Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.034687
Article

A 3D Geometry Model of Vocal Tract Based on Smart Internet of Things

Ming Li[1], Kuntharrgyal Khysru[2], Haiqiang Shi[2,*], Qiang Fang[3,*], Jinrong Hu[4] and Yun Chen[5]

[1]Tianjin Medical University General Hospital, Tianjin, 300041, China
[2]Key Laboratory of Artificial Intelligence Application Technology State Ethnic Affairs Commission, Qinghai Minzu University, Xining, 810007, China
[3]Chinese Academy of Sciences, Beijing, 300308, China
[4]School of Computer Science, Chengdu University of Information and Technology, Chengdu, Sichuan, 610225, China
[5]College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China
*Corresponding Authors: Haiqiang Shi. Email: shihaiqiang2022@126.com; Qiang Fang. Email: fangqiang@cass.org.cn
Received: 24 July 2022; Accepted: 19 October 2022

Abstract: The Internet of Things (IoT) plays an essential role in the current and future generations of information, network, and communication development and applications. This research focuses on vocal tract visualization and modeling, which are critical issues in realizing inner vocal tract animation. That is applied in many fields, such as speech training, speech therapy, speech analysis and other speech production-related applications. This work constructed a geometric model by observation of Magnetic Resonance Imaging data, providing a new method to annotate and construct 3D vocal tract organs. The proposed method has two advantages compared with previous methods. Firstly it has a uniform construction protocol for all speech organs. Secondly, this method can build correspondent feature points between different speech organs. There are less than three control parameters can be used to describe every speech organ accurately, for which the accumulated contribution rate is more than 88%. By means of the reconfiguration, the model error is less than 1.0 mm. Regarding to the data from Chinese Magnetic resonance imaging (MRI), this is the first work of 3D vocal tract model. It will promote the theoretical research and development of the intelligent Internet of Things facing speech generation-related issues.

Keywords: Virtual reality; vocal tract visualization; articulatory modeling; IoT

1 Introduction

As intelligent devices increase, Internet of Things (IoT) has attracted more and more researchers' attention and interest. In terms of the network of devices and objects, IoT makes the interactions of machine-to-machine come true. Besides the machine interaction, another critical component is the human-machine interaction. With a connected life, people, appliances, vehicles are connected together. Commonly, speech is used to during the connection and enhances the customer experience. As a IoT subset, innovative speech is a rapid-developing industry, including speech training, speech therapy, speech analysis and other speech production-related applications. This paper combines the relevant

technologies of the Internet of Things and builds a 3D vocal tract organ, which will provide a research basis for subsequent speech recognition and other issues.

Talking head has been studied widely in different fields for many decades. For example, Zhang et al. [1] A new Neural Machine Translation (NMT) based system is an essential technology for translation applications. Zhao et al. [2] proposed a method using latent regression Bayesian network (LRBN) to extract the shared speech feature for the input of the end-to-end speech recognition model. Nisar et al. [3] the proposed system uses an adaptive filter bank with weighted Mel-Frequency Cepstral Coefficients for feature extraction. Meanwhile, MRI data has already played an essential role in research. For example, Mengash et al. [4].

The facial animation has high naturalness. The animation of the inner vocal tract, however, has not been well developed so far. Vocal tract animation could offer wide applications in speech training, speech therapy, etc. Showing the synthesized articulatory movements, as well as the synthesized corresponding speech, would facilitate speech training and understanding.

In terms of the mechanism study for the speech production, two types of modeling strategies exist. They are physiologic and geometry modeling. Different from the physiologic strategy, geometric approach does not mimic mechanism and function of the tongue muscles. The geometric models have merits in high calculation efficiency and real-time response. The geometric model can be driven by parameters obtained by analysis of the labeled original dataset rather than using finite element mesh, a time-consuming process, as did in the physiological model. For speech applications, the geometric strategy seems to be a more promising approach.

Some researchers have built vocal tract models in the past couple of decades. Mustaqeem et al. [5] designed a one-dimensional Convolutional Neural Network (CNN) network to enhance the speech signals, which uses a spectral analysis. Maeda described the tongue with four parameters using factor analysis Maeda et al. [6]. Badin et al. analyzed the original dataset based on factor analysis and then achieved six parameters for describing the 3D model of the tongue [7–9]. Additionally, Peter also had computer tomography (CT) scans of oral-dental impressions. The CT scans were used to adapt the geometry of the maxilla, the jaw, and the teeth Peter et al. [10]. The engwall team built a three-dimensional tongue model based on EMA, MRI and EPG data fusion for Swedish data Olov et al. [11]. Gérard et al. [12] built a physiological tongue model by collecting data from a French speaker. Dang et al. (Dang & Honda, 1998) built a physiological articulatory model by using Japanese speakers' data Dang et al. [13]. However, there is no vocal tract model using Chinese data due to no Chinese Magnetic Resonance Imaging (MRI) data collected before. Therefore, our research aims to analyze and construct a Chinese-oriented vocal tract geometric model based on the Chinese MRI database. Engwall (2003) presents an MRI-adapted, 3D parametric tongue model, which can be externally controlled by EMA data. Still, this data is only two-dimensional, and the mapping from the EMA coils to model control parameters is somewhat simplified.

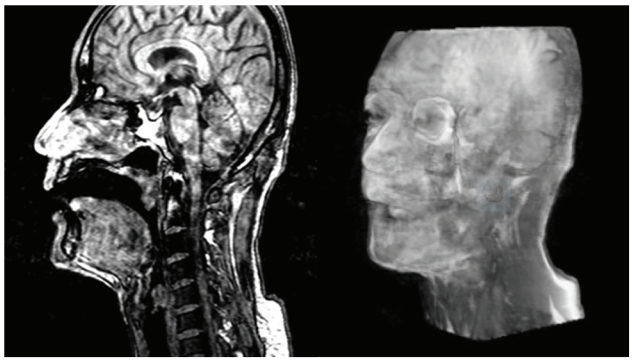
Wang et al. [14] proposes a novel pairing-free certificates scheme that utilizes the state-of-the-art block-chain technique and intelligent contract to construct a novel reliable, and efficient CLS scheme. Haddad et al. [15] introduce a novel, efficient and secure authentication and key agreement protocol for 5G networks using block-chain. Li et al. [16] propose a lightweight mutual authentication protocol based on a novel public key encryption scheme for innovative city applications. Xiong et al. [17] propose an efficient and large-scale batch verification scheme with group testing technology based on ECDSA. The application of the presented protocols in Bitcoin and Hyperledger Fabric has been analyzed as supportive and effective. Wang et al. [18] propose a lightweight and reliable authentication protocol for WMSN, composed of cutting-edge blockchain technology and physically unclonable functions (PUFs).

This paper provides a new method to annotate and construct 3D vocal tract organs. The proposed method has two advantages compared with previous methods. Firstly it has a uniform construction protocol for all speech organs. Secondly, this method can build correspondent feature points between different speech organs.

The paper is organized as follows, where the second part deals with the MRI database as well as the processing techniques, the 3D vocal tract model is presented in the third part, and related parameters, the modeling and experimental comparison results are then evaluated, and the last part summarizes and concludes.

2 MRI Database of Mandarin Speech Production

Experiments use the Mandarin MRI database, which contains static two-dimensional and three-dimensional vocal tract data. Databases include a total of eight people, and each one pronounces ten vowels and few consonants. The subject took a prone position during the scan, holding blueberry juice inside the mouth as an oral contrast medium for MRI Hiraishi et al. [19]. Each vowel group contains 51 frames of head cross section parallel to sagittal images (as shown in Fig. 1a), and all of the planes are put together to reconstruct the three-dimensional shape of the head (as shown in Fig. 1b shown). The plane size is 512 * 512 with a pixel as a unit, wherein a length of two pixels is equal to 1 mm; the center distance of two planes is 2 mm; in our study, we select the vowels and the tooth mold part of one of the objects.



(a) (b)

Figure 1: (a) An MR image of the original database; (b) A three-dimensional model

The tooth is a bone structure, and bone structures in the MRI acquisition process, the same as air, cannot be displayed directly (as shown in Fig. 1a). We want to describe the shape of a complete vocal tract, so the data needs to include the shape of the teeth. Therefore before labeling the organs, firstly, we take the dental filling method of Takemoto to fill MRI data Takemoto et al. [20], and then label the organs (as shown in Fig. 2).

Ideally, label each of the articulator pronunciations of each frame on the picture and then superimpose all 51 frames images to synthesize the three-dimensional shape of the articulatory organs. The boundary near both sides of the organs in the MR images not obvious, such as the tongue and pharyngeal wall. When the plane is perpendicular to the main direction of extension of the organ, the performance is the best. Still, the boundary on both sides of the tongue is almost parallel to the sagittal plane.

Contraposing this phenomenon, a set of planes has been selected at another angle to label the shape of both sides of the tongue. What is referred to here, the transverse plane, is achieved by re-segmenting the three-dimensional head in Fig. 1b. The head part in Fig. 3 has been divided into 50 sections between eyes and glottis. Wherein, there are 7.84 pixels between the center distance of neighboring two sections, i.e.,

3.92 mm, and there are 512 ∗ 512 pixels in each section. The tongue and pharyngeal wall boundaries are more precise, so we can annotate the outline of the tongue and get a more accurate profile on both sides of the tongue. Almost parallel to the tongue dorsum, transverse plane cannot accurately describe the shape of the tongue dorsum, so we need to integrate these two kinds of the plane to get the final shape of the tongue.

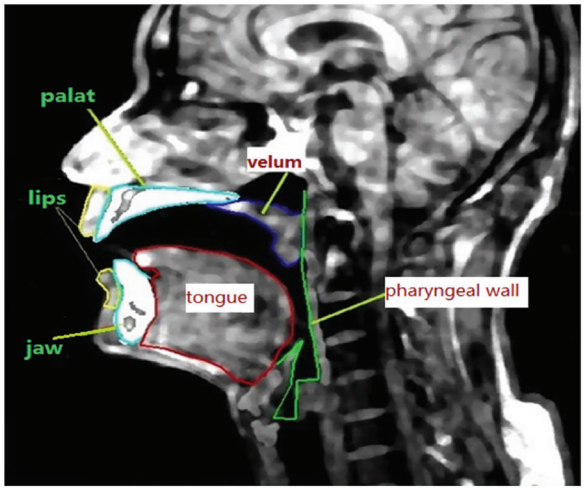
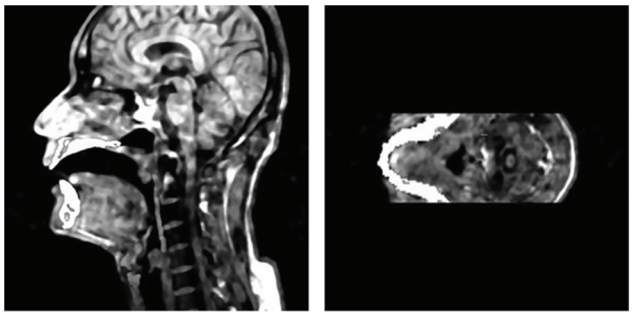


Figure 2: Outlines the organ of marked MR images in the database



(a) (b)

Figure 3: (a) is an MR image after adding the teeth, and (b) is the plane of the MR image after transverse

3 Three-Dimensional Vocal Tract Modeling

3.1 Data Annotations

In our annotating system, shown in Fang et al. [21] there are six modules are involved in the vocal tract, including the upper and lower lips, jaw, palate (including the hard palate and velum), tongue and pharyngeal wall. Annotating is mainly based on the sagittal image, but the tongue and pharyngeal wall need to annotate the outline of their transverse planes and then fuse both data. Meanwhile, the other four organs use the shape of their sagittal plane as the final one, owing to the boundary contour of the sagittal plane is accurate enough. In addition, it should be noted that since the lower jaw and tongue are connected with each other, the annotated points of their overlapping parts must be identical to ensure that the lower jaw and tongue correspond correctly.

The annotated data also needs to be preprocessed, including data calibration of the jaw and palate, smoothing of the organs' surface, and unified treatment of the physiological boundary of each organ.

1) Jaw modeling

From the physiological structure of humans, lower teeth are a kind of bone material fixed together with the jaw so that we will take lower teeth and the jaw as a whole to annotate and analyze, collectively called the jaw. Simultaneously the jaw connects the tongue and lower lip, thus affecting their movement, so a accurate description of the jaw is vital to the whole structure.

There are two methods to annotate the jaw. One is entirely manual, annotating each different pronunciation; another method is adding rotation based on the manual method, which only annotates one shape of the jaw, and the other shapes are achieved by shifting and rotating the annotated jaw. The second method is an improvement to the first one, and it is based on the invariance of the jaw itself when pronouncing different voices, which will also avoid introducing artificial annotating errors. Translation amounts and rotation angles are obtained by calculating the difference between two contours of the jaw in midsagittal planes. Fig. 4 displays a three-dimensional grid shape of the jaw.

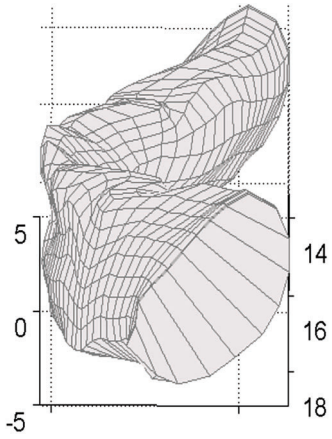


Figure 4: Mesh shape of the jaw when pronouncing /a/

2) Tongue Modeling

In the database, the bilateral portion of the tongue is not transparent in the sagittal plane, but in the re-segmented transverse plane, this part is obvious, while the blade of the tongue is missing (as shown in Fig. 5b). Therefore it needs to fuse these two kinds of data to obtain an accurate description of the tongue. Then we process the tongue in three levels:

First, we need to annotate the shapes in the sagittal planes and transverse planes separately. Second, we preprocess the annotated data to ensure that the physiological boundary points correspond between different images. There are three physiological boundary points: tongue tip, the junction between the fixed part and active part of the tongue and junction between the blade and root of the tongue. Thus, the tongue can be divided into four parts, and then we split each part equally. As a result, each tongue contains 39 points in each sagittal plane, the same as the transverse plane. Two kinds of tongue shape are shown in Figs. 5a and 5b. As shown, the shapes in the sagittal plane lack the shape of the bilateral tongue, and the transverse plane shapes lack information about the blade of the tongue.

Finally, we resample and smooth the fused tongue to improve the regularity and correspondence of surface points for the following analysis. Although the tongue after integration is complete, points on it

distribute irregularly, and the surface is rough. Therefore, we need to smooth and resample its surface to obtain the final shape of the tongue, as shown in Fig. 5c.

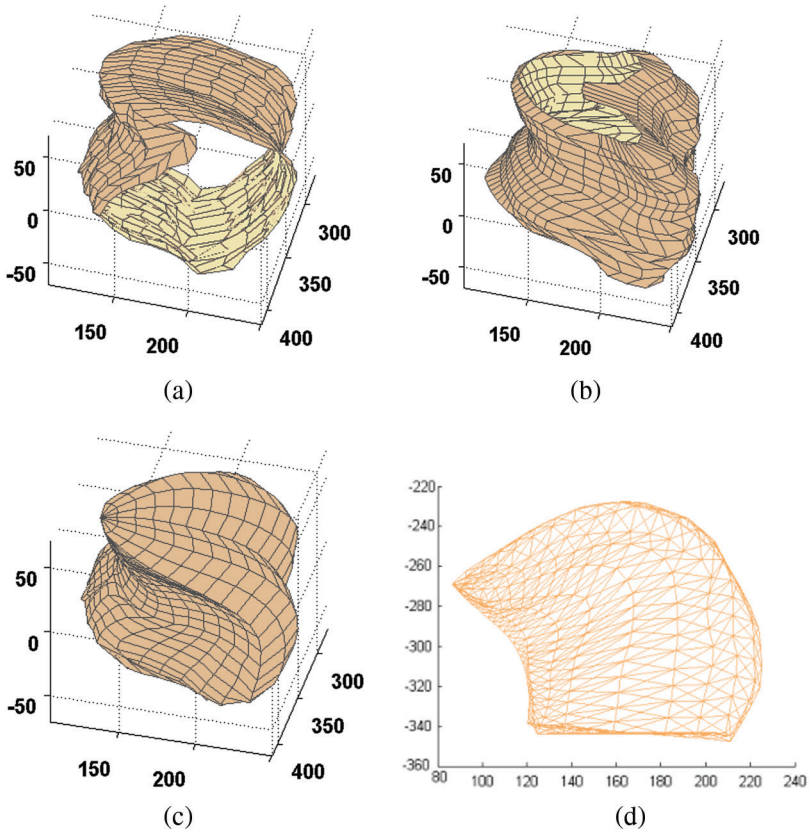


Figure 5: The forming process of the mesh shape of the tongue. (a) The annotated tongue of the sagittal plane; (b) The annotated tongue of the transverse plane, (c) The tongue after fusing and smoothing; (d) The 3D model of the tongue

3) Lips Modeling

Lips play a vital role in linguistic communication. Different voices show different lip shapes. The boundary of lips in the sagittal plane is clear enough, while the boundary in the transverse plane is not easy to identify. Therefore we use the annotated contour of lips in the sagittal plane as the final contour and display the upper and lower lip together, as shown in Fig. 6.

4) Lower lip Modeling

Analysis of the lower lip is consistent with the upper lip. Before adjusting the corner of the mouth, the cumulative contribution rate of the first three dimensions achieved 91.92%. After adjusting, the cumulative contribution rate of the first two dimensions is reached 92.58%. Then the reconstruction error is 0.0275 cm. The physical significance of each dimension is shown in Fig. 8.

As can be seen from the Fig. 7, the first dimension of the lower lip and upper lip has the same physical meaning, and both of them describe the narrowness and circle of the lips. The second dimension is mainly reflected in the description of up and down movement of the lips.

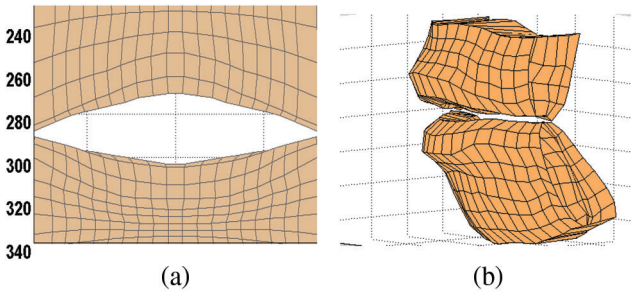


Figure 6: (a) The grid shape of lips when speaking /a/; (b) The 3D model of the upper and lower lips

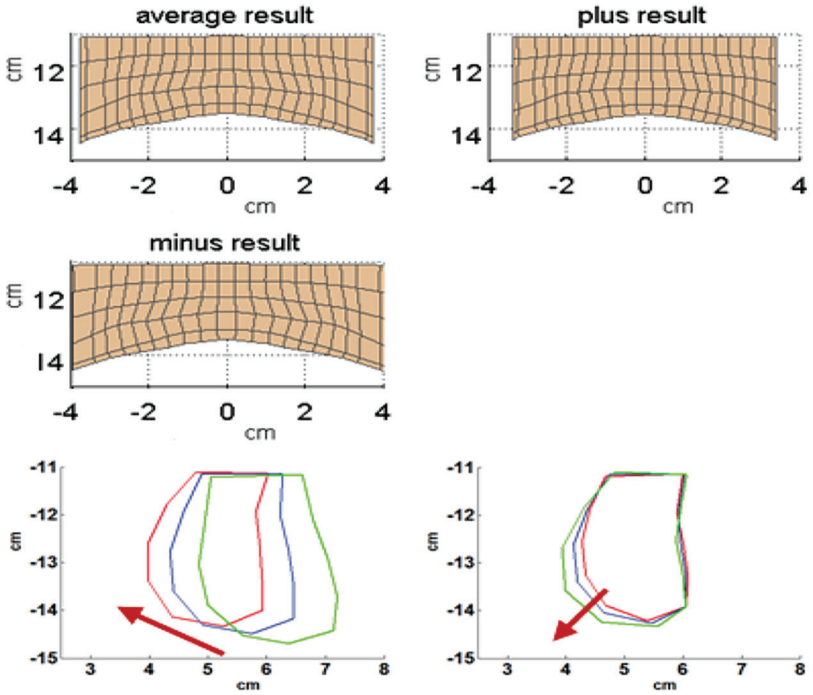


Figure 7: The physical significance of the first two dimensions of the upper lip. Above is the comparison chart of the first dimension at the three-dimensional shape, and the following is a comparison chart of the first two dimensions in the two-dimensional mid-sagittal plane

3.2 Baseline Method

We introduce our baseline system, which can be seen in Fig. 9. The baseline method is a conventional formant tracking algorithm based on dynamic programming.

The general scheme of the proposed VTR estimation procedure can be seen in Fig. 10.

3.3 Selection of Parameters

The DTW algorithm is employed in MFCC-DTW alignment to build MFCC vector pairs from comparable speeches. A frame width from 5 to 100 ms is usually used and shifted by 1/3 of the frame width. Euclidean distances are used to calculate two MFCC vector series in two recordings. By this way, MFCC-DTW algorithm obtains the warping path of minimum-distance. Each audio frame with each imaging time of repetition (TR) is measured via the nearest neighbor approach, which avoids any frame width and shifts size restriction in the synchronization8:

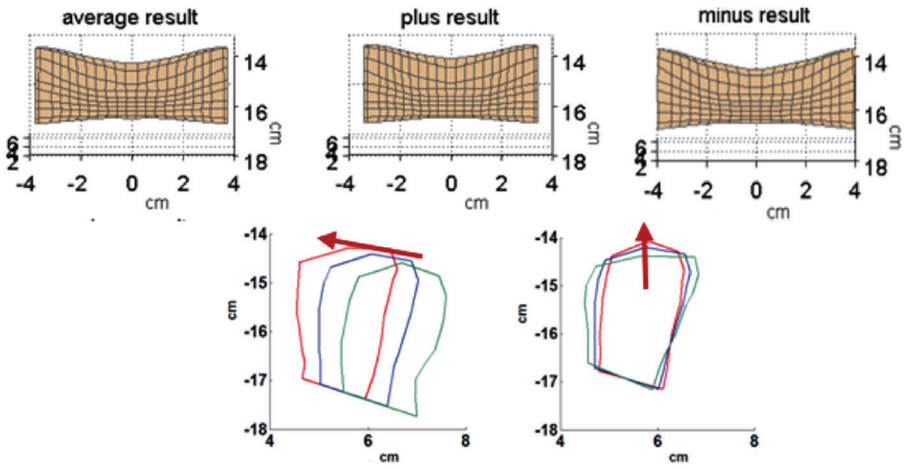


Figure 8: The physical meaning of the first two dimensions. Above is the comparison chart of the first dimension at the three-dimensional shape, and the following is a comparison chart of the two-dimensional sagittal of the first two dimensions

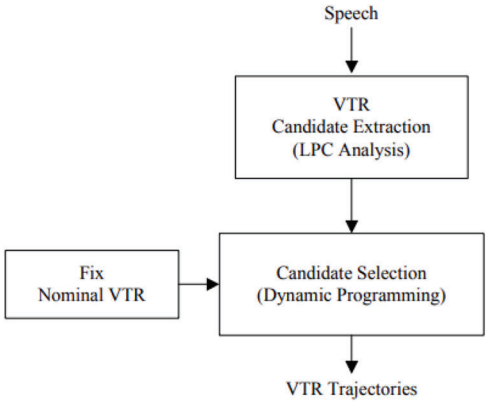


Figure 9: General scheme of baseline formant estimation procedure

$$\text{Mel}(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \tag{1}$$

where f is the frequency.

$$RMSE = \sqrt{\frac{1}{Ns}\sum_{i=1}^{Ns} ||v_r^i - v^i||^2} \tag{2}$$

where Ns is the number of the articulatory shape samples, Vr is the coordinates of a reconstructed articulatory vertex, and v is that of the original articulatory vertex.

Eqs. (1) and (2) are the core equations for this part. The flow of pairwise data processing can be seen in Fig. 11.

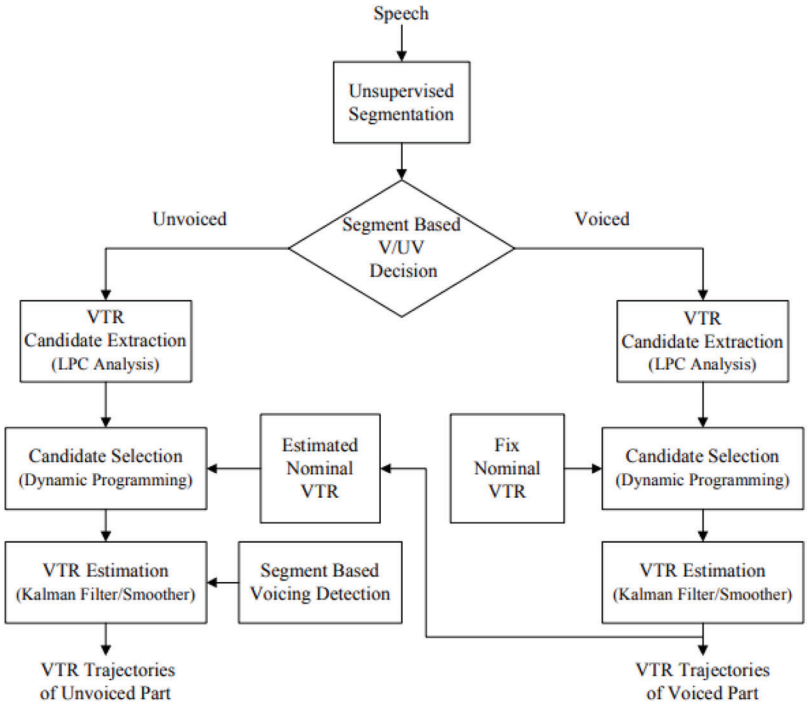


Figure 10: General scheme of proposed VTR estimation procedure

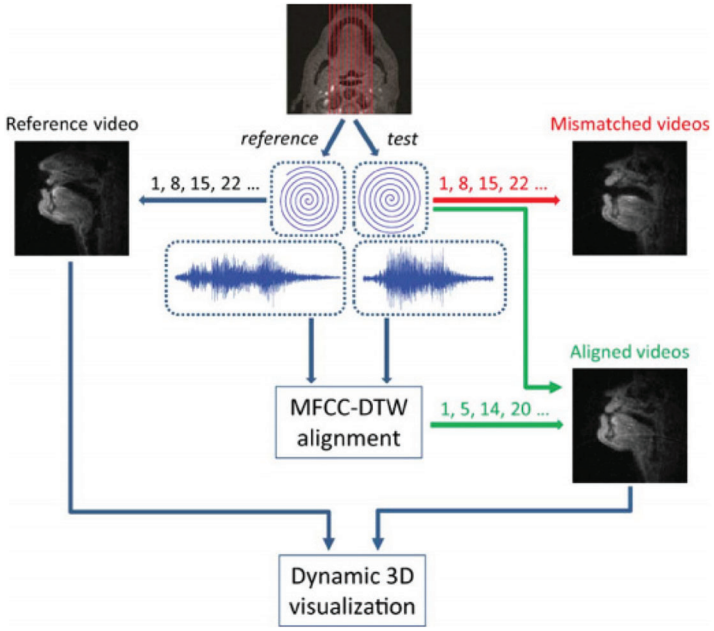


Figure 11: The data processing flow. Parasagittal scan planes shelters the upper airway. Real-time MR data and synchronized noise-canceled audio recordings are included. As the speech rate and duration change, unsuitable videos are constructed by the same sliding window. MFCC-DTW alignment synchronizes audio recordings. The resulting warping paths help the sliding window placement in aligned videos. Video and aligned videos from other sagittal planes contribute to dynamic 3-D visualizations of synthesized coronal movies and 3-D dynamics

In this step, a simultaneous estimation algorithm is implemented to accurately estimate a glottal source waveform and vocal tract shape based on the ARX-LF model.

4 Evaluation

4.1 Evaluation Scheme

Refactoring is based on the KTH tongue model (as shown in Fig. 13) of the Engwall team (as shown in Fig. 12). Their data is obtained by annotating three planes (sagittal plane, transverse plane, coronal plane). First, define a center on the tongue, and the parts under the center are extracted from transverse planes, a totally of five spacing equal transverse planes, denoted planes 1 to 5; Next, it is the 11 polar planes across the center, every ten angles between two planes, denoted plane 6 to 16; The remaining is four frame from the 16th plane to what is parallel to the 16th section and across the tip of the tongue, and distances between any two planes are the same, denoted plane 17 to 20; The junction outlines of these planes and the surface of the tongue are used to restructure the three-dimensional shape of the tongue.

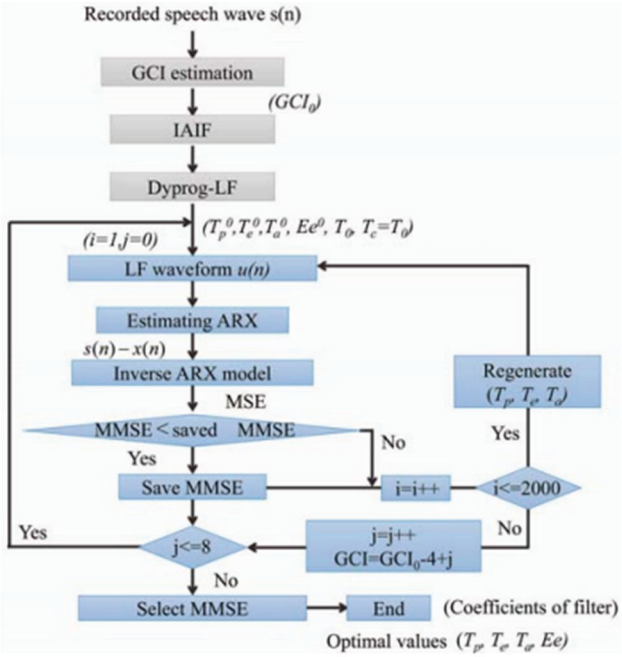


Figure 12: Estimation scheme of glottal source waveform and vocal tract shape

In the subsequent restructuring, limiting the tongue contours in the axial can remove the overlap (lines 1–5) and semi-polar parts (lines 6–15) of the grid to the parts that did not surpass the first grid plane in the second linear part of the grid. The trimmed contours were then resampled to have equally spaced points along the contour, such that the half-contours in the axial and the semi-polar parts of the grid each have 18 evenly spread points and those of the frontal part 30. This resulted in an ordered mesh consisting of 420 vertices Dang et al. [13].

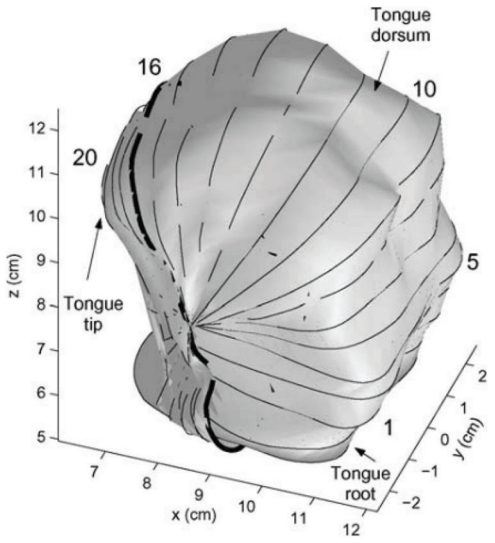


Figure 13: KTH tongue model Olov et al. [11] and the straight lines in the figure represent the position of the plane in this direction

4.2 Comparing Results

According to the above method, to build a three-dimensional shape of the tongue, we first need to define point C, the center of the tongue, and then according to the C, to define the position of the three segmentation sections, which can be seen in Fig. 14.

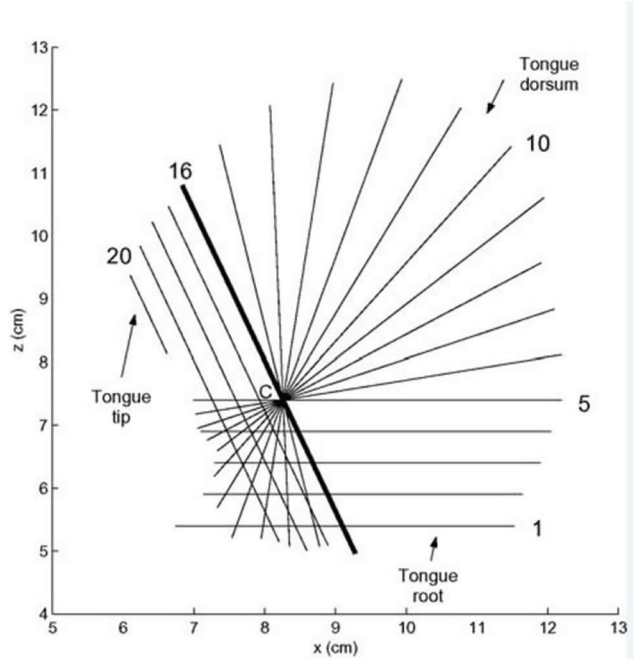


Figure 14: The traditional method to construct the three-dimensional shape of the tongue Olov et al. [11], and the straight lines in the figure represent the position of this plane in the direction

1) Definition of the center

There are two ways to define the position of the center point C. First, according to the position of the jaw, we can determine the center. The jaw will move with different pronunciations, and the corresponding coordinates of the center will change too. Second, choose three fixed points to determine a circle, and the position of the center of the circle is defined as the center. These three points are the tip of the upper teeth, the highest position of the palate and the position of a bone in the uppermost part of the pharynx, respectively. Thus the coordinate of the center is fixed, while the tongue can cause large deformation with the difference in pronunciation. Eventually, the position of the center on the tongue may be particularly up, close to the tongue surface, or close to the root of the tongue, thus making the spacing between planes different. Therefore, we choose the first method as the final definition method.

Description of the method is as follows: choose the intermediate position, P1, of the connection part of the tongue and jaw, a line segment P2P3 which is parallel to the bottom of the root of the tongue, and a level straight line from the position P1, then draw the perpendicular bisector of the line segment P2P3, and the intersection of two straight lines is the center C of the tongue as shown in Fig. 15.

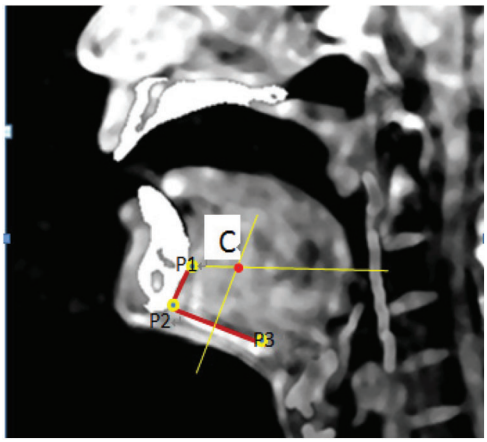


Figure 15: The definition of the tongue center C and the privileged position of P1, P2, and P3

2) Contrast of integrity

Finally, we got the plane as shown on the left in the following figure. Compared with Fig. 15, we found that the position of the two kinds of plane fit generally. In the selected course of the three planes, points beyond the planes were discarded, thus losing some data of the root and front of the tongue. As shown on the right in Fig. 16, the tongue shape, which removes the position beyond the planes (including the data of the front part and the root of the tongue), is presented.

The position of the root of the tongue varies with different pronunciations, so it contributes to the description of the features of the pronunciation to some degree, and the lack of data on the root of the tongue in traditional methods will reduce the contribution rate of description parameters of the tongue.

3) Contrast of correspondence

The stretching scale of the surface and front of the tongue varies with different pronunciations. As shown in Fig. 17. The 3D tongue shape has been displayed when pronouncing /i/ and /u/. The left one is the 3D shape of the tongue when speaking /i/, but the correct response is /u/. It can be clearly seen from the figure that the correct tongue extension is significantly greater than the left (arrow). However, when

considering the degree of tongue extension below the tip of the tongue, the right side is significantly smaller than the left side (circle).

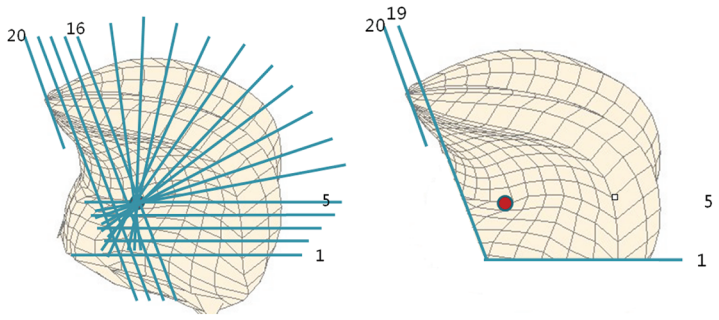


Figure 16: The position of planes constructed for tongue shape using the method already exists (left) and the contrast of this approach has brought the missing part of the tongue (right)

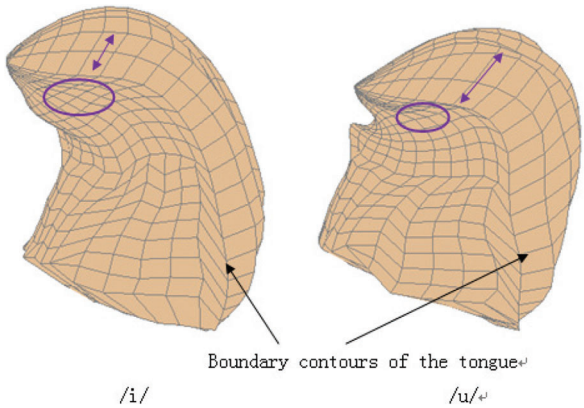


Figure 17: The static three-dimensional shape of the tongue of any vowel /i/ and /u/

In the traditional method, the surface of the tongue and part of the front of the tongue are equally divided as a whole. By default, the deformation between the surface of the tongue and the anterior part of the tongue is the same. Fig. 18 gives the polar plane (left) and the outline of the tongue's surface of the first five planes (right). The distance between the surface and the front of the tongue is accordant, considering that the shrinkage degree in the different pronunciations is the same.

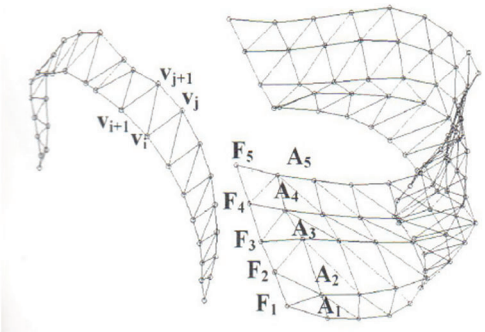


Figure 18: The tongue surface is divided and displayed in KTH model Olov et al. [11], the left is the grid display of the semi-polar plane, and the right is the outline grid of the first five planes

One-dimensional equivalents usually generate the first 3–4 formants. In more extensive frequency ranges, 3D simulation is better in reproducing resonance features. This is especially true when consider the shifting effect. The magnitude response of one-dimensional Kelly-Lochbaum model, the 3D equivalent and the benchmark PSD simulate Jeff's pronunciation vowels /A/Cross-section function is shown in Fig. 19. As is shown in the figure, both the 1D and 3D graphs are the power spectral densities. The relatively simple 1D model is a typical cylinder-like model, in which many spectral details inherent in the complex channel structure are not represented. Although the first four resonance peaks can be identified, there is little similarity between the response above 4 kHz and the measurements. In contrast, when considering the expected downward frequency shift, the PSD obtained from 3D simulation is very close to the measured PSD, which can reach 7 kHz above. Then, the benchmark reliability is questionable because of the time average effect of Welch PSD.

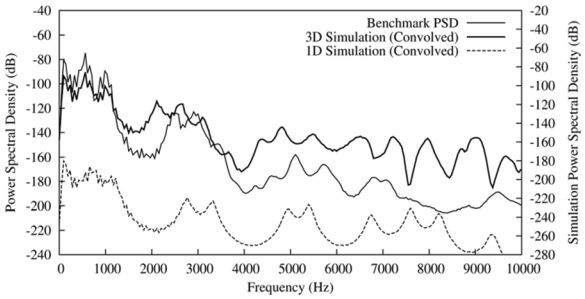


Figure 19: 1D Kelly-Lochbaum models Jeff's pronunciation vowels /A/Cross-section function from vocal tract model, which is compared with full 3D simulation and the Benchmark PSD. Simulations are provided after convolution of appropriate Lx source waveform (The dotted line is 1D simulation, the bold line is 3D simulation, and the thin line is Benchmark PSD)

The two-dimensional impedance mapping simulation is given in the frequency response diagram of Fig. 20, where the 2D and 3D simulations are compared with the Benchmark PSD. As shown in Fig. 19, before measuring Welch PSD, the impulse responses generated by simulation are convolved with Lx source waveform. Similar to 1D simulation, 2D impedance mapping simulation performs best at low frequencies. Although the reproduction of the first two resonance peaks is usually accurate, resonance modes are often affected by unstable frequency drift, which may be caused by the awkward mapping of 3D volume to a single plane. Although 3D simulation provides more accurate geometric simulation, it is limited by static vowel reproduction, while impedance mapping provides a framework for dynamic vowel pronunciation.

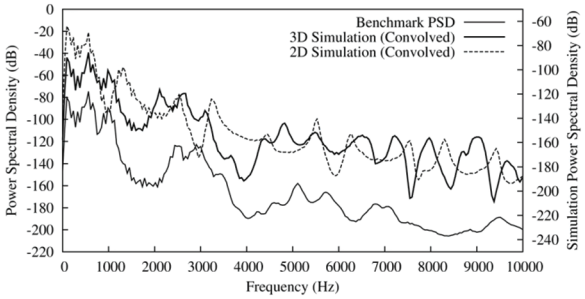


Figure 20: Two-dimensional impedance-mapped simulates Jeff's pronunciation vowels /A/Cross-section function derived from vocal tract model (2D Simulation). This is compared with full 3D simulation (3D simulation) and the acoustic PSD recorded before imaging (Benchmark PSD). Both simulations are shown after convolution with the Lx source waveform (The dotted line is 2D simulation, the bold line is 3D0 simulation, and the thin line is Benchmark PSD)

Therefore, it is essential to determine the characteristic physiological point of the tongue and annotate the parts of different deformation separately to improve the accuracy of the analysis results, enhancing the show of the physical significance of the principle components and representing the different deformation correctly.

5 Conclusions

This paper focuses on the method of modeling based on articulators. We extract the contours of vocal organs such as lips, jaw, tongue, soft palate, and pharynx from the recorded pronunciation database, establish a vocal organ model based on the physiological characteristics of each vocal organ, and quantify the movement and deformation2of the vocal organs. Based on the articulator model, we use statistical methods to analyze the control parameters of the vocal organs. We analyze and model the synergistic relationship between the vocal organs (such as the jaw and tongue, etc.) in the pronunciation process for controlling each articulator model. There are less than three control parameters can be used to describe every speech organ accurately, for which the accumulated contribution rate is more than 88%. By means of the reconfiguration, the average error between the model and accurate data is exactly less than 1.0 mm Warden et al. [22]. This is the first effort to construct a 3D vocal tract model based on Chinese MRI data. It will promote the theoretical research and application of the intelligent Internet of Things Li et al. [23] in speech generation-related issues Zhou et al. [24]. Pandya et al. [25] have shown that noise-robust heartbeat acoustic images are classified using long short-term memory (LSTM)-convolutional neural network (CNN), recurrent neural network (RNN), LSTM, Bi-LSTM, CNN, K-means Clustering, and support vector machine (SVM) methods.

So regarding future work, we will try to construct a device based on our 3D vocal tract model, including Chinese MRI data, but also some other Chinese dialect datasets, which different machine learning models such as CNN, RNN and SVM will be imported into and the results will be compared. What's more, we will develop a system that can show detailed information and results in time on our cell phones.

Funding Statement: This work was supported by the Regional Innovation Cooperation Project of Sichuan Province (Grant No. 2022YFQ0073).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Y. M. Zhang, J. Liu and X. Y. Lin, "Improve neural machine translation by building word vector with part of speech," *Journal on Artificial Intelligence*, vol. 2, no. 2, pp. 79–88, 2020.

[2] Y. Zhao, J. Yue, W. Song, X. X. Li, L. Wu *et al.,* "Tibetan multi-dialect speech recognition using latent regression Bayesian network and end-to-end mode," *Journal of Internet of Things*, vol. 1, no. 1, pp. 17–23, 2019.

[3] S. Nisar, M. Asghar Khan, F. Algarni, A. Wakeel, M. Irfan Uddin *et al*., "Speech recognition-based automated visual acuity testing with adaptive mel filter bank," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2991–3004, 2022.

[4] H. A. Mengash and H. A. H. Mahmoud, "Brain cancer tumor classification from motion-corrected MRI images using convolutional neural network," *Computers, Materials and Continua*, vol. 68, no. 2, pp. 1551–1563, 2021.

[5] Mustaqeem and Soonil Kwon, "1D-CNN: Speech emotion recognition system using a stacked network with dilated CNN features," *Computers, Materials and Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.

[6] Maeda, S. "Compensatory articulation in speech: analysis of x-ray data with an articulatory model," in *Proc. First European Conf. on Speech Communication and Technology (Eurospeech 1989)*, pp. 2441–2445, 1989. http://doi.org/10.21437/Eurospeech.1989-282.

[7] P. Badin and A. Serrurier, "Three-dimensional modeling of speech organs: Articulatory data and models," in *IEICE Technical Report*, vol. 106, no. 177, SP2006–26, pp. 29–34, 2006.

[8] P. Badin, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[9] D. Beautemps, P. Badin and G. Bailly, "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling, 2001," In: M. D. Smith, J. C. Wilcox, T. Kelly, A. K. Knapp (Eds.), *Dominance not Richness Determines Invisibility of Tallgrass Prairie*, Oikos, vol. 106, no. 2, pp. 253–262, 2004.

[10] B. Peter and J. K. Bernd, "Vocal tract model adaptation using magnetic resonance imaging," in *Proc. 7th Int. Seminar on Speech Production*, Belo Horizonte, Brazil, pp. 493–500, 2006.

[11] E. Olov, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Communication, 2003. Tilman D. Causes, Consequences and Ethics of Biodiversity*, Nature, vol. 405, no. 4, pp. 208–211, 2000.

[12] J. M. Gérard, W. T. Reiner, P. Pascal and P. Yohan. "A 3D dynamical biomechanical tongue model to study speech motor control," *Research Developments in Biomechanics*, vol. 1, pp. 49–64, 2003.

[13] J. Dang and K. Honda, "Speech production of vowel sequences using a physiological articulatory model," in *Proc. of ICSLP98*, Toyko, Japan, vol. 5, pp. 1767–1770, 1998.

[14] W. Wang, H. Xu, M. Alazab, T. R. Gadekallu, Z. Han *et al.,* "Blockchain-based reliable and efficient certificateless signature for IIoT devices," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7059–7067, 2022. http://doi.org/10.1109/TII.2021.3084753.

[15] Z. Haddad, M. M. Fouda, M. Mahmoud and M. Abdallah, "Blockchain-based authentication for 5G networks," in *2020 IEEE Int. Conf. on Informatics, IoT, and Enabling Technologies (ICIoT)*, Doha, Qatar, pp. 189–194, 2020. http://doi.org/10.1109/ICIoT48696.2020.9089507.

[16] N. Li, D. Liu and S. Nepal, "Lightweight mutual authentication for IoT and its applications," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 4, pp. 359–370, 2017. http://doi.org/10.1109/TSUSC.2017.2716953.

[17] H. Xiong , C. Jin, M. Alazab, K. H. Yeh, H. Wang *et al.,* "On the design of blockchain-based ECDSA with fault-tolerant batch verification protocol for blockchain-enabled IoMT," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 1977–1986, 2022. https://doi.org/10.1109/JBHI.2021.3112693.

[18] W. Wang, Q. Chen, Z. Yin, G. Srivastava, T. R. Gadekallu *et al.,* "Blockchain and PUF-based lightweight authentication protocol for wireless medical sensor networks," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8883–8891, 2022. https://doi.org/10.1109/JIOT.2021.3117762.

[19] K. Hiraishi, I. Narabayashi, O. Fujita, K. Yamamoto, A. Sagami *et al.,* Blueberry juice: Preliminary evaluation as an oral contrast agent in gastrointestinal MR imaging, *Radiology*, vol. 194, pp. 119–123, 1995.

[20] H. Takemoto, T. Kitamura, H. Nishimoto and K. Honda, "A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions," *Technical Report*, 2004.

[21] Q. Fang, J. Liu, C. Song, J. Wei and W. Lu, "A novel 3D geometric articulatory model," in *The 9th Int. Symp. on Chinese Spoken Language Processing*, Singapore, pp. 368–371, 2014. https://doi.org/10.1109/ISCSLP.2014.6936699.

[22] P. Warden, *"Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition,"* ArXiv e-prints, 2018. [Online]. Available: https://arxiv.org/abs/1804.03209.

[23] H. Li, K. Ota and M. Dong, "Learning iot in edge: Deep learning for the internet of things with edge computing," *IEEE Network*, vol. 32, no. 1, pp. 96–101, 2018.

[24] J. Zhou, Q. Zhang, B. Zhang and X. Chen, "TongueNet: A precise and fast tongue segmentation system using U-net with a morphological processing layer," *Appl. Sci.*, vol. 9, no. 15, pp. 3128, 2019.

[25] S. Pandya, T. R. Gadekallu, P. K. Reddy, W. Wang and M. Alazab, "InfusedHeart: A novel knowledge-infused learning framework for diagnosis of cardiovascular events," *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2022. https://doi.org/10.1109/TCSS.2022.3151643.