minimizing the computational cost to make it more flexible for a variety of different tasks.

Similarly, the algorithm proposed by [23] required an initial input from an expert to delineate the tongue contour on the first image frame to ease the snake algorithm optimization of the energy constraints that enforced the detection of tongue contour edges in the desired region of interest. Subsequent video frames were processed by adapting the initial contour edges to match the tongue deformation. External and internal energy functions were suggested to optimize the tongue contour’s external edges and concavity, respectively. Although the methodology showed some success in tongue contour detection, its performance dropped drastically in the case of noisy images due to its sensitivity to speckle noise. Moreover, in the case of rapid tongue movements, the external energy function could fail to adapt the edges and match the tongue boundaries’ deformation to the new position at the next frame. This, unfortunately, limited the ability of this methodology in real-time processing as it could fail suddenly during the video processing in real time.

Publicly available software EdgeTrack [1] proposed an improvement to the mentioned work in [23]. EdgeTrack implemented an enhanced methodology for the active contours that incorporated the gradient, local image information, and object orientation, unlike the classical methods that relied only on the gradient information [1]. This improvement optimized the contour's lower boundaries and rejected any undesirable edges unrelated to the tongue. EdgeTrack software had a few technical limitations, and like any other deformable models, it could misidentify the true tongue contour's edges. EdgeTrack did not have any preprocessing capability, reducing the snake algorithm's efficiency as it is sensitive to noise. The software program could not process a long video sequence with more than 80 frames, limiting it to short recordings. This is not beneficial in the case of long speech processing sessions or a real-time analysis. EdgeTrack was computationally expensive because the algorithm relied on complex optimization techniques. In some cases, when there was a rapid movement during the speech, the tongue contour had a visible deformation that looked like a concave arc; the software tool failed because it did not use temporal smoothness in the minimized internal energy function. EdgeTrack results were validated by two experts who delineated the tongue contour manually. The mean sum of distances (MSD) accuracy measure was used to compare the results between EdgeTrack and manual ground truth data. The reported results were in the range of 1.83–3.59 mm for the MSD.

The multihypothesis approach [3] combined the traditional motion model, snake algorithm, and particle filter to track the tongue contour. The first step toward building the algorithm was by deriving a motion model based on manually prelabelled images. Next, tongue contours were extracted and then normalized with respect to the length and position. Following that, a principal component analysis (PCA) and mean shape were estimated, then the covariance matrix was computed by using the information from the tongue motion information such as the scale, shape, and position.

The snake algorithm used in [3] required to be initialized to process the tongue tracker by manually identifying points on the contour at the first frame to segment the tongue. After that, the particle filter was created by copying the segmented contour for a defined number of so-called particles. Next, a multihypothesis approach was created from each copied particle of the previous frame based on the derived motion model of the tongue scale, position, and coarse shape. The derived tongue contour model was then adapted using the snake algorithm to fit the tongue contour accurately. A band of energy-optimized constraints was used to choose the best particle by ensuring that the tongue contour was below the bright white arc on the tongue's upper surface. Two groups of subjects with Steinert's disease (a form of myotonic dystrophy that causes slow speech, distorted vowels, and consonants) and healthy subjects were used to validate the research study. The reported accuracy was 1.69 ± 1.10 mm for the mean sum of distances (MSD). However, the approach claimed that it was not highly dependent on the training data. The segmentation accuracy was still dependent on the number of particles, which increased the snake algorithm's computational complexity [3].

To fully automate the tongue contour extraction without using training data or human interaction, some researchers designed multistage techniques [5]. Unlike other semiautomated methodologies such as those in [1][2][24], which required human interaction in the first frame, this methodology initiated the active contour model by automatically deriving candidate points on the tongue contour. These points were identified by applying the phase symmetry

method for image enhancement. Then, the image was skeletonized, and data points were clustered to select the best candidate points. These candidates were used as initialization points for the algorithm. The accuracy improved by implementing two methodologies for algorithm resetting or reinitialization in a frequent and timely manner order. According to the results, the measured mean sum of distances (MSD) accuracy measure was similar to that of other semiautomated techniques. They claimed that the MSD was 1.01 mm and 0.63 mm for their fully automated and reinitialized techniques, respectively. The reported results were highly accurate with some frames, but this may not be easy to achieve when processing videos in real time.

However, relying on the active contour model for tongue tracking in ultrasound images is error-prone and maybe not the most efficient technique. In some cases, it can lead to ultimate failure due to the number of constraints needed for the model adaption, which is difficult to predict for all cases accurately. Although the approach in [5] proposed a novel methodology for automating the process of identifying the active contour initialization and reinitialization parameters, this was still not enough to produce highly accurate results in a global and generalized context. There are many variations in ultrasound imaging modalities that produce different imaging qualities, making it difficult to track the tongue contour using the same active contour model constraints.

The similarity-constrained active-contour-based methodology for tongue tracking proposed in [25] suggested a technique that coped with the tongue contour tracking errors and missing data based on the tongue shape from previous contours to minimize the effect of missing data. In order to deal with the accumulated error during the continuous tracking of the tongue contour over a video sequence, a complex-wavelet image similarity index (CW-SSIM) was proposed to reinitialize the tongue tracker automatically. This algorithm showed an advancement compared to traditional techniques by handling missing data and using an automatic reinitialization. However, it was still based on the active contour, which is error-prone and sensitive to noise. Too many constraints would enhance the model accuracy but increase the computation cost. The best-reported results using similarity constraint + CW-SSIM were an MSD of 0.9912 ± 0.2537 mm.

As mentioned before, all methodologies that are based on the active contour may suddenly fail and the tongue tracker would stop. An initializer, either manual or automatic, is needed to enhance the accuracy of tongue tracking. The researchers in [26] conducted a comparative study on the effect of an automatic reinitialization technique to enhance the well-known traditional image segmentation. The automatic reinitialization enhanced the results from an MSD of 5–6 pixels to about 4 pixels (1 pixel = 0.295 mm). The MSD accuracy results without the need for automatic reinitialization for the well-known tongue tracking tools EdgeTrack and TongueTrack were 7.06 ± 2.77 pixels and 5.59 ± 3.04 pixels, respectively. The MSD accuracy after using the automatic reinitialization was 3.46 ± 1.04 pixels and 3.60 ± 0.96 pixels for EdgeTrack and TongueTrack, respectively.

2.2. Shape Consistency and Graph-Based Tongue Tracking Methodologies

Researchers derived an active appearance model to predict the tongue contour shape on ultrasound images in [27]. The active appearance model was inspired and estimated using a manual delineation and extraction of the tongue contour from tongue X-ray images. The results were compared to those of EdgeTrack [1] and the constrained snake

algorithm [28], which combined ultrasound, EMA, and recorded voice to predict the tongue shape. The work in [27] showed an improvement in root mean square error compared to that of [1][28]. The active shape model (ASM) was also evaluated and used in [18]; the authors showed that the ASM was efficient and powerful for phonological applications. It was able to capture the tongue motion variation by capturing the temporal information. It was also useful for either automated or semiautomated techniques.

Lingual ultrasound tracking was introduced in another well-known software called [2] TongueTrack, which could process a sequence of 500 frames. The methodology considered contextual information and advanced optimization techniques to estimate unpredictable tongue motion. The reported accuracy was 3 mm, making it acceptable for segmentation purposes. The tool used a higher-order Markov random field energy minimization framework. The results were validated with the ground truth data from two different groups of 63 acoustic videos [2].

The process of TongueTrack required an initial human interaction by manually delineating a few points on the first tongue contour to be used as an initializer for the algorithm. After that, the delineated points were fitted by using a curve-fitting polynomial function to build a continuous and smooth contour. Next, a solution-space label set was created by generating an estimation model for the dynamic tongue motion. This label set was used to compare each contour with the minimized Markov random field energy module in each subsequent frame. It processed it iteratively until reaching a predefined threshold; it was predefined as 2 mm in [2]. The tool obtained good results, but it had a few drawbacks. The software tool could not process long video frames. At the same time, the algorithm optimizer might not converge properly, leading to a sudden failure in tracking progress as it required 20 iterations to optimize nine parameters. Moreover, the algorithm needed a manual reinitialization by delineating the tongue contour by hand, limiting its efficiency for real-time processing.

Tongue contours are also tracked in ultrasound images by using graph-based analysis of the temporal and spatial information during speech [29]. Spatial information is essential to extract tongue features from each image on a single frame. At the same time, the temporal resolution is necessary to predict the intrarelationship between the entire sequence of image frames extracted from the video session of the speech. The tongue tracker was implemented as an optimization problem using a Markov random field energy minimization. The algorithm enforced temporal and spatial regularization constraints to ensure tongue tracking reliability.

In the landmark-based tongue contour tracking [24], the tongue shape was predicted based on the position of a few pellet plates used as landmarks on the tongue surface. The landmarks were extracted from the available articulatory database. The available landmark positions were smoothed using the spline function and compared to the ground truth data extracted by ultrasound images. Tongue contours extracted by ultrasound helped to identify the optimum number of required landmarks to get the desired accuracy of 0.2–0.3 mm for any future use.

Another research study coped with the tongue tracking problem by modelling it as a biomechanical method [30]. The methodology was initialized by manually drawing a closed contour around the external and internal edges of the tongue. The Harris feature detector was used to identify the one hundred most significant corners or edge features. The detected points were sorted in descending order based on the quality of the feature. An optical flow

algorithm was then used to estimate each point's displacement in the consequent frames. The corner feature displacement estimation was approximated only in the neighbour pixels (around 15–20 pixels) to minimize the displacement error in case of any missing data. In order to minimize the uncertainty of the estimated features, a covariance matrix was computed. The accuracy was measured by the mean sum accuracy, which was reported between 0.62 mm and 0.97 mm. However, the study faced many challenges. The algorithm required many parameters and constraints to be computed in order to estimate the displacement. Relying on the Harris feature detector may not have been efficient, especially in the case of rapid tongue movement, missing details, or extreme deformation, as it was almost impossible to guarantee that the same detected corner features were visible in the next frame within the neighbourhood pixel constrains.

An interactive approach for lingual ultrasound segmentation that incorporated four stages from preprocessing to the segmentation and postprocessing analysis was introduced in [4]. In the first stage, and unlike other methodologies that ignored an essential part of image denoising, the thesis implemented novel denoising techniques by using a combined curvelet transform and shock filter. In the second stage, the thesis derived an interactive model that predicted the tongue area of interest to minimize the computation complexity and contour tracking error. The third stage focused on tongue contour extraction and smoothness. The fourth stage proposed a new technique that transformed the extracted tongue contour from an image state to a continuous signal which resembled a full video for all frames. The advantage of this technique was that it enabled the researcher to extract a unique signature of each sound; this could be beneficial for training a machine learning model on sound pattern recognition. The tongue contour segmentation results were validated and compared to ground truth data. The mean sum of distances (MSD) was 0.955 mm.

3. Machine-Learning-Based Techniques for Tongue Contour Tracking

One of the early attempts to use deep learning for automatic tongue extraction was made by [31]. Their methodology, Autotrace, was implemented using a translational deep belief neural network (tDBN), which was based on restricted Boltzmann machines (RBMs). The network was trained based on human-labelled and generated sensor data. The hybrid data training methodology was efficient for improving tongue contour segmentation accuracy. However, there were discrepancies in the segmentation of some image frames and model-segmented tongue-unrelated parts. The results were validated by using a five-fold cross-validation, and the reported accuracy was measured by an average mean sum of distances (MSD) of 2.5443 ± 0.056 pixels (1 pixel = 0.295 mm [1]). The algorithm segmentation capabilities were fair enough; however, a postprocessing algorithm was needed to refine and enhance the final tongue contour segmentation.

To improve Autotrace [31], researchers in [32] proposed a new technique that automatically labelled the tongue contour, followed by training the algorithm in two phases. Using a deep autoencoder, the algorithm learned the relationship between the extracted contour and the original ultrasound image. By using the training data, the algorithm was able to reconstruct the tongue contour from ultrasound images without human intervention. The results were validated by comparing the average mean sum of distances between the hand-labelled and the deep-

learning-extracted contours. The average MSD was reported as 1.0 mm, making it applicable to lingual ultrasound applications.

Based on the principal component analysis (PCA) and a neural network, an automatic algorithm was designed to segment the tongue contour [33]. The PCA-based feature extractor, Eigen Tongue, was used to extract the tongue contour features from the ultrasound images. The visual features of the extracted Eigen Tongue were processed using an artificial neural network based on the PCA feature model. The model was evaluated by using 80 annotated images from nine speakers. The average error measured by the MSD was reported to be around 1.3 mm.

Typical convolutional neural networks were used to classify the tongue gesture from B-mode ultrasound images on the midsagittal plane in [34]. The researchers used data augmentation to increase the size and versatility of the data, which increased the algorithm's performance. The reported accuracy results for the classification task were 76.1%. Further improvements were suggested as future work. The recommended improvements were in the model optimization or combining the methodology with a hybrid technique such as the ensemble method.

The well-known U-net architecture [35] was used by [36] to automatically extract the tongue contour in ultrasound images. The algorithm was trained by using 8881 human-labelled images collected from three subjects. The results were validated by using the Dice score, which was 0.71. Relying on the Dice score only for validation is not enough. More validation is needed for their methodology, such as the mean sum of distances (MSD) measure, which has become a de facto standard in the lingual ultrasound accuracy measures. The MSD provides a reliable measure that considers the variation of the tongue contour length, which normalizes the sum of distances over the tongue contour length. To further enhance the performance, it might be needed to use a hybrid technique and larger dataset.

To automate tongue segmentation, a convolutional-neural-network-based architecture was utilized in [37]. They compared the efficiency of using the U-net [35] and Dense U-net [38] architectures to extract the tongue contour. These architectures have become de facto models of biomedical image segmentation and gained a wide popularity in the field. The results showed that Dense U-net was more generalizable for a wide variety of datasets. At the same time, the standard U-net architecture could perform the tongue extraction task faster. After extracting the tongue contour, it had to be postprocessed. In the postprocessing stage, the output was fed into a probability heat-map model, where the intensity of each pixel corresponded to the probability of each part of the tongue [37]. A 50% threshold was applied to filter out any undesired predictions. The remaining output was skeletonized to reduce the segment thickness. Following that, the results were smoothed and interpolated using the UnivariateSpline function in the SciPy package in Python. The final output was a hundred points to represent the predicted tongue. The algorithms were evaluated using the MSD for the 17,580-frame dataset. The reported MSD results for the 32×32

data size were 5.81 mm and 5.6 mm for U-net and Dense U-net, respectively. The research also showed that data augmentation and the loss function significantly affected model performance other than stacking more layers.

Two deep learning architectures were designed, BowNet and wBowNet, to extract the tongue contour from ultrasound in [39]. With the integrated multiscale contextual information, the decoding–encoding model had the ability for global prediction. The dilated convolution had the local searching capability of preserving image features more than standard convolution, making it valuable for medical imaging applications to retain fine image details. The two architectures enhanced the final prediction results by combining the local and global searching. The mean sum of distances for BowNet and wBowNet compared to the greyscale ground truth images was in a range of 0.2874–0.4014 in pixels for BowNet and 0.1803–0.3588 pixels for wBowNet. However, the reported results appeared to be almost perfect, which is not easy to achieve in the case of a complex analysis of lingual ultrasound. The researchers need to provide more information about the data validation in a generalized clinical context by using a dataset from a different source.

A simple approach to extracting the tongue contour by training a deep network on landmarks annotated on the tongue contour was developed in [40]. These landmarks were automatically and randomly selected on different points by using annotation software. The model architecture was called TongueNet, and the results were validated by the mean sum of distances which achieved 4.87 pixels.

Using U-net and the lighter version of sU-net in a thesis work, a deep learning approach was implemented to segment tongue contours [41]. In their thesis, the researcher emphasized the validity and performance of deep learning models to segment the tongue contours from ultrasound images. However, they suggested that the deep learning model they used only focus on the spatial information on a single image frame without considering the temporal information that handled the full speech in the video sequence. The thesis [41] also discussed the limitations of their deep learning model in their generalization capability of feature extraction, as they inherited the nongeneralization of convolutional neural networks (CNN) models, which is the core of a deep learning model such as the U-net architecture. The thesis suggested using data augmentation to enhance the model training by considering the variation and image transformation to handle different cases at different scales.

A denoising convolution autoencoder (DCAN) model to process B-mode ultrasound images was investigated in [42]. The model reported being able to extract image features due to its ability to denoise and retain the resolution of the reconstructed input from the ultrasound. It was tested on reconstructing ultrasound images in speech-related applications. The research compared the DCAN to other three well-known autoencoder architectures, the deep autoencoder (AE), the denoising autoencoder (DAE), and the convolutional autoencoder (CAE). The reported result showed that the DCAN had a 6.17% error rate in identifying words in a silent-speech recording test [42].

Researchers implemented a novel technique that harnessed the spatial–temporal analysis to predict future tongue movement based on a short recording of the past tongue motion in [43]. The research used a combination between a convolutional neural network (CNN) and long short-term memory (LSTM), which was called ConvLSTM. The advantage of this combination was that the CNN had the ability to segment tongue contour in each image frame to extract spatial information. However, it could not process the temporal information of ultrasound image sequence frames. On the other hand, LSTM was used in processing data sequence in one dimension, making it efficient for temporal information data prediction, but at the same time, it was unable to handle images in two dimensions (2D).

The ConvLSTM could handle image data in 2D and predict future data based on the history of tongue motion. The ConvLSTM results outperformed the three-dimensional convolutional neural network (3DCNN) in predicting future tongue contours. The ConvLSTM was able to predict the future nine frames based on data from the previous eight frames.

An algorithm combining an image-based segmentation model, U-net, and a shape consistency regularizer was proposed by [44]. The combination provided a solution to the missing data in ultrasound images by predicting the information based on the consideration of the sequential information of the shape regularizer. The regularizer was derived based on the similarity between adjacent image frames. The results were validated by computing the MSD of the tongue contour data segmented by the U-net algorithm using different loss functions. The quantitative validation showed that the combination between the regularizer and cross-entropy loss (CE) obtained the best results among the other compared losses such as the Dice coefficient (DC) or the active contour loss (AC). The CE+regularizer reported having an MSD of 2.243 ± 0.026 mm.

To improve the well-known U-net architecture, researchers proposed a tongue contour segmentation algorithm called wUnet [45]. The main modification of wUnet was replacing the skip connection in typical U-net with a VGG19 block. The researchers claimed that the new algorithm surpasses U-net by passing more information to the decoder to compensate for the information loss during the convolution within the encoder. The wUnet validation results showed an MSD of 1.18 mm compared to 2.26 mm in the U-net architecture.

A system based on a deep learning technique was designed to predict silent speech using ultrasound images in [46]. The system was trained on audio features recorded synchronously with ultrasound images using a deep convolutional neural network. The system was designed to predict the speech sound from the silent speech based on the training data. This methodology could be beneficial for human–machine interaction in smart devices.

To update an older silent-speech benchmark study [47], the work [48] used a deep learning approach for the same benchmark. The new study used a deep autoencoder to train the collected dataset from acoustic tongue and lips movement videos, which were collected at the same time.

The research [8] used ultrasound videos to extract tongue features using deep learning. The dataset was collected from 82 speakers and trained using the Kaldi speech recognition toolkit [49]. In terms of speech analysis, the research suggested two methodologies. The first one was the utterance or speech duration, which was measured based on the syllable rate. The second one was the articulatory area, which was measured by estimating the convex hull area, which was the area under the tongue contour spline that formed a convex-like shape when extracted from the ultrasound images using the MTracker tool [36]. Following that, a postprocessing was performed by the isolation forest method [50]. The research found that the silent articulation exhibited a longer time compared to the model speech.

References

1. Li, M.; Kambhamettu, C.; Stone, M. Automatic contour tracking in ultrasound images. Clin. Linguist. Phon. 2005, 19, 545–554.
2. Tang, L.; Bressmann, T.; Hamarneh, G. Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves. Med. Image Anal. 2012, 16, 1503–1520.
3. Laporte, C.; Ménard, L. Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. Med. Image Anal. 2018, 44, 98–114.
4. Al-hammuri, K. Computer Vision-Based Tracking and Feature Extraction for Lingual Ultrasound. Ph.D. Thesis, University of Victoria, Victoria, BC, Canada, 2019.
5. Karimi, E.; Ménard, L.; Laporte, C. Fully-automated tongue detection in ultrasound images. Comput. Biol. Med. 2019, 111, 103335.
6. Cai, J.; Denby, B.; Roussel-Ragot, P.; Dreyfus, G.; Crevier-Buchman, L. Recognition and Real Time Performances of a Lightweight Ultrasound Based Silent Speech Interface Employing a Language Model. In Proceedings of the Interspeech, Florence, Italy, 27–31 August 2011; pp. 1005–1008.
7. Lee, W.; Seong, J.J.; Ozlu, B.; Shim, B.S.; Marakhimov, A.; Lee, S. Biosignal sensors and deep learning-based speech recognition: A review. Sensors 2021, 21, 1399.
8. Ribeiro, M.S.; Eshky, A.; Richmond, K.; Renals, S. Silent versus modal multi-speaker speech recognition from ultrasound and video. arXiv 2021, arXiv:2103.00333.
9. Stone, M. A guide to analysing tongue motion from ultrasound images. Clin. Linguist. Phon. 2005, 19, 455–501.
10. Ramanarayanan, V.; Tilsen, S.; Proctor, M.; Töger, J.; Goldstein, L.; Nayak, K.S.; Narayanan, S. Analysis of speech production real-time MRI. Comput. Speech Lang. 2018, 52, 1–22.
11. Deng, M.; Leotta, D.; Huang, G.; Zhao, Z.; Liu, Z. Craniofacial, tongue, and speech characteristics in anterior open bite patients of East African ethnicity. Res. Rep. Oral Maxillofac. Surg. 2019, 3, 21.
12. Lingala, S.G.; Toutios, A.; Töger, J.; Lim, Y.; Zhu, Y.; Kim, Y.C.; Vaz, C.; Narayanan, S.S.; Nayak, K.S. State-of-the-Art MRI Protocol for Comprehensive Assessment of Vocal Tract Structure and Function. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 475–479.
13. Köse, Ö.D.; Saraçlar, M. Multimodal representations for synchronized speech and real-time MRI video processing. IEEE/ACM Trans. Audio Speech Lang. Process. 2021, 29, 1912–1924.
14. Isaieva, K.; Laprie, Y.; Houssard, A.; Felblinger, J.; Vuissoz, P.A. Tracking the tongue contours in rt-MRI films with an autoencoder DNN approach. In Proceedings of the ISSP 2020—12th International Seminar on Speech Production, Online, 14–18 December 2020.

15. Zhao, Z.; Lim, Y.; Byrd, D.; Narayanan, S.; Nayak, K.S. Improved 3D real-time MRI of speech production. Magn. Reson. Med. 2021, 85, 3182–3195.
16. Xing, F. Three Dimensional Tissue Motion Analysis from Tagged Magnetic Resonance Imaging. Ph.D. Thesis, Johns Hopkins University, Baltimore, MD, USA, 2015.
17. Stone, M.; Shawker, T.H. An ultrasound examination of tongue movement during swallowing. Dysphagia 1986, 1, 78–83.
18. Kaburagi, T.; Honda, M. An ultrasonic method for monitoring tongue shape and the position of a fixed point on the tongue surface. J. Acoust. Soc. Am. 1994, 95, 2268–2270.
19. Kass, M.; Witkin, A.; Terzopoulos, D. Snakes: Active contour models. Int. J. Comput. Vis. 1988, 1, 321–331.
20. Iskarous, K. Detecting the edge of the tongue: A tutorial. Clin. Linguist. Phon. 2005, 19, 555–565.
21. Akgul, Y.S.; Kambhamettu, C.; Stone, M. Extraction and tracking of the tongue surface from ultrasound image sequences. In Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231), Santa Barbara, CA, USA, 25 June 1998; pp. 298–303.
22. Akgul, Y.S.; Kambhamettu, C.; Stone, M. Automatic motion analysis of the tongue surface from ultrasound image sequences. In Proceedings of the Workshop on Biomedical Image Analysis (Cat. No. 98EX162), Santa Barbara, CA, USA, 27 June 1998; pp. 126–132.
23. Akgul, Y.S.; Kambhamettu, C.; Stone, M. Automatic extraction and tracking of the tongue contours. IEEE Trans. Med. Imaging 1999, 18, 1035–1045.
24. Qin, C.; Carreira-Perpinán, M.A.; Richmond, K.; Wrench, A.; Renals, S. Predicting Tongue Shapes from a Few Landmark Locations. Available online: http://hdl.handle.net/1842/3819 (accessed on 14 August 2022).
25. Xu, K.; Yang, Y.; Stone, M.; Jaumard-Hakoun, A.; Leboullenger, C.; Dreyfus, G.; Roussel, P.; Denby, B. Robust contour tracking in ultrasound tongue image sequences. Clin. Linguist. Phon. 2016, 30, 313–327.
26. Xu, K.; Gábor Csapó, T.; Roussel, P.; Denby, B. A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization. J. Acoust. Soc. Am. 2016, 139, EL154–EL160.
27. Roussos, A.; Katsamanis, A.; Maragos, P. Tongue tracking in ultrasound images with active appearance models. In Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; pp. 1733–1736.
28. Aron, M.; Roussos, A.; Berger, M.O.; Kerrien, E.; Maragos, P. Multimodality acquisition of articulatory data and processing. In Proceedings of the 2008 16th European Signal Processing

Conference, Lausanne, Switzerland, 25–29 August 2008; pp. 1–5.

29. Tang, L.; Hamarneh, G. Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 154–161.
30. Loosvelt, M.; Villard, P.F.; Berger, M.O. Using a biomechanical model for tongue tracking in ultrasound images. In Proceedings of the International Symposium on Biomedical Simulation, Strasbourg, France, 16–17 October 2014; pp. 67–75.
31. Fasel, I.; Berry, J. Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 1493–1496.
32. Jaumard-Hakoun, A.; Xu, K.; Roussel-Ragot, P.; Dreyfus, G.; Denby, B. Tongue contour extraction from ultrasound images based on deep neural network. arXiv 2016, arXiv:1605.05912.
33. Fabre, D.; Hueber, T.; Bocquelet, F.; Badin, P. Tongue tracking in ultrasound images using eigentongue decomposition and artificial neural networks. In Proceedings of the Interspeech 2015—16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
34. Xu, K.; Roussel, P.; Csapó, T.G.; Denby, B. Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images. J. Acoust. Soc. Am. 2017, 141, EL531–EL537.
35. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
36. Zhu, J.; Styler, W.; Calloway, I.C. Automatic tongue contour extraction in ultrasound images with convolutional neural networks. J. Acoust. Soc. Am. 2018, 143, 1966.
37. Zhu, J.; Styler, W.; Calloway, I. A CNN-based tool for automatic tongue contour tracking in ultrasound images. arXiv 2019, arXiv:1907.10210.
38. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
39. Mozaffari, M.H.; Lee, W.S. Encoder-decoder CNN models for automatic tracking of tongue contours in real-time ultrasound data. Methods 2020, 179, 26–36.
40. Mozaffari, M.H.; Yamane, N.; Lee, W.S. Deep Learning for Automatic Tracking of Tongue Surface in Real-Time Ultrasound Videos, Landmarks instead of Contours. In Proceedings of the 2020

IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea, 16–19 December 2020; pp. 2785–2792.

41. Wen, S. Automatic Tongue Contour Segmentation Using Deep Learning. Ph.D. Thesis, University of Ottawa, Ottawa, ON, Canada, 2018.
42. Li, B.; Xu, K.; Feng, D.; Mi, H.; Wang, H.; Zhu, J. Denoising convolutional autoencoder based B-mode ultrasound tongue image feature extraction. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7130–7134.
43. Zhao, C.; Zhang, P.; Zhu, J.; Wu, C.; Wang, H.; Xu, K. Predicting tongue motion in unlabeled ultrasound videos using convolutional LSTM neural networks. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5926–5930.
44. Feng, M.; Wang, Y.; Xu, K.; Wang, H.; Ding, B. Improving ultrasound tongue contour extraction using U-Net and shape consistency-based regularizer. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–12 June 2021; pp. 6443–6447.
45. Li, G.; Chen, J.; Liu, Y.; Wei, J. wUnet: A new network used for ultrasonic tongue contour extraction. Speech Commun. 2022, 141, 68–79.
46. Kimura, N.; Kono, M.; Rekimoto, J. SottoVoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–11.
47. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M.; Brumberg, J.S. Silent speech interfaces. Speech Commun. 2010, 52, 270–287.
48. Ji, Y.; Liu, L.; Wang, H.; Liu, Z.; Niu, Z.; Denby, B. Updating the silent speech challenge benchmark with deep learning. Speech Commun. 2018, 98, 42–50.
49. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
50. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.