scientific data

Check for updates

OPEN

DATA DESCRIPTOR

An Audio-Ultrasound Synchronized Database of Tongue Movement for Mandarin speech

Yudong Yang[1,5], Rongfeng Su[1,5], Shaofeng Zhao[2], Jianguo Wei[3], Manwa Lawrence Ng[4], Nan Yan[1] ✉ & Lan Wang[1] ✉

Ultrasound imaging has been widely adopted in speech research to visualize dynamic tongue movements during speech production. These images are universally used as visual feedback in interventions for articulation disorders or visual cues in speech recognition. Nevertheless, the availability of high-quality audio-ultrasound datasets remains scarce. The present study, therefore, aims to construct a multimodal database designed for Mandarin speech. The dataset integrates synchronized ultrasound images of lingual movement, and the corresponding audio recordings and text annotations elicited from 43 healthy speakers and 11 patients with dysarthria through speech tasks (including vowels, monosyllables, and sentences), with a total duration of 22.31 hours. In addition, a customized helmet structure was employed to stabilize the ultrasound probe, precisely controlling for head movement and minimizing displacement interference, The proposed database carries apparent values in automatic speech recognition, silent interface development, and research in speech pathology and linguistics.

Background & Summary

Articulators are essential in speech production. According to the Source-Filter Theory of Speech Production[1–3], all speech sounds that we produce everyday are products of the source generated by periodic vibration of vocal folds through respiratory support and resonances of the vocal tract with configuration set for the specific speech sound to be produced. Speech articulators such as the tongue, lips, and the palate play a crucial role in configuring the specific vocal tract shape for the sound to be made, thus shaping these raw glottal (vocal fold) sounds into meaningful speech sounds. The complex coordination between the two systems (the source and the filter) has a great impact on the precision and accuracy of speech production. In particular, articulation, which determines the resonance characteristics of the vocal tract, gives rise to accurate production of different speech sounds, in a continuous manner. Therefore, being able to visualize articulation is important as it allows us to observe and understand the intricate dynamic articulatory movements.

Magnetic resonance imaging (MRI) has been used as a tool to visualize articulation, thanks to its ability to capture the entire vocal tract, providing a comprehensive view of articulatory actions[4–7]. However, its relatively low temporal resolution greatly limits its effectiveness in capturing rapid movements that is essential in understanding the intricate dynamics of speech production. Electromagnetic articulography (EMA) provides precise trajectories of specific articulators by attaching sensors to them[8,9], but it only provides discrete location data of selected points and it is fairly invasive and uncomfortable, and time consuming. On the other hand, ultrasound imaging (UTI) of the moving tongue stands out as a non-invasive, real-time solution that provides dynamic visualization of tongue movements without health risks compared to other methods[10–13].

As the above-mentioned technologies are not easily available, data sharing among professionals could greatly accelerate research and advancement in the field. Currently, multiple ultrasound databases exist for English speakers, including the TAL corpus which consists of 82 English participants and provides ultrasound

[1]Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. [2]Department of Rehabilitation Medicine, The Eighth Affiliated Hospital of Sun Yat-sen University, Shenzhen, China. [3]College of Intelligence and Computing, Tianjin University, Tianjin, China. [4]Speech Science Laboratory, Faculty of Education, University of Hong Kong, Hong Kong SAR, China. [5]These authors contributed equally: Yudong Yang, Rongfeng Su. ✉e-mail: nan.yan@siat.ac.cn; lan.wang@siat.ac.cn

Datasets	No. of speakers	Type	Resolution	Modalities	Language	Duration	Task
TAL	82	Normal	64 × 842	Ultrasound, Lip Videos, Audio	English	13.5 hours(audio)	Sentence, Non-words
UltraSuite	86	Normal/Dysarthria (Children)	63 × 412	Ultrasound, Audio, Text	English	18.67 hours(audio)	Words, Sentence, Non-words
SSR7000	1	Normal	640 × 445	Ultrasound, Lip Videos	English	7484 samples	Sentence
AUspeech	54	Normal/Dysarthria (Adult)	920 × 700	Ultrasound, Audio, Text	Mandarin	22.31 hours	Vowels, Monosyllable, Sentence, Non-words

Table 1. Comparison of AUSpeech with other different databases in terms of resolution, duration, tasks, etc.

Category	Types	Male	Female	Total
Participants	Normal (Average Age: 24.2)	21 (Average Age: 23.6)	22 (Average Age:24.8)	43
	Dysarthria (Average Age: 60.0)	9 (Average Age: 59.5)	2 (Average Age: 62.0)	11
Total Participants	\	29	25	54

Table 2. Demographic information results of the main AUSpeech database. Number (N), sex.

images of tongue, lip videos, and 13.5 hours of audio data[10]. Similarly, the UltraSuite dataset includes 86 speakers of Scottish-accented English, providing ultrasound data and 18.67 hours of audio recordings. However, as shown in Table 1, the ultrasound images contained within these databases are consistently with low resolution, making it difficult to identify fine and detailed articulatory information[11]. Although the SSR7000 dataset has high-resolution ultrasound of the tongue, it is limited to only one English speaker. Mandarin is a tone language, and its tonal changes require rapid and precise oral and lingual movements during articulation[12]. It follows that constructing an ultrasound dataset of Mandarin is of great research value. In addition, such database can have obvious clinical and practical applications, including training systems for automatic speech recognition (ASR), early and accurate identification of types and severity of dysarthria (articulation disorders), teaching to speak Mandarin Chinese as a second language, childhood phonological disorders, etc[14–19].

In the present work, a multimodal Mandarin ultrasound dataset containing parallel UTI, text and speech data was established. The dataset consists of 43 healthy speakers and 11 patients with dysarthria, and all participants are native speakers of Mandarin Chinese. The UTI data were with a resolution of 920 × 700 pixels at 60 frames per second, and the total recording time was about 22.31 hours, which provides a comprehensive platform for investigating the dynamic articulatory mechanisms of Mandarin speech production. As a language with a rich tonal system, Mandarin has complex phonetic and articulatory features. In order to explore these features in depth, the dataset is designed with three types of tasks: vowel, monosyllable and sentence productions, which cover almost all common Chinese pronunciation patterns, with particular attention paid to the articulation patterns of key phonological phenomena, such as back consonants (e.g., [tʂ], [tʂ^{h}], [ʂ]), high front vowel (e.g., [i]) and rounded back vowel (e.g., [u], [y]). The dataset not only provides important and basic data for the study of Mandarin phonology, but also provides strong support for applications in cross-linguistic research, speech recognition, speech synthesis and clinical speech therapy[20–26].

Methods

Participants. As shown in Table 2, the AUSpeech dataset[27] contains two groups of participants: 43 healthy subjects and 11 individuals with dysarthria. The healthy participants consisted of 21 males and 22 females, with an average age of 24.2 years. The total duration of recordings of the healthy control group was 22.313 hours. All healthy participants had no reported history of speech, hearing, or neurological disorders. Inclusion criteria required participants to be native adult speakers of Mandarin Chinese who were between 20 and 30 years old and had normal vision and hearing. Exclusion criteria included any history of psychiatric or cognitive disorders, neurological conditions, or speech-related impairments. For the dysarthric patients, a dataset of 0.74 hours was included. The inclusion criteria were: (1) ages ranged from 45 to 70 years old, reflecting the typical age distribution of post-stroke dysarthria, as stroke-related speech disorders predominantly affect middle-aged and older adults; (2) native Mandarin speakers; (3) diagnosed with speech articulation disorders; 4) normal vision and hearing. Exclusion criteria of the patients with dysarthria included: (1) any history of psychiatric conditions; (2) other disorders that could affect speech production. The ethics of the study was approved by the Institutional Research Ethics Committee of Shenzhen Institute of Advanced Technology and the Eighth Affiliated Hospital of Sun Yat-Sen University. Written informed consent was obtained from all participants or their relatives.

Speech materials. To ensure the comprehensiveness of data, the speech materials were designed to include vowel, monosyllable, and sentence productions, capturing various aspects of Mandarin phonetics. The dataset is systematically constructed from two primary linguistic resources: a curated collection of 405 high-frequency monosyllabic lexical items representing fundamental phonological units and six primary simple finals in Mandarin Chinese, and 17,500 unique sentence-level samples extracted from the Chinese Linguistic Data Consortium (CLDC) corpus. This comprehensive compilation encompasses the full spectrum of Mandarin phonological structures, including complete coverage of permissible syllable onsets.

Vowel sustention task. Participants were asked to produce six primary simple finals (/a/, /o/, /e/, /i/, /u/, and /ü/) continuously for approximately two seconds each. This continuous pronunciation facilitated the acquisition of stable imaging of tongue movements. These vowels were selected as they represented a collection of speech sounds requiring a wide range of tongue positions, and yielded data for studying both anterior and posterior tongue positions in Mandarin. The UTI data captured dynamic articulatory configurations while complemented by acoustic data.

Monosyllable production task. The monosyllable task was designed to encompass a diverse array of mandarin phoneme combinations, providing insight into the articulation of both consonants and vowels. Each participant produced 15 distinct monosyllables selected from a set of 405 unique syllables, with each syllable repeated three times (totaling 45 recordings per participant), which included syllables that are frequently used in real life. The selected syllables represented a balanced distribution of consonants and vocalic sounds, capturing essential variations for phonetic analysis. For example, the syllable "当" highlights tongue articulation, with the plosive [t] produced by pressing the tongue tip against the alveolar ridge to block airflow before abruptly releasing it, smoothly transitioning into the vowel [a]. These syllables exemplify the intricate coordination between consonants and vowels in Mandarin, emphasizing the crucial role of tongue position in phoneme production. Including such speech stimuli ensures that the monosyllable set captures essential articulatory variations, providing a rich foundation for phonetic analysis.

Sentence-reading tasks. Participants were presented with 375 unique sentences selected from a larger database of about 17,500 Mandarin sentences. The sentence prompts were sourced from the CLDC corpus, which was designed to reflect everyday Mandarin speech. Each set of 375 sentences was tailored to the participant, ensuring no repetition across the dataset and maintaining a diverse sampling of Mandarin phonetic patterns. The variety of sentence structures and syllable frequency provided rich data for analyzing natural speech production across different contexts. The sentences encompassed various phonetic and syntactic complexities, facilitating the exploration of articulatory behavior in continuous speech.

Data acquisition device. The AUSpeech dataset contains synchronized audio and UTI data. The audio recordings were obtained by using a BOYA BY-WM4 PRO wireless lapel microphone, using a sampling rate of 16 kHz, 16-bit encoding, and single-channel audio. The UTI data was captured by using the Focus & Fusion Finus 55 ultrasound device, equipped with a phased array probe (P5-2). The probe was placed under the chin of the participant to capture images of the tongue in the sagittal plane. Key ultrasound parameters were optimized for articulatory analysis: (1) Sampling rate: 60 frames per second (fps). (2) Spatial resolution: 920 × 700 pixels, providing high-definition visualization of tongue contours. (3) Dynamic range (DR): 114 dB, enhancing contrast between soft tissue interfaces (e.g., tongue surface vs. oral cavity). (4) Line density: 1.8–4.6 MHz, adjusting transmit frequency to optimize penetration depth and image clarity. (5) Soft tissue thermal index (TIS): 0.91, maintaining safety standards for prolonged exposure. (6) Maximum depth: 11 cm, ensuring full coverage of the adult tongue and adjacent structures. (7) Focal point: 5.8 cm, aligned with the mid-tongue position in adults to maximize focal zone precision. Furthermore, ultrasound images were streamed via HDMI in full-screen mode and vertically flipped to match anatomical orientation, ensuring accurate spatial representation.

To synchronize the audio and ultrasound signals, AVerMediaGC553 4 K data acquisition card was used. This card can capture multiple data streams concurrently. The HDMI input from the ultrasound device was connected to the acquisition card, which interfaced with the computer via a USB 3.1 connection, ensuring high-speed data transmission and stable recording. In addition, a customized support system was developed to stabilize the ultrasound probe. This system utilized two modified mechanical mounts: one was integrated into the helmet to fit the skull structure, minimizing unnecessary head movements that could affect the stability of the imaging plane, while the other was attached to the ultrasound probe. The probe mount featured an adjustable position and angle, allowing precise alignment with the tongue's mid-sagittal plane, and was secured using hot-melt adhesive and locking mechanism such as screws to ensure stability throughout the recording process. During the experiment, participants were facing a computer screen with their chins extended slightly forward to enable clear imaging of tongue movement. The schematic diagram is shown in Fig. 1(b).

Experimental paradigm. This section provides a systematic characterization of the AUSpeech dataset's collection process. The AUSpeech was collected in a controlled acoustic environment, and the acquisition of speech data was divided into different subsets (named "Normal" and "Patient"). Figure 2 schematically illustrates (1) the integrated acquisition framework comprising multi-channel recording apparatus, and (2) the operational workflow governing multi-modal data capture.

Participants were seated facing a computer screen displaying speech prompts and instructed to read each item aloud sequentially. The ultrasound probe position was fixed and adjusted to match anatomical landmarks. During the experiment, ultrasound probe stability and image quality were rigorously monitored. The session was paused if significant displacement of tongue imaging occurred (e.g., due to head movement). Both the affected item and the preceding one were re-recorded to maintain reliability between audio, ultrasound, and text modalities throughout the dataset.

Throughout the procedure, participants were instructed to maintain natural speech patterns with prohibited vocal modulation or exaggerated articulation. As shown in Fig. 2, In each articulation trial, participants were asked to perform three standardized swallowing maneuvers to serve as temporal markers delineating session initiation and termination phases. Each articulation trial comprised three chronologically defined stages: (1) Preparation phase (1,500 ms): Participants read on-screen instructions while baseline tongue positioning was recorded via ultrasound imaging. This phase established initial articulatory postures for subsequent analysis. (2) Production phase:

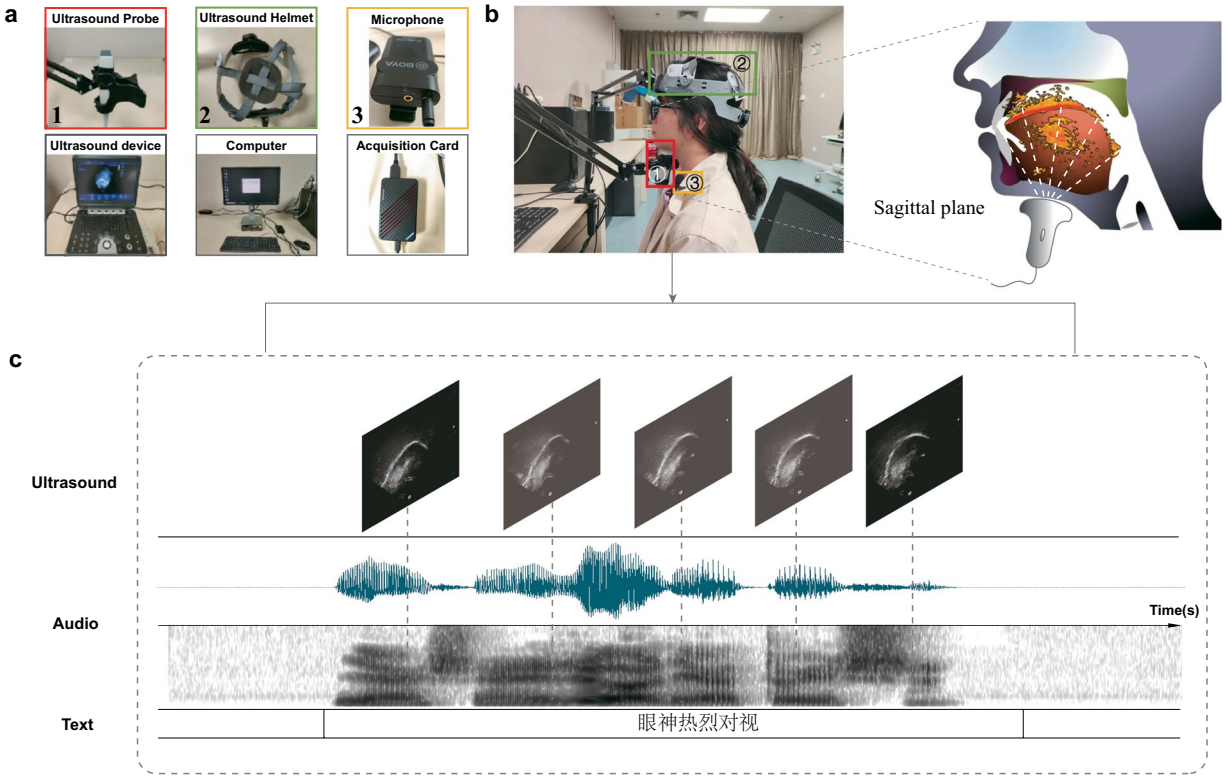

Fig. 1 System overview and data collection process. (**a**) Equipment used for data acquisition: ① Ultrasound probe: captures real-time tongue movement during speech production. ② Ultrasound helmet: stabilizes the probe to ensure consistent imaging. ③ Microphone: records synchronized speech audio. Additional components include the ultrasound imaging device, a computer for stimulus display, and an acquisition card for synchronizing data streams. (**b**) Experimental setup and sagittal view of tongue imaging. The participant wears the ultrasound helmet to keep the probe stable while reading aloud from a computer screen and the individual has provided consent for their image to be shown in the paper. (**c**) Speech tasks and multimodal data synchronization. Participants performed speech tasks: vowels (Task A), monosyllables (Task B), and sentences (Task C). Each task's output includes synchronized ultrasound frames (top row), corresponding speech waveforms (middle row), and text annotations (bottom row).

Upon text presentation, participants performed verbally cued tasks using natural prosody while maintaining head stabilization for optimal ultrasound tongue imaging (UTI) acquisition. The hierarchical speech protocol included: Vowel sustention, Monosyllable production, and Sentence-reading tasks. (3) Inter-trial interval (1,500 ms): Participants maintained a neutral oral posture (closed mouth position) during blank screen displays, preparing for subsequent trials. The experimental protocol required healthy participants to perform three swallowing maneuvers as temporal markers at both the initiation and termination phases of the session. This approach enabled a detailed comparison of articulatory movements between patients and healthy individuals, providing valuable insights into differences in speech production patterns, particularly among those diagnosed with articulation disorders.

Data annotation. On the basis of ensuring the quality of UTI data, the Montreal Forced Aligner (MFA) and Voice Activity Detection (VAD) tool was used to force align and automatically label the speech and text data and generate the related TextGrid annotation files[28]. In addition, manual inspection annotation was performed to ensure the reliability of the alignment. This allowed accurate calibration of the correspondence between speech and tongue movement in the temporal dimension. The MFA and VAD can be used to accurately align each pronunciation in the audio signal to the time point of the UTI to achieve the correlation. Furthermore, for patient data, all annotations were performed manually with detailed precision to ensure accuracy.

Data Records

Database description. The AUSpeech dataset is available at https://cstr.cn/31253.11.sciencedb.18722[27] with a total size of approximately 676.16 GB. As shown in Table 3, the dataset consists of 22.31 hours of synchronized audio and ultrasound data collected. Data obtained from the Normal sessions, totaling 21.57 hours (10.39 hours from males and 11.18 hours from females), contained swallowing, vowel vowel, monosyllable, and sentence productions. Swallowing tasks comprised 0.75 hours, with 0.35 hours from males and 0.40 hours from females. Recordings of vowels accounted for 0.609 hours, almost equally contributed by males (0.30 hours) and females (0.30 hours). Monosyllables made up 1.22 hours, with 0.58 hours from males and 0.64 hours from females. Sentences represented the largest portion of the dataset, totaling 19.00 hours, contributed by 9.16 hours from

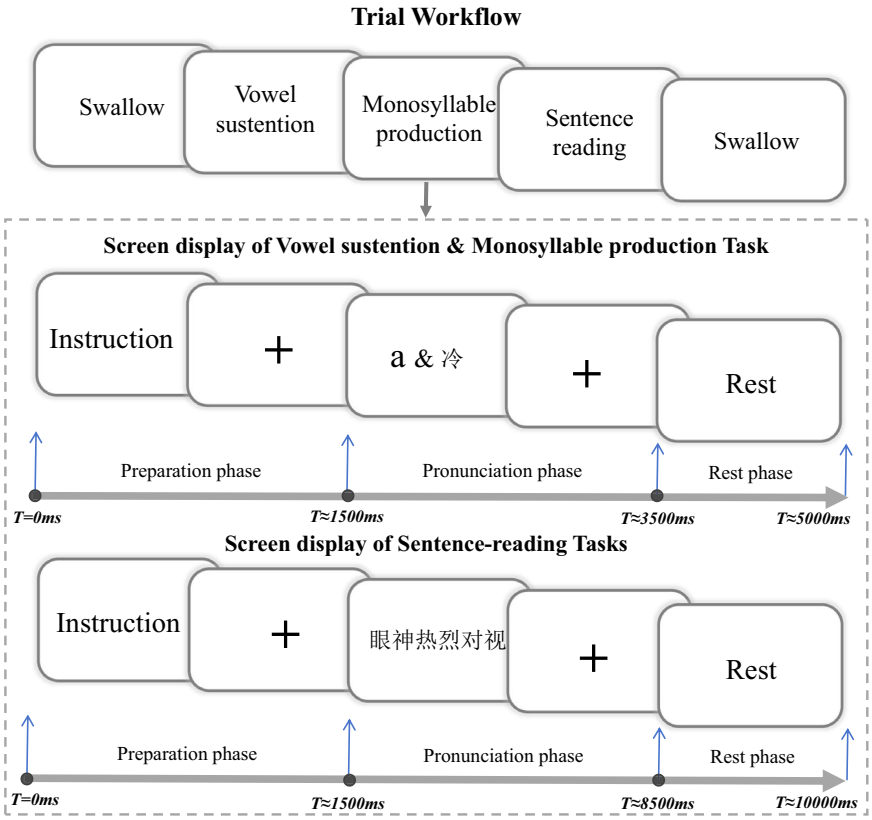

Fig. 2 Schematic diagram of the recording procedure.

Session	Category	Man (h = hour)	Female (h = hour)	Total (h = hour)
Normal	Swallowing	0.35 h	0.40 h	0.75 h
	vowel	0.30 h	0.30 h	0.60 h
	Monosyllables	0.58 h	0.64 h	1.22 h
	Sentences	9.16 h	9.84 h	19.00 h
	Total duration	10.39 h	11.18 h	21.57 h
Patient	vowel	0.04 h	0.01 h	0.05 h
	Monosyllables	0.34 h	0.04 h	0.38 h
	Words	0.16 h	0.15 h	0.31 h
	Total duration	0.54 h	0.20 h	0.74 h
All	Total duration	10.93 h	11.38 h	22.31 h

Table 3. Detailed duration results of the main AUSpeech database. Number(N), sex.

males and 9.84 hours from females. The Patient session added 0.74 hours to the dataset (0.54 hours from males and 0.20 hours from females) and includes vowels (0.05 hours), Monosyllables (0.38 hours), and Word tasks (0.31 hours). This comprehensive dataset offers a balanced representation across genders and a variety of speech tasks, making it a valuable resource for research in speech dynamics, phonetics, and clinical speech studies.

Data organization and storage. The AUSpeech dataset is systematically organized into a hierarchical directory structure to ensure accessibility and efficient data retrieval. As illustrated in Fig. 3, the dataset is divided into two primary session-level folders: Normal/, which contains data from 43 healthy participants, and Patient/, which contains data from 11 dysarthric participants. Within these sessions, the data is further subdivided based on participant-specific tasks and modalities, with individual folders for each speaker that facilitate easy identification of the speaker, session, and corresponding speech tasks.

Each participant folder (e.g., Speaker0001_M/) is organized into three subfolders: (1) Audio/: Speech recordings in .wav format, (2) Ultrasound/: Sagittal-plane tongue motion images in .dcm format and (3) Text/: Transcripts files in .lab and .TextGrid formats. A strict naming convention is employed for clarity and consistency. for example, an audio file is named according to the format speaker[ID]_[Gender]_[Session]_[stn][ID].

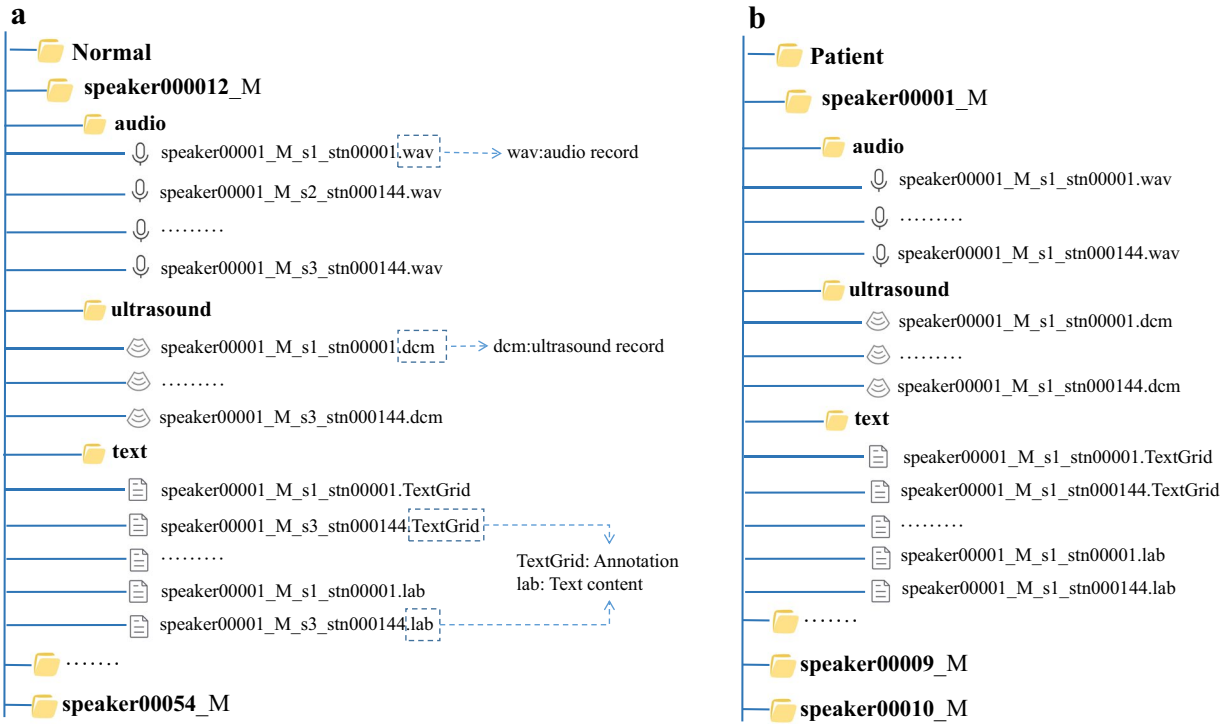

Fig. 3 Organization structure of the AUSpeech database. (**a**) General overview of the normal dataset directory structure. (**b**) Content of the patient dataset participant directories.

wav (e.g., speaker00012_M_s1_stn00001.wav), which clearly indicates the speaker's unique identifier, gender, session (s1 for Normal or s2 for Patient), and speech item. This structured approach, along with clearly defined metadata elements such as [ID], [Gender], [Session], [Task], and [ItemID], ensures that all aspects of the dataset are well-organized and readily available for further analysis.

Technical Validation

The quality of UTI data plays a crucial role in the dynamic analysis of tongue movement[29–31]. To ensure the temporal consistency and signal integrity of the data, abnormal UTI data that might result from errors during the acquisition process were strictly screened out. Two types of abnormalities were identified: frames with no signal and frames with articulatory movement that was significantly stationary. First, a similarity-checking script (examples can be found on the datasets website) was used to detect and mark frames with no signal in the UTI-acquired results. This process involved matching frames against a template of signal-free examples. Subsequently, the marked frames were reviewed manually by professionals to ensure that these anomalies did not affect the consistency of tongue motion trajectories or the alignment between audio and ultrasound data.

Additionally, small probe displacements or missing device signals during acquisition could result in static frames, disrupting the dynamic consistency of the data. To address this, a static frame detection algorithm was employed to automatically identify still frames by analyzing the similarity between consecutive frames and marking those with significant motion stagnation as outliers. These flagged frames were manually verified and discarded to ensure the stability and reliability of the data for analysis. Through these rigorous data screening and quality control procedures including systematic rejection of abnormal data and consistency checks, robust support for the study of tongue motion and speech correlation were in place. A sample image is shown in Fig. 4.

Furthermore, in our previous work we used AUSpeech normal subset to perform an acoustic-to-articulatory inversion generation task[13]. The model was designed to generate ultrasound tongue imaging data solely from the audio signal, and the generated tongue movement patterns and images were remarkably similar to the original ultrasound images in both spatial detail and temporal dynamics. This close resemblance to the original data not only demonstrates the effectiveness of our inversion methods but also serves as compelling evidence for the reliability of our parallel speech and ultrasound recordings. This technical validation is crucial for downstream applications, including silent speech interfaces, articulatory synthesis, automatic speech recognition, and clinical speech pathology studies.

Usage Notes

The AUSpeech dataset, along with the provided code, serves as resource for researchers interested in a range of speech-related tasks. It can be utilized for applications such as ultrasound generation, automatic speech recognition, and pathological research. The accompanying Python scripts demonstrate key operations for processing the dataset.

The code samples make use of several core libraries, including librosa for audio analysis, matplotlib for data visualization, numpy for numerical operations, textgrid for handling transcription files, and pydicom for

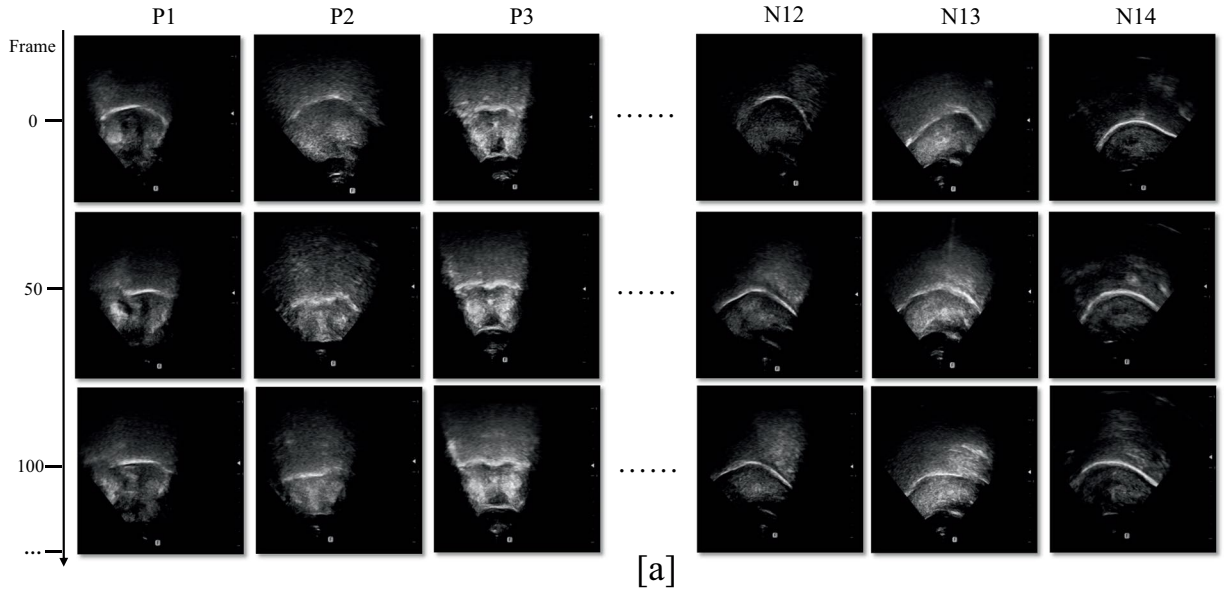

Fig. 4 Example ultrasound tongue images (sagittal-plane) randomly selected from normal (N12–N14) and patient (P1–P3) speakers producing the vowel [a], to visualize tongue motion trajectories, one image was showed every 50 frames.

processing ultrasound images. For instance, the data_processing.ipynb script illustrates how to load an audio file and parse transcription files, displaying their contents for easy verification. Additionally, scripts enable the display of sample frames from ultrasound images, facilitating detailed analysis of tongue motion.

Code availability

The scripts for preprocessing the audio and video data as well as those used in the validation section are available at https://github.com/huanraozhineng1/AUSpeech_code.

Received: 2 January 2025; Accepted: 26 March 2025;
Published online: 11 April 2025

References

1. Raphael, L. J., Borden, G. J. & Harris, K. S. Speech Science Primer – Physiology, Acoustics, and Perception of Speech (6th edition). Glendale, CA: Lippincott Williams & Wilkins (2011).
2. Honda, K. Physiological processes of speech production[J]. Springer handbook of speech processing, 7–26 (2008).
3. Lacroix, A. Speech Production—Acoustics, Models, and Applications[M]//Communication Acoustics. Berlin, Heidelberg: Springer Berlin, Heidelberg, 321–337 (2004).
4. Lim, Y. *et al.* A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images[J]. *Scientific data* **8**(1), 187 (2021).
5. Narayanan, S. *et al.* Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *J. Acoust. Soc. Am.* **136**, 1307–1311 (2014).
6. Isaieva, K. *et al.* Multimodal dataset of real-time 2D and static 3D MRI of healthy French speakers[J]. *Scientific Data* **8**(1), 258 (2021).
7. Kim, J. *et al.* USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging. In Proc. the 10th Int. Semin. Speech Prod. 226–229 (2014).
8. Ji, A., Berry, J. J. & Johnson, M. T. The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 7719–7723 (2014).
9. Richmond, K., Hoole, P. & King, S. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus[C]//Twelfth Annual Conference of the International Speech Communication Association. (2011).
10. Ribeiro, M. S. *et al.* TaL: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos[C]//2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 1109–1116 (2021).
11. Eshky, A., *et al.* UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions. Proc. Interspeech 2018, 1888–1892, https://doi.org/10.21437/Interspeech.2018-1736 (2018).
12. Kimura, N. *et al.* SSR7000: A Synchronized Corpus of Ultrasound Tongue Imaging for End-to-End Silent Speech Recognition[C]// Proceedings of the Thirteenth Language Resources and Evaluation Conference. 6866–6873 (2022).
13. Yang, Y. *et al.* An Audio-Textual Diffusion Model for Converting Speech Signals into Ultrasound Tongue Imaging Data[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2170–2174 (2024).
14. Preston, J. L. *et al.* Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors[J]. Journal of visualized experiments: JoVE, (119) (2017).
15. Davidson, L. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance[J]. *The Journal of the Acoustical Society of America* **120**(1), 407–415 (2006).
16. Ménard, L. *et al.* Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model. *JJ. Folia Phoniatrica et Logopaedica* **64**(2), 64–72 (2012).
17. Hueber, T. *et al.* Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips[J]. *Speech Communication* **52**(4), 288–300 (2010).

18. Gosztolya, G. *et al*. Applying dnn adaptation to reduce the session dependency of ultrasound tongue imaging-based silent speech interfaces[J]. *Acta Polytechnica Hungarica* **17**(7), 109–124 (2020).
19. Ding, H. *et al*. Ultraspeech: Speech enhancement by interaction between ultrasound and speech[J]. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **6**(3), 1–25 (2022).
20. Kabakoff, H. *et al*. Extending ultrasound tongue shape complexity measures to speech development and disorders[J]. *Journal of Speech, Language, and Hearing Research* **64**(7), 2557–2574 (2021).
21. Allen, J. E., Cleland, J. & Smith, M. An initial framework for use of ultrasound by speech and language therapists in the UK: Scope of practice, education and governance[J]. *Ultrasound* **31**(2), 92–103 (2023).
22. Zheng, R. C., Ai, Y. & Ling, Z. H. Incorporating Ultrasound Tongue Images for Audio-Visual Speech Enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2024).
23. Chiu, C. *et al*. Sibilant fricative merging in Taiwan Mandarin: An investigation of tongue postures using ultrasound imaging[J]. *Language and speech* **63**(4), 877–897 (2020).
24. Chen, S. Whalen, D. H. & Mok P. P. K. Production of the English/ɹ/by Mandarin–English Bilingual Speakers[J]. Language and Speech, 00238309241230895 (2024).
25. Lee, J., Littlejohn, M. A. & Simmons, Z. Acoustic and tongue kinematic vowel space in speakers with and without dysarthria[J]. *International Journal of Speech-Language Pathology* **19**(2), 195–204 (2017).
26. Sugden, E. & Cleland, J. Using ultrasound tongue imaging to support the phonetic transcription of childhood speech sound disorders[J]. *Clinical linguistics & phonetics* **36**(12), 1047–1066 (2022).
27. AUSpeech: An Audio-Ultrasound Synchronized Database of Tongue Movement for Mandarin speech. Science Data Bank, https://cstr.cn/31253.11.sciencedb.18722 (2025).
28. McAuliffe, M. *et al*. Montreal forced aligner: Trainable text-speech alignment using kaldi[C]//Interspeech. **2017**: 498–502 (2017).
29. Ohkubo, M. & Scobbie, J. M. Tongue shape dynamics in swallowing using sagittal ultrasound[J]. *Dysphagia* **34**(1), 112–118 (2019).
30. Dugan, S. *et al*. Tongue part movement trajectories for/r/using ultrasound[J]. *Perspectives of the ASHA special interest groups* **4**(6), 1644–1652 (2019).
31. Xu, K. *et al*. Robust contour tracking in ultrasound tongue image sequences[J]. *Clinical linguistics & phonetics* **30**(3-5), 313–327 (2016).

Acknowledgements

This work is supported by National Natural Science Foundation of China (U23B2018, NSFC 62271477), Shenzhen Science and Technology Program (JCYJ20220818101411025, JCYJ20220818102800001, JCYJ20220818101217037) and Shenzhen Peacock Team Project (KQTD20200820113106007).

Author contributions

All authors contributed extensively to the work presented in this paper. Y.Y. was involved in data collection and writing. S.R. was responsible for data preprocessing and organization. Z.S. carried out subject recruitment and assessment. W.J. engaged in equipment research and development and experimental design. M.N.L. participated in discussion and writing. W.L. and Y.N. was involved in experimental protocol design, planning, and writing. We and our co-authors followed standard research ethical procedures in the conduct of the study and do not have any interests that might be interpreted as influencing the research. All authors have read the manuscript, and the paper has not been published and is not under simultaneous consideration by another journal. There has been no ghost writing by anyone not named on the authors list.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Correspondence and requests for materials should be addressed to N.Y. or L.W.