Real-Time MRI Synthesis from Text Using a Two-Stage Pipeline for Pronunciation Training

***Abstract*—Providing visual or animated information on how to position and move the tongue and other speech articulators is beneficial for children with hearing impairments to improve their pronunciation. However, since these articulatory movements occur inside the mouth, they are not visible. Real-time MRI of the midsagittal plane of the mouth can be used to extract this information. However, acquiring MRI data for every word is both time consuming and impractical. In this work, we propose a method to synthesize MRI sequences for a given text. Our approach involves a two-stage pipeline. In the first stage, we perform text-to-speech synthesis that replicates the speaker's voice characteristics. In the second stage, the synthesized speech is mapped to an MRI sequence using a speech-MRI mapping model trained on the speaker recordings. The synthesized speech and corresponding rtMRI sequences are then integrated to generate a synchronized rtMRI video, providing a simultaneous representation of both articulatory movements and acoustic output. A key advantage of the proposed method is that it eliminates the requirement for phoneme-to-MRI frame alignment during the training of phoneme-to-MRI mapping models, a process that is frequently error-prone when relying on forced-alignment techniques. We employed the Tortoise-TTS system for speech synthesis from the text, ensuring that the generated speech matches the target speaker's voice characteristics. A bidirectional Long Short-Term Memory network was trained to map speech to the corresponding MRI image sequence. To evaluate the synthesized MRI sequences, we used both objective measures, such as Frechet Inception Distance, and subjective evaluations. The results demonstrate that the generated MRI sequences are realistic and closely resemble natural recordings.**

***Index Terms*—MRI, Text-to-Speech synthesis, articulation, acoustic-to-articulatory mapping.**

I. Introduction

The ability to pronounce words clearly is a significant challenge for children with hearing impairments, as they often lack auditory feedback, which is crucial for learning correct articulation. Visual or animated feedback on the positioning and movement of speech articulators, such as the tongue, lips, and jaw, can provide an effective means of improving their pronunciation skills [1]–[3]. Such feedback helps these children understand how to physically shape their vocal tract to produce the correct sounds, compensating for the auditory cues they are unable to perceive. However, a major limitation is that these articulatory movements occur inside the mouth, making them difficult to observe directly.

Various imaging techniques, such as X-ray, Electromagnetic Articulography, ultrasound, and Magnetic Resonance Imaging (MRI), can be employed to capture the articulatory movements that occur inside the mouth [4]–[7]. Real-time Magnetic Resonance Imaging (rtMRI) of the midsagittal plane of the mouth provides the complete visualization of internal articulatory movements during speech. rtMRI offers detailed insights into the dynamic movements of the tongue and other speech articulators, providing a clear visual representation of how speech sounds are formed [8], [9]. Despite its potential, acquiring MRI data for every word in a vocabulary set is highly resource-intensive. The process of capturing real-time MRI data requires specialized equipment, expert supervision, and significant time, making it impractical for large-scale use in speech therapy applications. Therefore, there is a need for methods that can generate such visualizations without the constraints of continuous MRI acquisition for every individual utterance. Additionally, estimating rtMRI data from text has promising applications in speech synthesis [10] for generating natural-sounding speech, as well as in advancing the understanding of articulation processes [11].

Previous research has made several attempts to estimate real-time MRI (rtMRI) sequences from phoneme sequences [12]–[15]. These methods either use cocantenative synthesis or deep learning techniques to obtain the rtMRI sequences from phonemes. These approaches typically rely on forced alignment methods to link specific phoneme sequences to corresponding MRI frames. Forced alignment is a technique used in automatic speech recognition (ASR) to align speech audio with its transcription by matching audio segments to phonemes.

However, forced alignment methods are primarily designed for transcribing speech, not for precisely capturing the exact timing of each phoneme. Since ASR systems do not require highly accurate phoneme-level alignments to produce correct transcriptions, the forced alignments obtained for phoneme-to-MRI frame mapping often contain errors. Additionally, the dynamic nature of speech makes it difficult to map a continuous sequence of MRI frames to a single phoneme. This mismatch introduces further challenges in generating accurate rtMRI sequences using these methods. Therefore, this limitation need to be addressed to improve the accuracy of rtMRI synthesis from phoneme sequences.

To address the challenges in generating rtMRI sequences from text, we propose a two-stage pipeline. This approach aims to overcome the limitations of previous methods that relied on forced alignments. In the first stage, speech is synthesized using the Tortoise-TTS system, ensuring that the generated speech closely matches the target speaker's characteristics. In the second stage, this synthesized speech is mapped to rtMRI image sequences through a trained speech-to-MRI model. The synthesized speech and corresponding

rtMRI sequences are then combined to produce a synchronized rtMRI video. A key advantage of our method is that it bypasses the need for explicit alignments between text and MRI frames required for training the phoneme-to-MRI mapping, which have been a source of errors in earlier approaches. Since there is no direct mapping from text to MRI frames, forced phoneme alignments are not necessary. Instead, we leverage the capabilities of Text-to-Speech systems, which can generate high-quality speech from text without requiring precise alignment at the phoneme level. This significantly simplifies the process of generating MRI sequences from text, while maintaining the accuracy of the visualized articulatory movements. For our experiments, we utilize the rtMRI dataset [8], which serves as the foundation for training and testing the proposed pipeline.

II. Related works

Previous research has explored the use of concatenative approaches to generate word-level rtMRI frames, as well as deep learning-based methods to map phoneme sequences to corresponding rtMRI frame sequences. Desai et al. [12] used a concatenative synthesis approach to create articulatory videos from audio without the need for simultaneous articulatory recordings. This method builds a repository of phoneme-specific articulatory images from existing rtMRI videos, then selects and combines the appropriate image sequences to generate smooth, high-quality videos for new audio inputs. In tests using unseen words from the MRI-TIMIT database, the synthesized videos received a subjective quality rating of 3.78 out of 5. Chandana et al. [13] improved upon this work by using an augmentation method based on pixel intensities that reflect muscle movements, highlighting the regions enclosed by air-tissue boundaries. The method involved synthesizing air-tissue boundaries from a few selected frames containing articulatory movements. Subjective evaluations show that the augmented videos for 50 words are more visually appealing, with a rating of 3.75 out of 5. Although these concatenative approaches produced high-quality rtMRI videos, the quality diminished as the number of phonemes in a word increased, making them unsuitable for generating rtMRI videos for full utterances. [14]

Udupa and Ghosh [14] proposed a sequence-to-sequence model that generates MRI video frames based on phonemes derived from text input. Their approach maps phoneme sequences to articulatory movements, producing video frames that correspond to predefined text. They used transformer-based phoneme encoder and convolutional frame decoder to generate the rtMRI frames from phoneme sequence. This method used the forced alignment to obtain the phoneme and rtMRI frame alignments. They showed that this deep learning model was able to generate realistic rtMRI video for unseen utterances. Ribeiro et al. [15] proposed an encoder-decoder network designed to map phoneme sequences to the contours of articulators in rtMRI images. The researchers utilized forced alignments to establish the correspondence between phonemes and rtMRI frame sequences. They extracted the contours of the articulators from the rtMRI images, which were then used to train the mapping model. The results demonstrated that the model effectively generated high-quality vocal tract shapes, exhibiting strong correlations between the predicted and target articulatory variables.

The generation of rtMRI sequences often depends on forced alignment methods to associate specific phoneme sequences with corresponding MRI frames. However, these forced alignments can introduce errors, as they are derived from automatic speech recognition systems that do not prioritize precise alignments. In contrast, accurate time alignments are essential for phoneme-to-rtMRI mapping. To address these challenges, we propose a two-stage pipeline strategy that incorporates Text-to-Speech and speech-to-MRI models, eliminating the need for forced alignments. The following sections present the dataset, proposed methodology, experimental setup, results, and conclusions.

III. Dataset

In this work, we utilize the real-time MRI data from the USC-TIMIT Speech Production Database [8]. This dataset provides synchronized speech and rtMRI image sequences for ten American English speakers, comprising five male and five female participants. Each speaker articulates 460 sentences, selected from the MOCHA-TIMIT corpus, capturing both speech audio and corresponding articulatory movements. For our study, we focused on the data from four speakers (two male speakers: M2, M3, and two female speakers: F1, and F2). These speakers were selected based on the quality of their recordings and the synchronization between their speech and rtMRI image sequences, which was essential for the accuracy of our analysis and the effectiveness of our proposed method.

IV. Proposed Methodology

In this work, we develop speaker-dependent models for generating real-time MRI sequences directly from text. Our methodology employs a two-stage pipeline, as illustrated in Figure 1. In the first stage, speech is synthesized from the input text using tortoise-tts Text-to-Speech model, ensuring that the generated speech aligns with the characteristics of the target speaker. The second stage involves mapping the synthesized speech to rtMRI sequences using speaker-dependent models trained on specific speech-MRI data. Finally, the synthesized speech and corresponding rtMRI sequences are combined to create a synchronized rtMRI video, capturing both the articulatory movements and the acoustic output.

A. Text to speech synthesis

In this work, we employ Tortoise TTS [16], a state-of-the-art Text-to-Speech (TTS) system, to generate speech that mimics the voice characteristics of specific speakers by using reference audio samples. Tortoise is a highly expressive Text-to-Speech system known for its advanced voice cloning capabilities. It utilizes an autoregressive acoustic model, similar to generative pre-training (GPT) architectures, to convert input text into discretized acoustic tokens. These tokens are then transformed into mel-spectrogram frames using a diffusion model. This

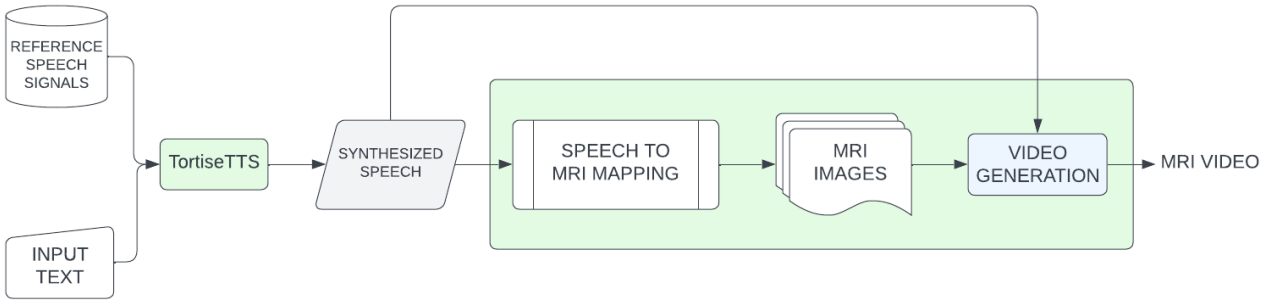


Fig. 1: flowchart of our Pipeline model

system achieves state-of-the-art performance in voice cloning, offering highly realistic prosody and intonation, closely replicating the natural characteristics of the target speaker's voice. To begin, the Tortoise TTS repository is cloned into a local environment, providing the necessary codebase for generating speech from text. The system is then supplied with reference audio files containing speech from the target speaker. The key voice attributes such as pitch, tone, rhythm, and other prosodic features are captured by the model during adaptation. Once the voice profile of the speaker is established, the system accepts text input and synthesizes speech that closely resembles the target speaker's voice. The synthesized speech maintains the distinct voice characteristics identified from the reference audio, producing high-fidelity output that accurately reflects the speaker's identity.

The Tortoise-TTS model is trained on all available audio files for each speaker, excluding the test set. This training enables the system to generate high-quality synthetic speech that remains true to the reference speaker's voice, even when new text is provided. The synthesized speech is then passed to the subsequent stage for generating synchronized speech-to-video output.

B. Speech to video generation

We employed a bidirectional Long Short-Term Memory (BLSTM) model to map the acoustic features to the rtMRI images. This model was chosen due to its proven effectiveness in providing state-of-the-art performance for acoustic-to-articulatory mapping tasks. To create the acoustic inputs for the model, we extracted set of 24 mel-frequency cepstral coefficients (MFCC) from the speech signal, with a frame shift of 43.14 ms corresponding to (1/23.18 fps). The MFCCs were selected as the input features as they are shown to result in lowest error in acoustic-to-articulatroy mapping. These spectral features served as the input for training the BLSTM network. The network architecture used for mapping the accostic features to the rtMRI image is shown in the figure 2. BLSTM model architecture consists of multiple layers designed to process acoustic features and predict the corresponding MRI images. It begins with a three layers of fully connected netowrks containing 575 neurons each, using ReLU activation to capture the detailed representations of the input features. The model then incorporates two BLSTM layers, both with 575 neurons and ReLU activation. The BLSTM layers capture temporal patterns in the data. Finally, the output layer is a fullly connected layer with a linear activation function, predicting the MRI frames with dimensions corresponding to the width and height of the rtMRI images. A separate speaker-dependent speech-to-MRI model was trained for each of the four speakers selected for our study.

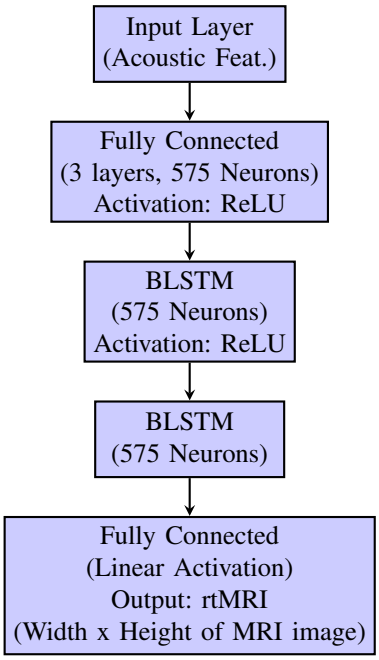


Fig. 2: BLSTM Model Architecture for MRI Prediction

C. Generating videos from the text using the trained models

Figure 3 illustrates the process of generating videos from text using trained models. The Tortoise TTS model, which has

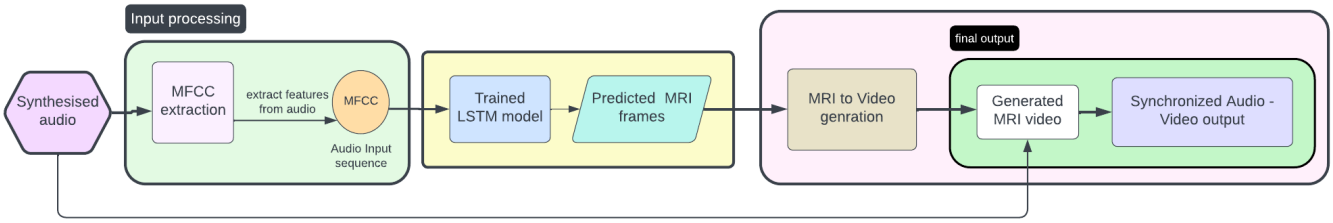


Fig. 3: The Figure shows the flow for Generating Videos

been trained on reference input recordings, generates speech signals that correspond to the provided text, emulating the speaker's voice characteristics. Acoustic features are extracted from the synthesized speech. The trained speech-to-MRI mapping model then produces the rt-MRI sequence based on the input acoustic features. Finally, the rt-MRI image sequence and the synthesized speech signal are combined to create a video corresponding to the input text.

V. Experimental Setup

We utilized data from four speakers (two male and two female) from the rtMRI dataset, which comprises 460 sentences of speech, along with their corresponding audio recordings and text transcriptions. The dataset was divided into two subsets: 410 samples were designated for training, while 50 samples were reserved for testing. This division was consistently applied throughout the experimental process, encompassing both the Text-to-Speech conversion and the speech-to-rtMRI tasks.

Both the Text-to-Speech and speech-to-MRI models were trained as speaker-dependent models, with a separate model developed for each of the four speakers. In the first stage, the Tortoise TTS model was trained using the 410 speech sentences as reference recordings, allowing it to synthesize speech for the 50 test sentences. In the second stage, the speech audio was processed to extract relevant features necessary for video generation. The speech was sampled at 20 kHz, and the window size for feature extraction was set to 25 milliseconds, with a frame shift of 43.14 milliseconds corresponding to the 23.18 frames per second of the rtMRI data. A set of 24 MFCCs was obtained from each speech segment. Subsequently, the speech-to-MRI mapping model was trained using the Adam optimization algorithm, with mean squared error loss and early stopping applied to the validation loss over five steps.

To produce the final output video, the predicted frames were compiled into a continuous sequence using OpenCV at a frame rate of 23.18 fps. This approach ensured smooth transitions between frames, thereby maintaining temporal coherence with the input speech. FFmpeg was employed to synchronize the original audio with the generated rtMRI frames. The resulting videos effectively demonstrated the MRI-based vocal tract movements corresponding to the provided speech audio.

The results of our proposed approach were evaluated using two distinct methods: subjective and objective evaluation. In the subjective evaluation, human participants rated the naturalness and coherence of the generated videos. For the objective evaluation, we used the Fréchet Inception Distance (FID) [17] metric to assess the visual quality of each generated video in comparison to real data. This combination of perceptual and quantitative assessments provided a thorough evaluation of the system's performance.

1) Subjective Evaluation: For the subjective evaluation, we engaged four evaluators—two with prior knowledge of the work and two without to ensure a balanced assessment of the generated videos. This diverse evaluator pool allowed us to capture a range of perspectives, enhancing the reliability of our findings. We focused on five key features essential for assessing video quality: naturalness, realism, clarity, temporal smoothness, and overall satisfaction.

Each evaluator rated videos corresponding to speakers M2, M3, F1, and F2. The evaluation employed a structured scale from 1 to 5, where a score of 1 indicated poor quality with significant flaws, while a score of 5 represented excellent quality, reflecting videos that met high standards of realism and clarity. This rating scale provided a clear framework for evaluations and enabled nuanced feedback.

Upon collecting the ratings, we calculated the average score for each feature across all evaluators, offering a comprehensive understanding of the model's performance across various quality dimensions. These average scores were critical in identifying strengths and areas for improvement. The final results of this evaluation are summarized in Table 1, which presents the overall Mean Opinion Score (MOS) values for the five features. This table is a vital component of our analysis, providing insights into the effectiveness of the model in generating realistic and high-quality videos.

2) Objective Evaluation: For the Objective evaluation, we evaluated the performance of our rtMRI video generation model using the Frechet Inception Distance (FID) metric. FID is commonly used in generative models to quantify how similar the distribution of generated samples is to that of real samples. We utilized the FID as our evaluation metric because there are no parallel rtMRI images available in the ground truth and the generated sequence due to the mismatch in the duration of the phonemes in the original speech and the synthesized speech. Lower FID scores indicate better performance, reflecting that the generated videos are closer

to the ground truth in terms of visual features.

We computed the FID scores for each speaker's video dataset. The generated videos were compared to their corresponding ground truth videos on a per-speaker basis.

VI. Results and Discussion

Results of the subjective and objective evaluation are presented below

1) Subjective Evaluation: Evaluators rated the videos for speakers M2, M3, F1, and F2 across five features: naturalness, realism, clarity, temporal smoothness, and overall satisfaction, using a scale of 1 to 5, where 1 indicated poor quality and 5 indicated excellent quality. The average scores, summarized in Table 1, highlight key insights. Clarity was a notable strength, with scores ranging from 3.47 (F1) to 3.66 (M2), indicating consistent visual sharpness. Naturalness and smoothness showed more variability, with M3 scoring 3.56 in naturalness, while M2 was lower at 3.4, suggesting room for improvement in making the videos appear more natural. Realism scores were moderate, between 3.425 (M2) and 3.6 (M3), indicating that while videos were generally realistic, further refinement is needed, particularly for M2.

Overall satisfaction scores were fairly close, reflecting consistent quality but also areas needing enhancement. These results provide clear direction for future improvements, focusing on boosting naturalness, smoothness, and realism to achieve a higher Mean Opinion Score across all features.

Figure 4 and figure 5 show the image sequences of the ground truth and the generated ones for words "Agriculture" and "Fresh" respectively. It should be observed that there is no exact one-to-one alignment between the two as the phoneme durations are difference between the original speech recoding in the rtMRI dataset and that generated using the Tortoise-TTS. We can see that there is good similarity in the images particularly for the specific shapes of the consonants such as 'k', 'ch', and 'sh'. We have obtained the consonant images for the easy comparison. Figure 6 shows the ground truth images from rtMRI dataset and generated images for consonants /ch, sh, p/. We can see that the generated images match closely with the ground-truth images.

TABLE I: Mean Opinion Scores (MOS) for Evaluated Features

attr/speaker	M2	M3	F1	F2
naturalness	3.4	3.56	3.39	3.38
realness	3.43	3.6	3.46	3.5
clarity	3.67	3.66	3.47	3.65
smoothness	3.5	3.48	3.47	3.4
overall	3.5	3.49	3.5	3.5

MOS for the generated videos evaluated across five key features: naturalness, realism, clarity, temporal smoothness, and overall satisfaction. Scores represent the average ratings from four evaluators—two with prior knowledge of the project and two without—across speakers M2, M3, F1, and F2.

2) Objective Evaluation: Table 2 presents the FID scores for the samples of each individual speaker. We computed the FID scores for the video datasets corresponding to each speaker, comparing the generated videos with their respective ground truth videos on a per-speaker basis. The table includes FID scores for ten individual speakers, revealing that the mean values range from 94.23 to 135.2, with standard deviations varying from 5.83 to 13.09.

To assess the significance of these scores, we obtained FID scores for the image sequences generated by [14], as reported in [18]. Since the text sequences used to generate these videos were not provided by the authors, it is not feasible to calculate the distances for these videos. The mean values from [14] ranged from 75.50 to 97.57, with standard deviations between 8.88 and 13.27, indicating lower values than those observed in our study. However, our scores are relatively comparable to theirs. The higher FID scores in our case may be attributed to the longer duration of the videos.

TABLE II: Summary Statistics of MRI FID distances for Different Speakers

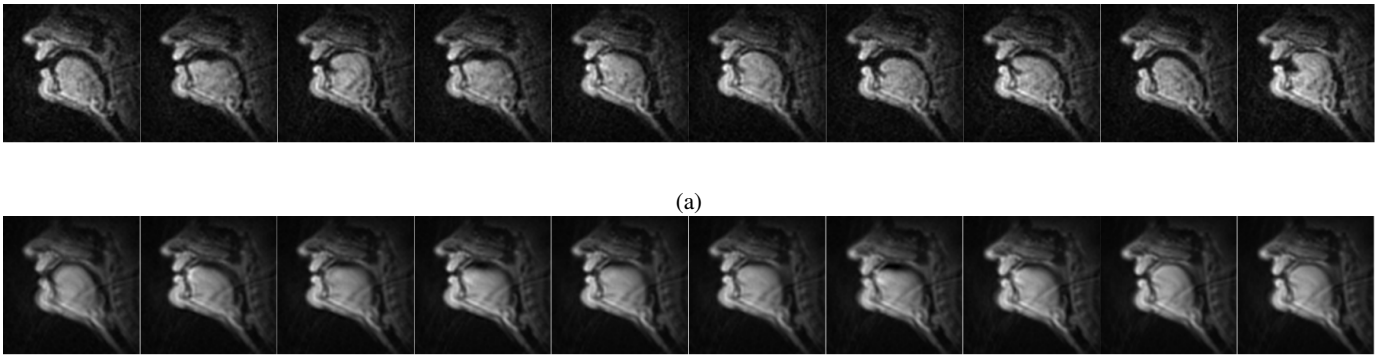
Test samples	M2	M3	F1	F2
sample 1	104.28	126.57	111.02	129.54
sample 2	107.22	85.33	95.92	134.86
sample 3	100.99	93.50	107.16	118.27
sample 4	105.23	83.07	94.23	146.53
sample 5	99.51	88.41	110.05	128.20
sample 6	104.38	110.35	112.02	148.95
sample 7	110.01	87.89	93.53	143.79
sample 8	120.14	85.73	96.45	143.43
sample 9	108.83	89.18	106.53	142.35
sample 10	102.52	100.38	104.37	120.97
Mean	106.38	94.23	102.89	135.28
Std Dev	5.83	13.09	6.98	10.39

Fid values for the Generated videos and their corresponding Ground-truth test samples for every speaker . Mean and standard deviation are calculated for each speaker.(rounded upto 2 decimals)

Conclusion

In this study, we introduced a novel method for synthesizing real-time MRI sequences from text. By employing a two-stage pipeline that integrates Text-to-Speech synthesis with speech-MRI mapping, we successfully generated synchronized MRI videos that depict articulatory movements in conjunction with their corresponding acoustic outputs. This approach addresses the challenge of acquiring MRI data for every word, which is both time-consuming and impractical. A key advantage of our method is that it eliminates the need for forced alignments during the training of our models. Our evaluation comprised both subjective and objective measures. Subjective assessments indicated that the synthesized videos displayed good clarity and overall satisfaction compared to ground truth videos. Objective assessments further validated the quality of the generated sequences, demonstrating a close resemblance to natural recordings.

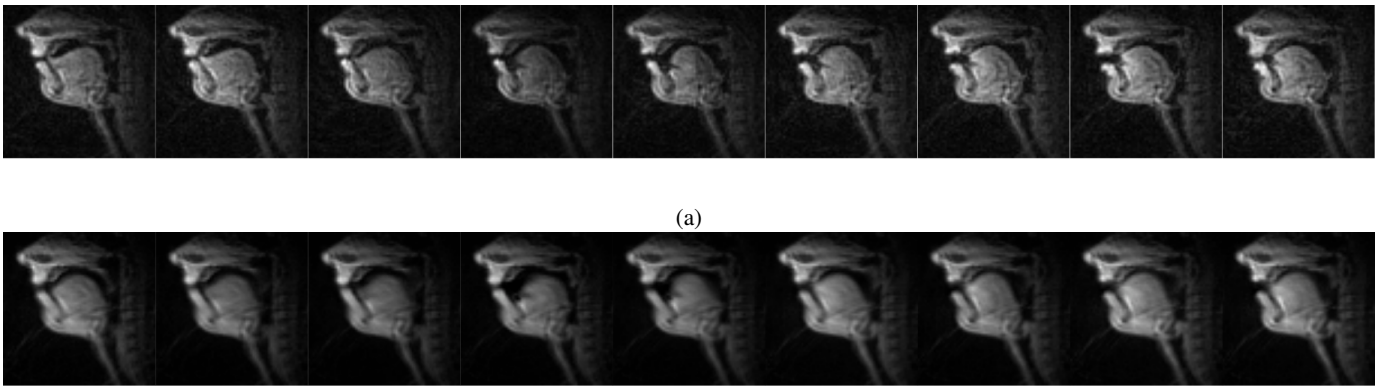
Overall, our findings suggest that this synthesized rtMRI approach holds significant potential for developing tools aimed



(a)

(b)

Fig. 4: Comparison of ground-truth and generated video frames for the word ”Agricultural: a) Ground truth sequence, b) Generated sequence ”.



(a)

(b)

Fig. 5: Comparison of ground-truth and generated video frames for the word ”Fresh: a) Ground truth sequence, b) Generated sequence ”.

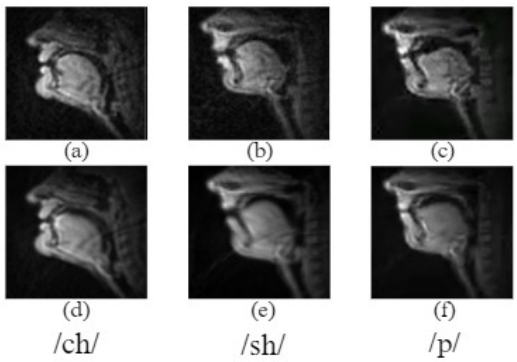


Fig. 6: Comparison of ground-truth and generated video frames for the consonant sounds ’/ch’,’/sh’,’/p/’ : a, b, c - original video frames, d, e, f - generated video frames.

at aiding pronunciation training. Future work will focus on refining the synthesis process to enhance the naturalness and realism of the videos, ultimately leading to more effective educational resources for children with hearing impairments.

REFERENCES

[1] O. Engwall, "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher," *Computer Assisted Language Learning*, vol. 25, pp. 37–64, 2012.

[2] I. Wilson, "Using ultrasound for teaching and researching articulation," *Acoustical Science and Technology*, vol. 35, pp. 285–289, 2014.

[3] W. F. Katz and S. Mehta, "Visual feedback of tongue movement for novel speech sound learning," *Frontiers in Human Neuroscience*, vol. 9, pp. 1–13, 2015.

[4] J. R. Westbury, *X-ray Microbeam Speech Production Database User's Handbook*, Version 1.0, Waisman Center on Mental Retardation and Human Development, University of Wisconsin, Madison, Wisconsin, USA, 1994. [Online]. Available: https://berkeley.app.box.com/v/xray-microbeam-database-data/

[5] J. Zhang, *Articulograph AG100 Electromagnetic Articulation Analyzer*, Version 1.2, UCLA Phonetics Lab, Los Angeles, California, USA, 1997. [Online]. Available: https://phonetics.linguistics.ucla.edu/facilities/physiology/Emamual.html

[6] P. Bacsfalvi and B. M. Bernhardt, "Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography," *Clinical Linguistics Phonetics*, vol. 25, pp. 1034–1043, 2011.

[7] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 100, pp. 537–554, 1996.

[8] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, 2014.

[9] I. Douros, J. Felblinger, J. Frahm, K. Isaieva, A. Joseph, Y. Laprie, F. Odille, A. Tsukanova, D. Voit, and P. A. Vuissoz, "A multimodal real-time MRI articulatory corpus of French for speech research," in *INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association*, 2019.

[10] A. Toutios, T. Sorensen, K. Somandepalli, R. Alexander, and S. S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data," in *Interspeech*, 2016, pp. 1492–1496.

[11] V. Ramanarayanan, L. Goldstein, D. Byrd, and S. S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, 2013.

[12] U. Desai, C. Yarra, and P. Ghosh, "Concatenative articulatory video synthesis using real-time MRI data for spoken language training," in *ICASSP*, 2018, pp. 4999–5003.

[13] S. Chandana, C. Yarra, R. Aggarwal, S. K. Mittal, N. K. Kausthubha, K. T. Raseena, A. Singh, and P. K. Ghosh, "Automatic visual augmentation for concatenation-based synthesized articulatory videos from real-time MRI data for spoken language training," in *Proc. Interspeech*, 2018.

[14] S. Udupa and P. K. Ghosh, "Real-Time MRI Video Synthesis from Time Aligned Phonemes with Sequence-to-Sequence Networks," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10094797.

[15] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and Y. Laprie, "Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated," *Speech Communication*, vol. 141, pp. 1-13, 2022. doi: 10.1016/j.specom.2022.04.004.

[16] J. Betker, "Better speech synthesis through scaling," 2023. [Online]. Available: https://arxiv.org/abs/2305.07243. [Accessed: Oct. 20, 2024].

[17] M. Seitzer, "pytorch-fid: FID Score for PyTorch," Aug. 2020. [Online]. Available: https://github.com/mseitzer/pytorch-fid. [Accessed: Oct. 20, 2024].

[18] S. Udupa, "text_to_rtmri_synthesis," GitHub, 2023. [Online]. Available: https://github.com/bloodraven66/textToRtmriSynthesis. [Accessed: Oct. 20, 2024].