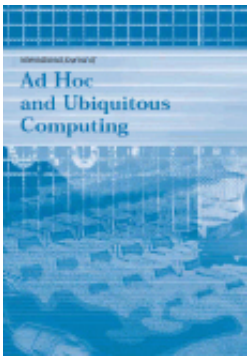


International Journal of Ad Hoc and Ubiquitous Computing

ISSN online: 1743-8233 - ISSN print: 1743-8225
https://www.inderscience.com/ijahuc

Visualisation of Chinese phonemes based on three-dimensional tongue model and ultrasound images

Shaochuan Zhang, Yihuai Zhang, Ping He

DOI: 10.1504/IJAHUC.2026.10076521

Article History:

Received:	28 July 2025
Last revised:	04 January 2026
Accepted:	10 January 2026
Published online:	02 March 2026

Visualisation of Chinese phonemes based on three-dimensional tongue model and ultrasound images

Shaochuan Zhang*, Yihuai Zhang and Ping He

Beihang University,
Beijing, 100083, China
Email: shaoc.zhang@outlook.com
Email: zhangyh6171117@163.com
Email: heppping@163.com
*Corresponding author

Abstract: This study introduces a high-precision 3D tongue motion visualisation model using ultrasound imaging to aid pronunciation training for hearing-impaired individuals and Chinese learners. Unlike static 2D images or generic 3D representations, this approach reconstructs authentic tongue motion by integrating 2D ultrasound data with a high-fidelity 3D model. It provides side-by-side 2D/3D comparisons for commonly confused phonemes and evaluates accuracy using curve-similarity metrics instead of point-error measures. Results show 91.7% reconstruction accuracy, capturing subtle dynamic tongue movements during pronunciation. The model dynamically and accurately visualises articulation, offering a more precise tool for pronunciation training in special education and second language acquisition.

Keyword: tongue motion visualisation; ultrasound imaging; pronunciation training; curve-similarity; statistic model.

Reference to this paper should be made as follows: Zhang, S., Zhang, Y. and He, P. (2026) 'Visualisation of Chinese phonemes based on three-dimensional tongue model and ultrasound images', *Int. J. Ad Hoc and Ubiquitous Computing*, Vol. 51, No. 5, pp.1–11.

Biographical notes: Shaochuan Zhang received her BS in Biomedical Engineering from Shanghai University in 2015. She began her Master's and PhD combined program at the School of Biological Science and Medical Engineering, Beihang University, in 2016, and graduated in 2024. Her research focuses on biomedical signal processing and speech rehabilitation.

Yihuai Zhang received his BS from Yanshan University in 2022. He is currently a PhD candidate focusing on medical physiological signal analysis and processing.

Ping He received his BS in Biomedical Engineering from Nanjing University of Aeronautics and Astronautics in 2023. He is currently a Master candidate focusing on the efficient numerical simulation of blood flow.

1 Introduction

Multimodal speech cognition research has shown that visual cues significantly influence speech perception. Early studies observed that listeners tend to gaze at the speaker's face during conversation (Dodd, 1977; Liberman and Mattingly, 1985; Sumby and Pollack, 1954; Summerfield, 1987), and the McGurk effect highlighted the important role of visual cues in speech comprehension (Macdonald and McGurk, 1978; McGurk and MacDonald, 1976). For individuals with hearing impairments, integrating visual information such as mouth movements and facial gestures markedly enhances spoken language understanding (Bernstein et al., 2004; Marinetti et al., 2011). In deaf education, traditional pedagogy relies on students observing the instructor's articulatory movements and facial expressions while mimicking their speech, with teachers providing corrective feedback (Easterbrooks and Baker, 2002; Massaro and Bosseler, 2002). However, this teacher-centric approach is limited by the instructor's expertise, high demand for qualified teachers, and dependence on direct guidance, which hinders independent learning. According to existing empirical studies, the visualisation of articulatory organ movements has a significant facilitative effect on language acquisition in deaf-mute individuals (Gibson and Lee, 2021; Al Ani et al., 2025).

Advancements in computer technology have enabled visualisation of articulatory movements (Olson, 2022; Revina, 2019). In Chinese language learning, videos and animations of mouth shape changes have been used to aid pronunciation imitation, but their effectiveness is limited as the tongue remains largely invisible (Chen, 2012; Ye, 2014). Some studies employed sound waves and spectrograms to illustrate frequency and amplitude

variations, assisting deaf children in understanding tone and articulation, but these methods require substantial prior speech knowledge and are not child-friendly (Johnson, 2022). More recently, visualisation of tongue movements has gained attention. For English, Wang Lan's team at the Shenzhen Institute of advanced technology developed a speech rehabilitation system using electromagnetic articulography (EMA) data to synthesise tongue movements (Wang et al., 2009, 2012). The Speaker Baldi system, developed by Massaro's team at the University of California, Santa Cruz, demonstrated improved speech comprehension and pronunciation in deaf children (Massaro, 1998, 2003; Massaro and Bosseler, 2002). For Swedish, Engwall's team at the Royal Institute of Technology created the ARTUR virtual speaker, modelling 43 tongue shapes (Massaro and Light, 2003, 2004; Engwall and Badin, 1999). For French, Badin's team developed the ATH virtual articulatory model, supporting speech rehabilitation (Engwall, 2002, 2012). However, research on Chinese remains limited. The web-based IPA visualisation system by the University of Science and Technology of China approximates tongue movements with a surface model, reducing the authenticity of pronunciation synthesis (Engwall, 2000, 2004; Badin et al., 2008). Initially, Tianjin University's 3D visualisation of Chinese vowel phonemes was static, lacking dynamic pronunciation processes. Subsequent EMA-driven tongue models, while enabling dynamic visualisation of articulatory motion, relied on only three tongue surface points, resulting in limited accuracy (Serrurier and Badin, 2008; Yu and Wang, 2015).

To overcome the limited model fidelity and inadequate three-dimensional (3D) motion accuracy of existing VR-based pronunciation training systems, tongue movement trajectories for selected phonemes in a specialised corpus were reconstructed and visualised by integrating authentic 2D ultrasound motion sequences with a high-precision statistical tongue model (Yu et al., 2013; Liu et al., 2013). By manually tuning every control parameter, the tongue model's mid-sagittal plane was aligned with two-dimensional ultrasound tongue contours, generating a 3D model that closely corresponds to the ultrasound data. The accuracy of these two-dimensional to 3D alignments was validated using a curve similarity metric. Analysis of tongue position variations across different phonemes was conducted, yielding results that offer valuable guidance for individuals with hearing impairments and learners of Chinese as a second language.

The following is the structure of this article: Section 2 elaborates on the construction of the database and the mapping methodology between ultrasound images and 3D models; Section 3 presents the similarity results between ultrasound tongue contours and the corresponding 3D models; Section 4 analyses tongue position variations across different phonemes based on the modelled mid-sagittal plane; Section 5 discusses the research findings and directions for future research.

Table 1 Corpus

	Consonant – vowel		
	a	i	u
b	ba	bi	bu
p	pa	pi	pu
d	da	di	du
t	ta	ti	tu
m	ma	mi	mu
n	na	ni	nu
f	fa		fu
l	la	li	lu
g	ga		gu
k	ka		ku
h	ha		hu
j		ji	
q		qi	
x		xi	
zh	zha	zhi	zhu
ch	cha	chi	chu
sh	sha	shi	shu
r		ri	ru
z	za	zi	zu
c	ca	ci	cu
s	sa	si	su

Table 2 Phoneme classification

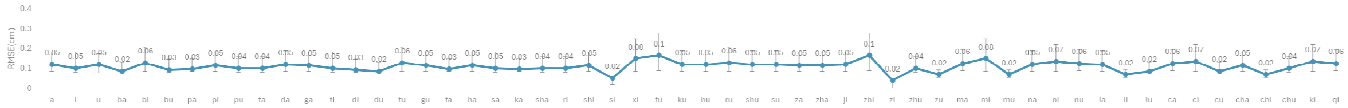
Phonemes	
Vowels	/a/, /i/, /u/
Plosives	/ba/, /pa/, /ta/, /da/, /ga/, /ti/, /di/, /tu/, /du/, /gu/, /ka/, /ku/
Fricatives	/fa/, /ha/, /sa/, /sha/, /shi/, /si/, /fu/, /hu/, /shu/, /su/, /za/, /xi/
Affricates	/zha/, /zhi/, /zi/, /zhu/, /zu/, /ca/, /ci/, /cu/, /cha/, /chi/, /chu/, /ji/, /qi/
Nasals	/ma/, /mi/, /mu/, /na/, /ni/, /nu/
Laterals	/la/, /li/, /lu/, /ri/, /ru/

2 Materials and methods

2.1 Corpus

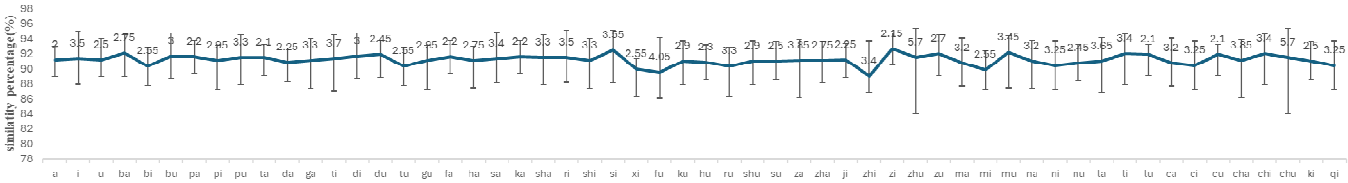
To comprehensively represent tongue morphology variations during Mandarin pronunciation, a corpus was constructed comprising three representative vowels and 52 consonant-vowel combinations (Table 1). This corpus was designed to encompass the primary phonemes and associated tongue shape changes in Mandarin Chinese. Phonemes were categorised by articulation manner into plosives, fricatives, affricates, nasals, and laterals (Table 2).

Figure 1 The mean RMSE of each phoneme between ultrasound image and the corresponding tongue model (see online version for colours)



Notes: The median of the error is indicated/labelled in the figure.

Figure 2 The average similarity degree between each phoneme ultrasound image and the corresponding tongue model (see online version for colours)



Note: The median of the error is indicated/labelled.

2.2 Ultrasonic image acquisition and contour extraction

The data-acquisition system comprised two synergistic modules: ultrasound imaging and acoustic recording. Ultrasound videos were captured using a Clover B-mode scanner (Wisonic, China) in conjunction with a GC573 frame-grabber card (AVerMedia, China) and a 4.5 MHz curved transducer. Probe-to-jaw coupling was maintained via an Ultrafit stabilisation headset (Articulate Instruments, UK). Acoustic signals were recorded by an ECM8000 condenser microphone (Behringer, Germany) routed through a Quad-Capture audio interface (Roland, Japan). Both ultrasound and audio streams were synchronised by a time triggered coordination scheme. All data were collected in a sound-proof studio from a healthy, native Mandarin speaker with no speech or hearing deficits. The transducer was positioned against the submandibular region to acquire midsagittal tongue images at 100 fps and a resolution of 1920 × 1080 pixels, while audio was sampled at 44.1 kHz.

Data collection followed a structured protocol. The subject pronounced sequences in an 'a-consonant-vowel' format, articulating the vowel /a/ followed by a consonant-vowel combination, with a 1-second pause between each sequence. This process was repeated until all 55 sequences in the corpus were completed. The continuous video stream was segmented into 55 one-second clips, each corresponding to a corpus sequence. Ultrasound image sequences were obtained through frame sampling, and synchronised speech signals were extracted in the time domain.

Given the VR-based tongue modelling focus, the canonical 3D tongue model template was derived from a single subject's data. In practical deployment, individual tongue geometries can be mapped onto this template through a straightforward registration algorithm.

In the ultrasound database, the bright curves observed in the images represent the tongue surface contour. Due to the limited reliability of current automatic tongue contour extraction methods, manual annotation was employed to ensure accuracy (Figure 1). For each tongue contour, 15 points were manually annotated on the tongue surface, followed by spline interpolation to generate 100 points for a detailed representation. To minimise annotation variability, two annotators with expertise in ultrasound imaging independently performed the annotations twice. Inter-annotator and intra-annotator errors (unit: pixels) were calculated, with results presented in Table 3. Ultimately, the mean of the two annotation sets was adopted as the tongue contour data. For practical purposes, a 60-frame interval was used for each phoneme, with 60 images representing the articulatory motion process of each phoneme.

Table 3 Inter-annotator and intra-annotator errors

	Laber1	*Laber2*
Laber1	2.32	2.75
Laber2	2.58	2.40

2.3 3D tongue model

This study utilised a 3D tongue statistical model developed based on prior work (Zhang et al., 2024). The statistical model was built by adapting the framework of to extract control parameters through a regulated combination of principal component analysis (PCA) and multiple linear regression (MLR). This approach is grounded in two fundamental hypotheses: first, that the control parameters function independently of one another; and second, that there is a linear relationship between changes in the organ's morphology and the control parameters. On the basis of these two assumptions, The tongue shape vectors (DV) can be represented as linear combinations of associated coefficients (AC), each modulated by its respective control factor (CF), apart from the mean neutral shape $(\overline{DV})$. The specific computation method is presented in equation (1).

$$DV = \overline{DV} + CF * AC \tag{1}$$

In the (1), $\overline{DV}$ is the average shape of all static tongue models of the 49 articulations of the corpus, while *CF* and AC are determined iteratively in the following way:

1 the *CF* is mainly determined using PCA

2 the associated coefficient *AC* is determined by the linear regression of the current residue data for the whole corpus over *CF*

3 the contribution of the *CF* refers to its product with the associated matrix *AC*, and is finally subtracted from the current residue in order to provide the next residue for determining the next component, to ensure independence between control parameters.

Using the aforementioned modelling method, parameters were extracted by taking the vertices of the tongue's mid-sagittal contour as variables. Due to the physiological connection between the tongue and the mandible, it is difficult to distinguish tongue movements caused by the mandible from those initiated by the tongue itself. In this study, the height difference between the upper and lower incisors-standardised by mean and variance-was defined as the first parameter, namely the jaw height parameter (JH). The corresponding relationship matrix was obtained via MLR between the vertex coordinates of the 49 tongue 3D models and JH. After removing the contribution of JH, PCA was performed on the tongue body and tongue dorsum regions. The first two principal components were designated as the second and third control parameters, namely the tongue body parameter (TB) and tongue dorsum parameter (TD), respectively. The coefficient matrices for these two parameters were also derived using MLR. Similarly, after regressing out the effects of TB and TD, PCA was applied to the tongue tip region, resulting in three parameters: tongue tip vertical (TTV), tongue tip horizontal (TTH), and tongue tip upturned (TTU). Another mandibular degree of freedom, jaw width, was omitted as its contribution to tongue movement was negligible and could be captured by the TTH parameter. The tongue motion simulation in this study was implemented by combining the aforementioned six control parameters.

Based on equation (1), different tongue shapes were generated by adjusting the values of each parameter according to their roles in controlling tongue movement. Specifically, JH controls the rotation angle of the tongue around a point on the dorsum; TB governs the extent of anterior protrusion and posterior retraction of the tongue body; TD determines the degree of upward arching and downward concavity of the tongue dorsum. TTH and TTV control the vertical and horizontal movements of the tongue tip, while TTU controls the degree of upturning of the tongue tip around a point near its front. During the parameter extraction process, the control parameter values and their corresponding two-dimensional coefficient matrices for the 49 phonemes in the corpus were obtained. By employing MLR, the linear relationships between the six control parameters and all vertices on the 3D tongue mesh were computed, resulting in a control relationship matrix for overall tongue movements in three dimensions.

2.4 Method

Although 3D tongue models can simulate all possible tongue configurations during speech, the correspondence between 3D shapes and phonemes remains unclear. Consequently, 3D tongue models were derived from ultrasound images collected during actual articulation to capture the 3D motion trajectories of phonemes. To establish the relationship between 2D ultrasound images and 3D tongue models, the 2D-to-3D matching was simplified to a 2D-to-2D contour matching process, focusing on the similarity between the tongue contour and the mid-sagittal plane of the model. This matching was implemented using MATLAB software, with a visualisation interface developed for the process. Within this interface, the upper surface contour of the model's mid-sagittal plane can be manually adjusted by tuning control parameters to align closely with the ultrasound tongue contour. As the model data and ultrasound images were obtained from the same subject, normalisation was not required. However, since the model's mid-sagittal plane includes both the tongue's upper surface and the tongue root region, the overlapping segment between the mid-sagittal plane and the ultrasound tongue contour was first identified. Using a specific frame of the reference vowel /a/ as a baseline, the ultrasound tongue contour was aligned with the upper surface of the model's mid-sagittal plane through a unified coordinate system and translation. Based on clockwise annotation of the tongue contour points, points 13 to 28 of the model's mid-sagittal plane were found to correspond exactly with the ultrasound tongue contour, and the coordinates of points 12 to 27 were defined as the model contour. To assess the similarity of curve shapes between modalities, the ultrasound images and the tongue model were unified to the same coordinate space. The original coordinate system for the tongue contour extracted from ultrasound images differs from that of the tongue model, necessitating translation and inversion operations. The new y-coordinate is computed as $y_{new} = height – y_{ori}$ where *height* denotes the ultrasound image height and y_{ori} represents the original y-coordinate. Due to differences in the number of points and coordinate references between the ultrasound tongue outline and the model contour, conventional distance metrics cannot accurately quantify their errors. To enhance the similarity evaluation between ultrasound and model contours, in addition to adopting the traditional distance error, a curve similarity metric (Eiter and Mannila, 1994) was also employed as an evaluation index to compute the degree of overlap between the ultrasound and model tongue contours. The specific formula is presented in the equation (1).

$$\overline{Sim} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-d_{frechet}(c_1,c_2)}{curv_{mean}(c_1,c_2)}\right) \quad (1)$$

where n denotes the number of ultrasound images for each tone; c_1 denotes the model tongue contour (x_i, y_i), i = 1, 2,

…, 15; and c_2 denotes the ultrasound tongue contour (u_i, v_i), $i = 1, 2, \ldots, 100$. $d_{frechet}$ denotes the normalised two-curve c_1 and c_2 Frechet distance. $curv_{mean} = \sqrt{d_{c1}d_{c2}}$ denotes the average distance between the two curves.

3 Results

3.1 The results of all phonemes' distances difference between model and ultrasound images

Using a pixel-based distance metric, the distance differences between the ultrasound tongue contours of each phoneme and the manually matched model tongue contours were computed. Specifically, the distance difference, measured in pixels, was calculated for each frame's ultrasound tongue contour against the model tongue contour, and the average distance difference was determined for each phoneme. The mean RMSE values for each phoneme are presented in Figure 1, with error bars indicating the maximum range of distance differences. The overall mean distance difference across all phonemes was approximately 5.96 pixels, corresponding to an error of 0.149 cm at a spatial resolution of 0.25 mm/pixel.

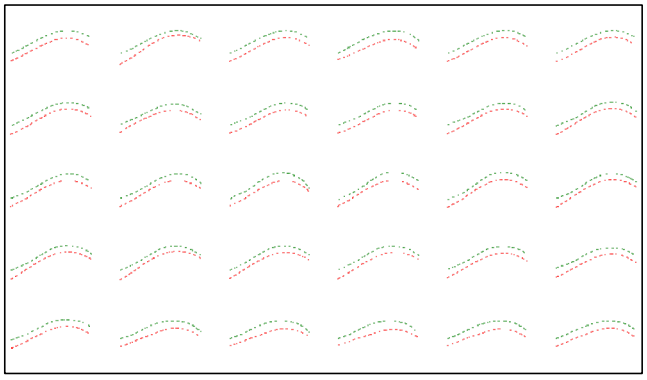
3.2 The results of all phonemes' shape similarity between model and ultrasound

Using the Frechet distance-based curve similarity metric (Eiter and Mannila, 1994), the similarity between the ultrasound tongue contours of each phoneme and the manually matched model tongue contours was computed. Specifically, the curve similarity value was calculated for each frame's ultrasound tongue contour against the model tongue contour, and the average $\overline{sim}$ was determined for each phoneme. The mean $\overline{sim}$ values for each phoneme are presented in Figure 2, with error bars indicating the maximum range of $\overline{sim}$. In this study, the average similarity across all types of phonemes reached 91.7%(±0.64%), reflecting the overall accuracy of the high-fidelity 3D tongue model visualisation library in capturing the dynamic features of articulation. The average similarity for plosives was 91.94%, demonstrating high mapping accuracy; for fricatives, it was 91.44%, indicating stable performance; for affricates, it achieved 91.99%, the best among all phoneme categories. Nasals showed 91.32%, which, while slightly lower than other categories, still indicates high precision. Laterals reached 91.81%, suggesting that the model effectively captures their dynamic features. For vowels, the similarity was 91.72%, demonstrating the model's strong adaptability to simpler articulation structures. These results indicate that the model achieves high accuracy across different phoneme categories, providing reliable technical support for language learning and rehabilitation training. Figure 3 illustrates the contour variations of the /ba/ sound in ultrasound images and the corresponding tongue model across dynamic frames.

3.3 Comparison with results from other methods

Regarding the extraction of model control parameters from ultrasound images, Table 4 compares this neural network-based approach with current popular methods in terms of accuracy of synthesised control parameters. The proposed direct mapping methodology demonstrates competitive performance with only slightly higher root mean square error (RMSE) than the global mixture regression (GMR) method, while achieving superior correlation coefficients. Importantly, unlike the GMR method which relies on EMA technology as an intermediary step, the proposed approach directly get control parameters from ultrasound images, eliminating the need for additional sensing modalities. This end-to-end framework reduces methodological complexity and equipment requirements while maintaining accuracy.

Figure 3 Contour comparison for the /ba/ sound across frames (see online version for colours)



Notes: Frame 1 to Frame 60, sampled at 1-frame intervals, from top-left to bottom-right; green: ultrasound contour; red: model contour.

Table 4 Comparison of mapping methods

GMR (*Fabre et al.,* 2017)			*Own method*		
Parameter	*RMS*	R^2	*Parameter*	*RMSE*	R^2
tiph	2.3	0.89	JH	1.78	0.95
tipv	2.0	0.85	TB	4.37	0.96
midh	1.9	0.89	TD	2.53	0.97
midv	1.7	0.9	TTV	4.32	0.96
backh	1.6	0.94	TTH	2.35	0.95
backv	2.2	0.79	TTU	3.48	0.95

4 Discussion

4.1 Static tongue analysis

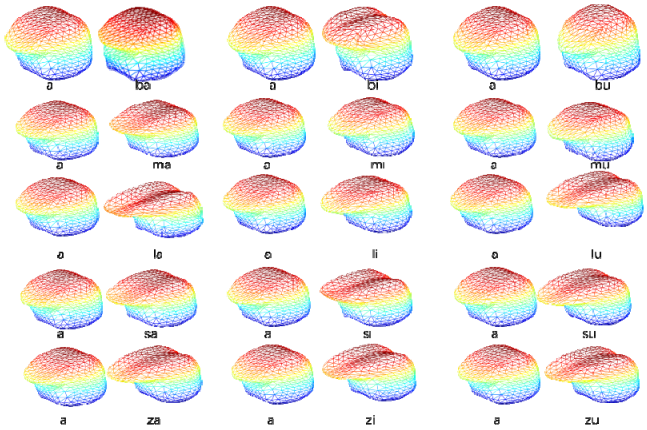
The tongue morphology of different types of phonemes at the moment of articulation into a stop (the moment when the change in tongue position is most pronounced) was compared with the onset sound a (Figure 4), and the change in tongue position was observed. In the stops /ba/, /bi/, /bu/, there is little difference concerning the vertical placement of the tongue between /ba/ and /a/, but the position is pulled

back; while tongue tip is moved forward in the /bi/ sound compared to the /a/ sound, the tongue position of the vowel /i/ itself is higher than that of the /a/ sound; and the tongue position of the /bu/ sound is shifted back and lifted up compared to the /a/ sound, due to the fact that the tongue position of the /u/ sound is higher and more anteriorly influenced than that of the /a/. The tongue morphology of /ba/, /bi/, /bu/ is shown in the following table. A side-by-side comparison of /ba/, /bi/, /bu/ reveals that the order of tongue position is /bu/, /ba/, /bi/, which is different from the real order of /bi/, /bu/, /ba/. The reason for this is that the tongue model obtained based on the ultrasound image has a defect, and the reconstruction of the tip of the tongue is poor, which is due to the fact that the tip of the tongue is not imaged due to the obstruction of the mandible when the tongue is imaged by the ultrasound, which leads to a poor reconstruction of the tongue tip. The high tongue position of the /bi/ sound is mainly based on the proximity of the tongue tip to the mandible in the anterior part of the tongue, while the tongue position based on the middle and posterior part of the tongue is indeed /bu/, /ba/, /bi/, which is still a guide to the pronunciation of the sound, and the subsequent analysis of tongue position is based on the middle and posterior part of the tongue by default and does not include the tongue tip. For the nasal sounds /ma/,/mi/,/mu/, the tongue position of /ma/ and /a/ is lowered, with a slight tip of the tongue; the tongue position of /mi/ is raised compared to that of /a/, with a forward tilt of the tongue; and the tongue position of /mu/ is raised compared to that of /a/, with a slight backward shift of the tongue. A side-by-side comparison of /ma/, /mi/, and /mu/ revealed that mi had the highest tongue position, followed by /mu/, and finally /ma/. In the fricatives /sa/, /si/, and /su/, /sa/ had a lowered tongue position and a cocked tip when compared to the /a/ sound, /si/ had a lowered and flattened tongue position with a slightly cocked tip when compared to the /a/ sound, and /su had an elevated posterior portion of the tongue position with a slightly lowered tip when compared to the /a/-sound. A side-by-side comparison of /sa/, /si/, /su/ based on tongue position in the central and posterior regions demonstrated that /sa/ had the highest tongue position, followed by /su/, and finally /si/. The /za/ sounds in the fricatives /za/, /zi/, /zu/ were both lowered and forwardly tilted compared to the /a/ sounds due to the influence of /z/. In a side-by-side comparison of /za/, /zi/, /zu/, the /za/ sound had the highest tongue position, followed by /zu/, and finally /zi/.

4.2 Dynamic tongue position analysis

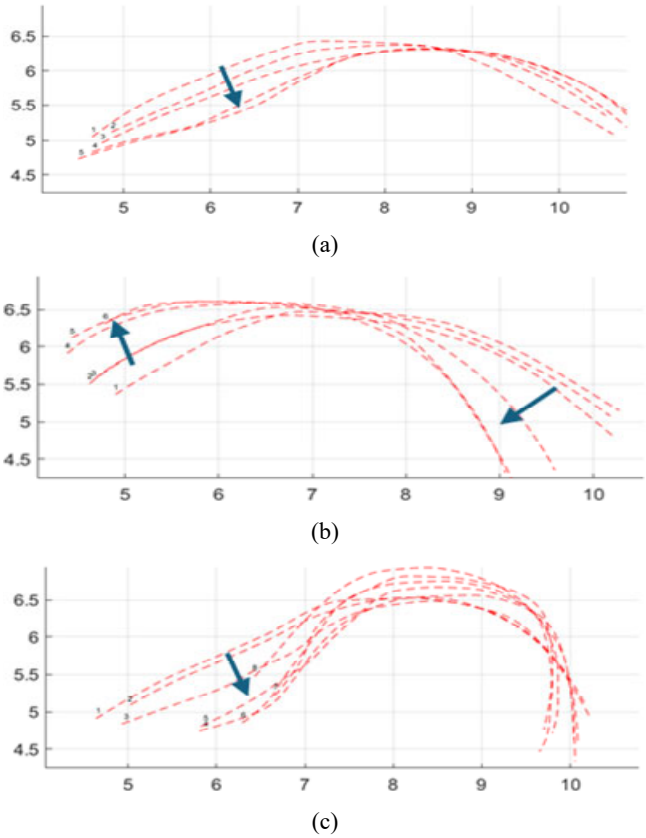
The trajectories of the tongue median sagittal plane movements from the resting state to /a/,/i/,/u/ are shown in Figure 5. For the /a/ sound, compared with the resting tongue position, the anterior part of the tongue is moved down, the tongue body is slightly depressed, and the tongue body is shifted backward; for the /i/ sound when the articulation is stabilised, the tongue tip is lifted high, the tongue body is advanced, and the overall tongue position is elevated; for the /u/ sound when the articulation is stabilised, compared with the resting state, the tongue is contracted backward, the tip is lowered, and the tongue position is elevated. When comparing the tongue position of the three representative phonemes /a/, /i/, /u/ horizontally, there is not much difference between the tongue position of /i/ and /u/ sounds, and the /a/ sound has the lowest tongue position, and at the same time, it can be seen that the /u/ sound shifts backward to belong to a back vowel, the /i/ sound shifts forward to belong to a front vowel, and the /a/ sound has a centralised tongue position to belong to a central vowel. Statistical analysis is conducted on the y-coordinate data of the contours, and ANOVA is employed to assess the significance of differences at each point for /a/, /i/, and /u/. The p-value for each point is determined to be less than 0.05, with the average p-value across all points calculated as 0.0083.

Figure 4 Comparison of the three-dimensional tongue shapes during pre-pronunciation of different phonemes with that of the sound /a/ (see online version for colours)



The articulatory movement trajectories from the /a/ sound to /za/, /ca/, /sa/ in the apical prosodic /za/, /ca/, /sa/ are shown in Figure 6, and it is observed that the apical prosodic sounds have an overall tendency for the tongue front to rise and the tongue body to fall; among them, the tongue body of the /za/ sound falls significantly, and that of the /ca/ and /sa/ sounds falls slightly, but all of them have a greater elevation of the tongue front. A lateral comparison of the apical consonants shows that there is no obvious difference in the overall tongue position, with the /sa/ sound having a slightly higher tongue body and a slightly lower tongue tip, but the difference is almost negligible, so it is difficult to distinguish the differences between different apical consonants from observation of the tongue position alone, but it is a guide to pronunciation by the obvious change in the tongue position from the reference sound /a/ to the apical consonants. Statistical analysis is conducted on the y-coordinate data of the contours, and ANOVA is applied to evaluate the significance of differences at each point for /za/, /ca/, and /sa/. Based on the p-values of all points, an average value of 0.6878 is calculated. This suggests that distinguishing the vertical differences among the three phonemes through tongue contours is challenging.

Figure 5 (a)–(c) The midline movement trajectories of the tongue for /a/, /i/, /u/ from left to right (see online version for colours)



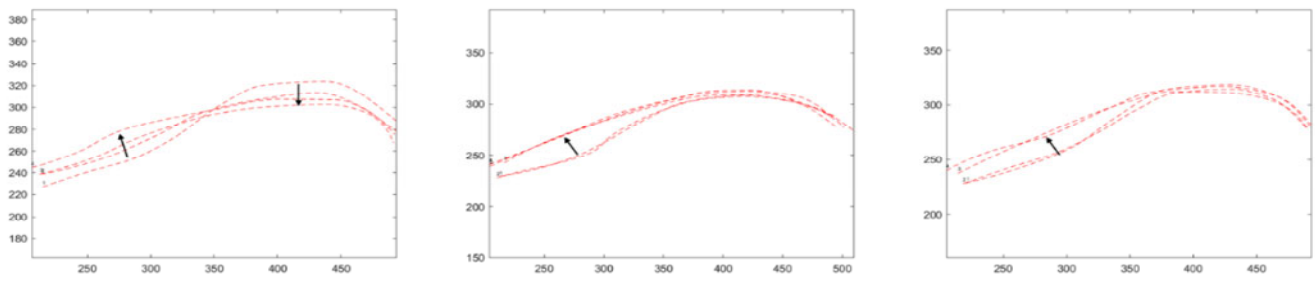
Notes: arrows indicate the direction of movement; the scale of the horizontal and vertical axes is in centimeters (cm), representing the size dimensions of the tongue model.

The trajectories of the apical medial consonants /da/, /ta/, /na/, /la/ from the reference sound a to /da/, /ta/, /na/, /la/ are shown in Figure 7. The lifting of the tongue tip for the da was not accompanied by a decrease in the height of the tongue body; the lifting of the tongue tip for the ta was accompanied by the forward extension of the tongue body, resulting in a slight decrease in the tongue body's lingual position; and the tips of the tongues of the sounds /na/ and /la/ were lifted but the tongue body was in an overall low position, with a significant decrease in the lingual body. A side-by-side comparison of the tongue-tip medial consonants /d/, /t/, /n/, /l/ showed that the overall tongue position of the tongue body was low, rising from /la/, /na/, /ta/, /da/, where the differences in the tongue positions of /ta/ and /da/ were not significant, but were generally higher than those of the /la/ and /na/ consonants. Statistical analysis is conducted on the y-coordinate data of the contours, and ANOVA is employed to assess the significance of differences at each point for /da/, /ta/, /na/, and /la/. Based on the p-values of all points, an average value of 0.5730 is calculated. This indicates that the overall tongue contours exhibit similarity.

The articulatory motion trajectories of the apical velar sounds /zhi/, /chi/, /shi/, /ri/ from the a sound to /zhi/, /chi/, /shi/, /ri/ are shown in Figure 8, and it is found that the apical velar sounds have a similar change in tongue position, with the overall tongue position dropping and the tongue tip lifting up, but the degree of tongue position dropping varies among different sounds, with /zhi/, /shi/ having a relatively large drop in the tongue position; /chi/ and /ri/ have a relatively small The degree of decline is different for different sounds. Cross-sectional comparisons of the postalveolar consonants show that /zh/ has the highest tongue position and anterior extension of the tongue body compared to the other postalveolar consonants, while /sh/ has the lowest tongue position and posterior movement of the tongue body. The /ch/ and /r/ consonants do not differ much and are difficult to distinguish from each other in terms of tongue position. Statistical analysis is conducted on the y-coordinate data of the contours, and ANOVA is utilised to evaluate the significance of differences at each point for /zhi/, /chi/, /shi/, and /ri/. Based on the p-values of all points, an average value of 0.6524 is calculated.

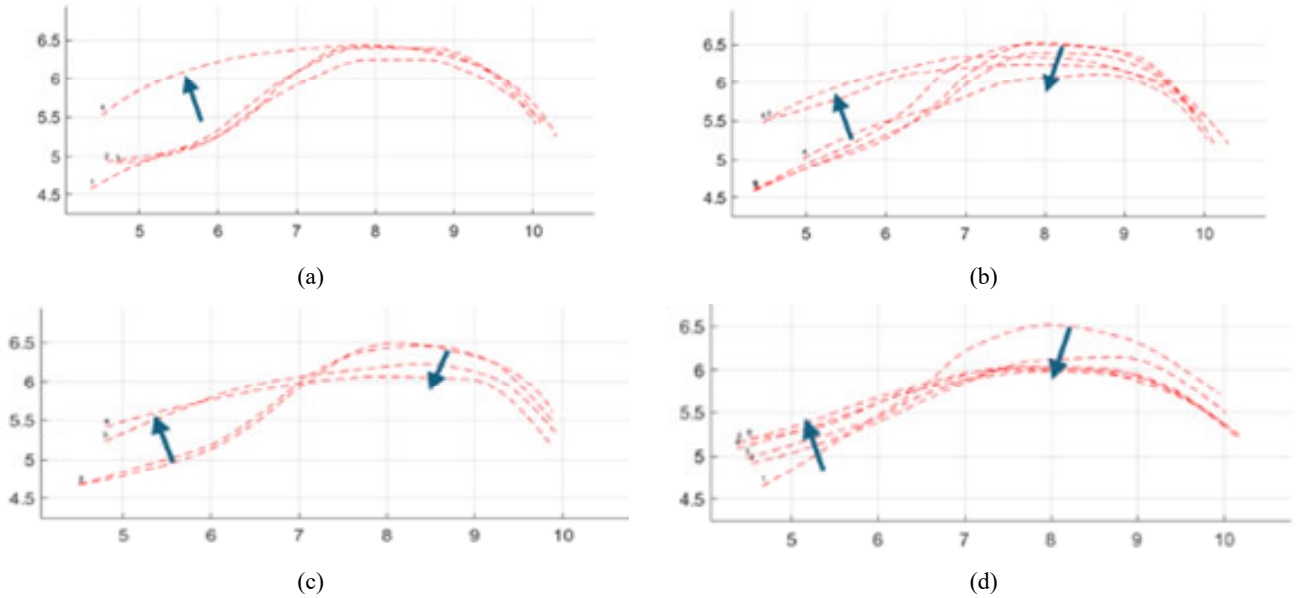
The trajectory of the tongue-facing sounds /ji/, /qi/, /xi/ from the reference sound a to /ji/, /qi/, /xi/ is shown in Figure 9. The tongue position of the tongue-face sound remains the same overall, but the tongue body is tilted forward, making the tongue position go from a backward to a central position while the tongue tip is raised. A side-by-side comparison of the tongue-side consonants indicate that the overall tongue position does not differ. Statistical analysis is conducted on the y-coordinate data of the contours, and ANOVA is employed to assess the significance of differences at each point for /ji/, /qi/, and /xi/. Based on the p-values of all points, an average value of 0.7337 is calculated.

Figure 6 (a)–(c) The midline movement trajectories of the tongue for /za/, /ca/, /sa/ from left to right (see online version for colours)



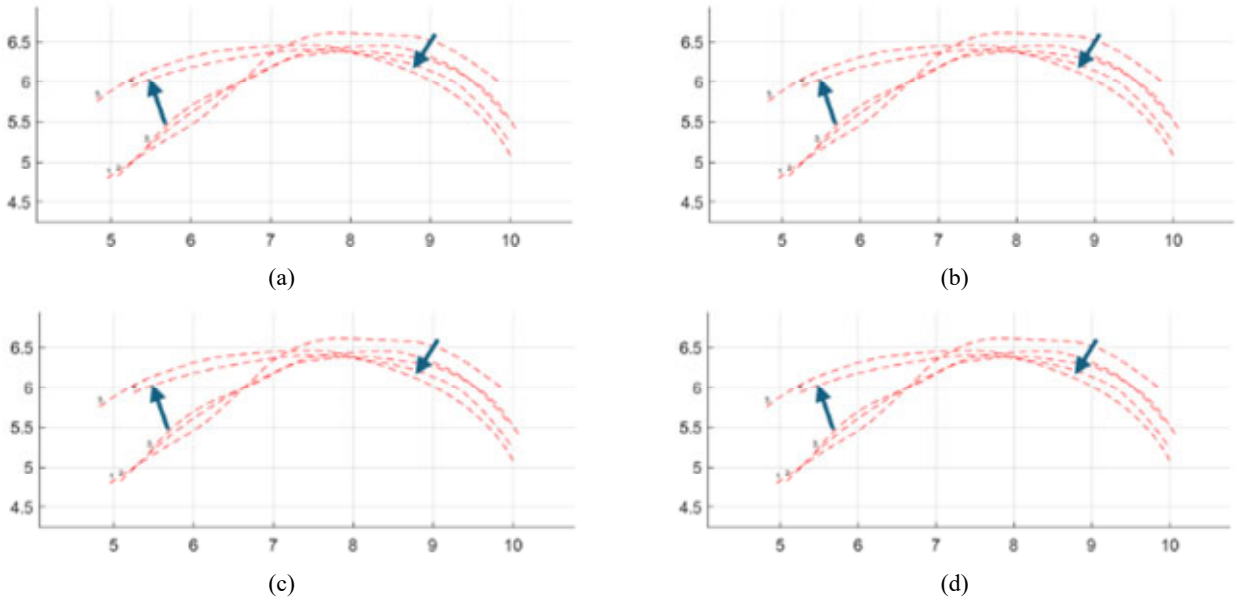
Notes: arrows indicate the direction of movement; the scale of the horizontal and vertical axes is in centimeters (cm), representing the size dimensions of the tongue model.

Figure 7 (a)–(d) Trajectories of the tongue median plane from upper left to lower right /da/, /ta/, /na/, /la/ (see online version for colours)



Notes: Arrows indicate direction of motion; the scale of the horizontal and vertical axes is in centimeters (cm), representing the size dimensions of the tongue model.

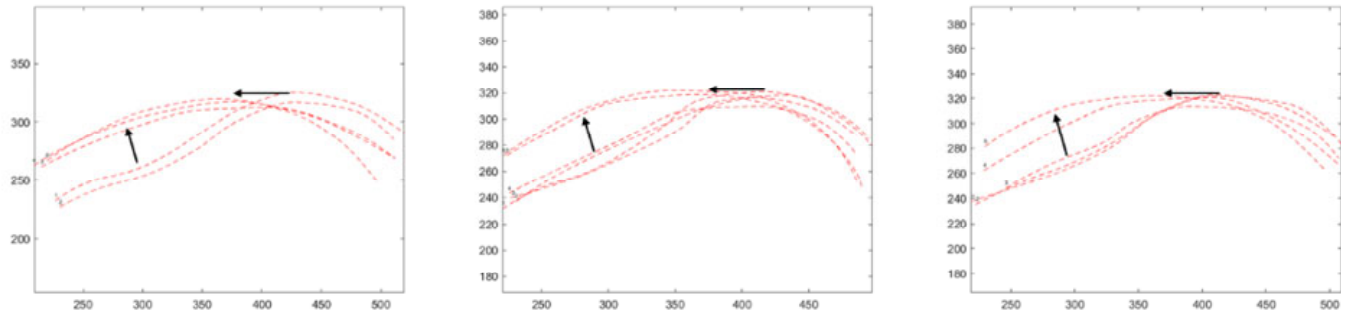
Figure 8 (a)–(d) Trajectories of the tongue median plane from upper left to lower right /zhi/, /chi/, /shi/, /ri/ (see online version for colours)



Notes: Arrows indicate direction of motion; the scale of the horizontal and vertical axes is in centimeters (cm), representing the size dimensions of the tongue model.

The changes in tongue position for the tongue-following sounds /ga/, /ka/, /ha/ from the reference sound /a/ to /ga/, /ka/, and /ha/ are shown in Figure 10. Among them, the tongue tip is found to rise and tongue body to fall when transitioning from the articulatory /a/ to that of the /ga/, indicating that the tongue body has an overall tendency to extend forward; when going from the articulatory /a/ to the /ka/ sound, the same tip of the tongue is found to rise and the tongue body to fall, and the tongue body has an overall tendency to extend forward, making it difficult to distinguish between the /ga/ and /ka/ sounds only by the change in the tongue position; there is not a great change of the tongue position from the articulatory /a/ to the /ha/ sound, and it only has a slight tendency to move backward, which is almost This is almost negligible. A side-by-side comparison of the root sounds /ga/, /ka/, and /ha/ shows that /ga/ is slightly higher than /ka/ and has the highest tongue position, while the /ha/ sound has the lowest tongue position and a posterior tongue body. Statistical analysis is conducted on the y-coordinate data of the contours, and ANOVA is employed to evaluate the significance of differences at each point for /ga/, /ka/, and /ha/. Based on the p-values of all points, an average value of 0.3647 is calculated.

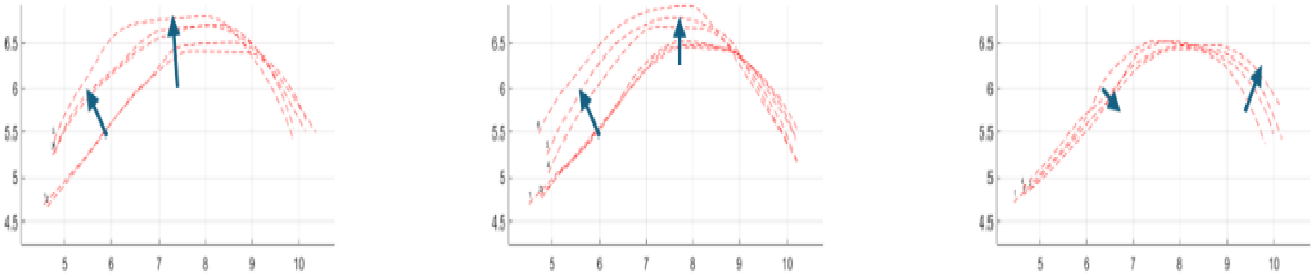
Figure 9 (a)–(c) The midline movement trajectories of the tongue for /ji/, /qi/, /xi/ from left to right (see online version for colours)



(a) (b) (c)

Note: arrows indicate the direction of movement; the scale of the horizontal and vertical axes is in centimeters (cm), representing the size dimensions of the tongue model.

Figure 10 (a)–(c) The midline movement trajectories of the tongue for /ga/,/ka/,/ha/ from left to right (see online version for colours)



(a) (b) (c)

Notes: Arrows indicate the direction of movement; the scale of the horizontal and vertical axes is in centimeters (cm), representing the size dimensions of the tongue model.

5 Conclusions

In the present study, a high-fidelity 3D tongue-model visualisation library was developed and a manual-matching–based 2D-to-3D mapping of ultrasound image sequences was implemented, yielding an average similarity of 91.7%. Such accuracy indicates that dynamic tongue movements during phoneme articulation can be faithfully captured, thus providing an intuitive and effective tool for language learning. Previous research Liu et al. (2013) has demonstrated that virtual-reality–based interactive pronunciation visualisation can partially enhance speech comprehension in deaf-mute children. However, existing 3D pronunciation-guidance modules are characterised by low model precision – a limitation addressed by the proposed approach.

Kinematic analyses of commonly confused phonemes were conducted, with focus on mid-posterior tongue configurations for Mandarin /ba/, /bi/ and /bu/. Distinctive features were identified, including maximal tongue elevation in /mi/ and posterior tongue height increase in /sa/. These findings can inform targeted guidance for individuals with hearing impairments and for non-native speakers of Mandarin, particularly in mastering articulatory gestures that are difficult to observe, such as plosives and fricatives.

The primary objective of this study is to develop a standardised reference model for pronunciation instruction. By utilising data from a standard speaker, the proposed model ensures normative quality while maintaining manageable data acquisition complexity, thus aligning with the practical requirements of educational applications. This standardised tongue model provides an intuitive articulatory reference for speech training among hearing-impaired individuals and second-language learners, offering several key benefits:

1 Compensating for the absence of auditory feedback by facilitating pronunciation learning through visual channels

2 Demonstrating tongue configurations for sounds that do not exist in the learner's native language

3 Supplying a quantitative benchmark of normal articulation for phonetic and speech pathology research.

Despite these advantages, the present work has several limitations.

1 The current speech corpus is confined to isolated monosyllables and does not incorporate disyllabic words, phrases, or continuous speech, where coarticulation effects play a crucial role.

2 Parameter optimisation for certain phonemes relies on manual adjustment, which limits automation and reproducibility.

3 While the standardised model serves effectively as a teaching reference, applications in personalised pronunciation correction require adaptation to inter-speaker anatomical variability.

4 Further validation through user studies involving hearing-impaired participants and Mandarin L2 learners is necessary to assess the model's effectiveness in improving pronunciation accuracy, interface usability, and long-term learning outcomes.

Future work will focus on expanding the speech corpus to include continuous speech, developing deep learning-based algorithms for automatic parameter optimisation, and exploring transfer learning strategies for personalised model adaptation. These efforts aim to further enhance both the practical applicability and scientific robustness of the proposed model.

Declarations

Conflicts of interest: all authors declare that they have no conflicts of interest.

References

Al Ani, S., Cleland, J. and Zoha, A. (2025) 'Deep learning in ultrasound tongue imaging: a systematic review toward automated detection of speech sound disorders', *Frontiers in Artificial Intelligence*, Vol. 8, p.1631134, DOI: 10.3389/frai.2025.1631134.

Badin, P., Elisei, F., Bailly, G. et al. (2008) 'An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data', *5th Conference on Articulated Motion and Deformable Objects* (*AMDO 2008*, LNCS 5098), pp.132–143.

Bernstein, L.E., Auer, E.T. and Takayanagi, S. (2004) 'Auditory speech detection in noise enhanced by lipreading', *Speech Commun*., Vol. 44, No. 1, pp.5–18, https://doi.org/10.1016/j.specom.2004.10.011.

Chen, R. (2012) 'Application of multimedia animation technology in teaching speech therapy for cleft palate', *Beijing Journal of Stomatology*, Vol. 20, No. 4, pp.231–232.

Dodd B. (1977) 'The role of vision in the perception of speech', *Perception*, Vol. 6, No. 1, pp.31–40, https://doi.org/10.1068/p060031.

Easterbrooks, S.R. and Baker, S. (2002) *Language Learning in Children who Are Deaf and Hard of Hearing: Multiple Pathways*, Allyn & Bacon, Boston.

Eiter, T. and Mannila, H. (1994) *Computing Discrete Fréchet Distance*, May.

Engwall, O. (2000) 'A 3D tongue model based on MRI data', *6th International Conference on Spoken Language Processing* (*ICSLP* 2000), ISCA, pp.901–904.

Engwall, O. (2002) 'Evaluation of a system for concatenative articulatory visual speech synthesis', *7th International Conference on Spoken Language Processing* (*ICSLP* 2002), ISCA, pp.665–668

Engwall, O. (2004) 'From real-time MRI to 3D tongue movements', *Interspeech 2004*, ISCA, pp.1109–1112.

Engwall, O. (2012) 'Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher', *Computer Assisted Language Learning*, Vol. 25, No. 1, pp.37–64.

Engwall, O. and Badin, P. (1999) 'Collecting and analysing two- and three-dimensional MRI data for Swedish', *Quarterly Progress and Status Report – Royal Institute of Technology*, Department of Speech, Music and Hearing, Vol. 3.

Fabre, D., Hueber, T. and Girin, L. (2017) 'Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract', *Speech Communication*, Vol. 93, pp.63–75, DOI: 10.1016/j.specom.2017.08.002.

Gibson, T. and Lee, S.A.S. (2021) 'Use of ultrasound visual feedback in speech intervention for children with cochlear implants', *Clinical Linguistics and Phonetics*, Vol. 35, No. 5, pp.438–457, DOI: 10.1080/02699206.2020.1792996.

Huang, D., Wu, X., Wei, J. et al. (2013) 'Visualization of mandarin articulation by using a physiological articulatory model', *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp.1–4.

Johnson, A. (2022) 'An integrated approach for teaching speech spectrogram analysis to engineering students', *The Journal of the Acoustical Society of America*, Vol. 152, No. 3, pp.1962–1969.

Liberman, A.M. and Mattingly, I.G. (1985) 'The motor theory of speech perception revised', *Cognition*, Vol. 21, No. 1, pp.1–36, https://doi.org/10.1016/0010-0277(85)90021-6.

Liu, X., Yan, N., Wang, L. et al. (2013) 'An interactive speech training system with virtual reality articulation for mandarin-speaking hearing impaired children', in *2013 IEEE International Conference on Information and Automation* (*ICIA*), August, pp.191–196, DOI: 10.1109/ICInfA.2013.6720294.

Macdonald, J. and McGurk, H. (1978) 'Visual influences on speech perception processes', *Percept Psychophysics*, Vol. 24, No. 3, pp.253–257, https://doi.org/10.3758/BF03206096.

Marinetti, C., Moore, P., Lucas, P. and Parkinson, B. (2011) 'Emotions in social interactions: unfolding emotional experience', in *Cognitive Technologies*, pp.31–46, https://doi.org/10.1007/978-3-642-15184-2_3.

Massaro, D. and Bosseler, A. (2002) 'Perceiving speech by ear and eye: multimodal integration by children with autism', *J. Dev. Learn. Disord.*, Vol. 7, pp.111–146.

Massaro, D.W. (1998) *Perceiving Talking Faces: From Speech Perception to A Behavioral Principle*, p.11, The MIT Press, Cambridge, MA, US, ISBN 978-0-262-13337-1.

Massaro, D.W. (2003) 'A computer-animated tutor for spoken and written language learning', Proceedings of the *5th International Conference On Multimodal Interfaces*, Association for Computing Machinery, New York, NY, USA, pp.172–175

Massaro, D.W. and Light, J. (2003) 'Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/', *8th European Conference on Speech Communication and Technology*, (*Eurospeech* 2003), ISCA, pp.2249–2252

Massaro, D.W. and Light, J. (2004) 'Using visible speech to train perception and production of speech for individuals with hearing loss', *Journal of Speech, Language, and Hearing Research*, Vol. 47, No. 2, pp.304–320

McGurk, H. and MacDonald, J. (1976) 'Hearing lips and seeing voices', *Nature*, Vol. 264, No. 5588, pp.746–748, https://doi.org/10.1038/264746a0.

Olson, D.J. (2022) 'Visual feedback and relative vowel duration in L2 pronunciation: the curious case of stressed and unstressed vowels', *Pronunciation in Second Language Learning and Teaching Proceedings*, Vol. 12, No. 1, https://doi.org/10.31274/psllt.13353.

Revina, E. (2019) 'The role of visualization in learning a foreign language', *Psychological and Pedagogical Sciences*, Vol. 16, pp.145–154, *Vestnik of Samara State Technical University*, https://doi.org/10.17673/vsgtu-pps.2019.2.11.

Serrurier, A. and Badin, P. (2008) 'A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data', *The Journal of the Acoustical Society of America*, Vol. 123, No. 4, pp.2335–2355.

Sumby, W.H. and Pollack, I. (1954) 'Visual contribution to speech intelligibility in noise', *J. Acoust. Soc. Am.*, Vol. 26, No. 2, pp.212–215, https://doi.org/10.1121/1.1907309.

Summerfield, Q. (1987) 'Some preliminaries to a comprehensive account of audio-visual speech perception', in: Dodd, B. and Campbell, R. (Eds.): *Hearing By Eye: the Psychology of Lip-Reading*, p.3–51, Lawrence Erlbaum Associates, Hillsdale, NJ.

Wang, L., Chen, H. and Ouyang, J. (2009) 'Evaluation of external and internal articulator dynamics for pronunciation learning: Interspeech 2009', *ISCA*, pp.2247–2250.

Wang, L., Chen, H., Li, S. et al. (2012) 'Phoneme-level articulatory animation in pronunciation training', *Speech Communication*, Vol. 54, No. 7, pp.845–856.

Ye, P. (2014) 'The application of accompanying animation videos in discourse teaching', *Foreign Language Teaching in Primary and Secondary Schools: Second Half of the Month*, Vol. 2014, No. 12, pp.17–20.

Yu, J. and Chen, C.W. (2017) 'From talking head to singing head: A significant enhancement for more natural human computer interaction', *2017 IEEE International Conference on Multimedia and Expo* (*ICME*), pp.511–516.

Yu, J. and Li, A. (2014) '3D visual pronunciation of Mandarine Chinese for language learning', *2014 IEEE International Conference on Image Processing* (*ICIP*), pp.2036–2040.

Yu, J. and Wang, Z.F. (2015) 'A video, text, and speech-driven realistic 3-D virtual head for human–machine interface', *IEEE Transactions on Cybernetics*, Vol. 45, No. 5, pp.991–1002.

Yu, J., Li, A., Hu, F. et al. (2013) 'Data-driven 3D visual pronunciation of Chinese IPA for language learning', *2013 International Conference Oriental COCOSDA* held jointly with *2013 Conference on Asian Spoken Language Research and Evaluation*, (O-COCOSDA/CASLRE), Gurgaon, India, IEEE, pp.1–6.

Zhang, J., Wei, J., Zhang, C. et al. (2014) 'Visualization of mandarin articulation driven by ultrasound data', *The 9th International Symposium on Chinese Spoken Language Processing*, IEEE, Singapore, pp.363–366.

Zhang, S.C., Liu, C., Li, F.J. et al. (2024) 'Tongue visualization model for mandarin pronunciation based on MRI', Wang, G., Yao, D. and Gu, Z. (Eds.): *12th Asian-Pacific Conference on Medical and Biological Engineering*, pp.363–370, Springer Nature Switzerland, Cham.