Computer Speech and Language 99 (2026) 101928

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Speech acoustics to rt-MRI articulatory dynamics inversion with video diffusion model

Xuan Shi [a,*], Tiantian Feng [a], Jay Park [a], Christina Hagedorn [b], Louis Goldstein [c], Shrikanth Narayanan [a,c]

[a] Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, 3740 McClintock Ave, Los Angeles, 90089, CA, United States
[b] College of Staten Island (CUNY), 2800 Victory Blvd, Staten Island, 10314, NY, United States
[c] Department of Linguistics, University of Southern California, 3601 Watt Way, Los Angeles, 90089, CA, United States

ARTICLE INFO

Keywords:
Speech production
Acoustic to articulatory
Diffusion model
Video generation

ABSTRACT

Inverting speech acoustics to articulatory dynamics presents a multidisciplinary challenge, spanning clinical, linguistic, and engineering domains, with applications including speech therapy and second language learning. Despite its significance, existing methods lack a systematic approach for generating articulatory dynamics of a more complete vocal tract from speech acoustics. Availability of spatio-temporally rich video covering the entire oro-pharynx and laryngeal region of the vocal tract during speech at high frame rates (83 frames/second) using real-time MRI (rt-MRI), alongside linguistic-theory guided computational frameworks, offers new possibilities to improve speech to articulatory inversion. In this work, we propose a novel system for inverting speech acoustics to articulatory dynamics using an rt-MRI driven video diffusion model. Additionally, we introduce a new evaluation method, a linguistic knowledge guided pixel intensity correlation within articulatory regions of interest (ROIs), for video generative model. Our results demonstrate the system's competitive performance in generalizing to unseen reference speech audio.

1. Introduction

The Acoustic-to-Articulatory Inversion (AAI) problem refers to the process of inferring articulatory dynamics from speech signals. As the inverse of the speech production process of generating acoustics from articulation, AAI has a wide range of practical applications in fields such as speech understanding, speech sound disorder intervention, second language acquisition, and even complementing audio information in improving speech and speaker recognition (Li et al., 2016; Ghosh and Narayanan, 2011). Previous studies have explored various acoustic and articulatory representations and models for the AAI task from the classic statistical ones (Ghosh and Narayanan, 2010, 2013) to contemporary neural approaches (Cho et al., 2024; McGhee et al., 2024; Yang et al., 2024; Csapó, 2020; Oura et al., 2024; Udupa and Ghosh, 2023; Nguyen et al., 2024). Specifically, articulatory information has been represented by methods such as Electromagnetic Articulography (EMA) (Cho et al., 2024; McGhee et al., 2024), ultrasound (Yang et al., 2024), X-ray (McGhee et al., 2024), or real-time magnetic resonance imaging (rt-MRI) signals (Csapó, 2020; Oura et al., 2024; Udupa and Ghosh, 2023; Nguyen et al., 2024). Among these, rt-MRI offers a unique advantage in providing a spatio-temporally rich video covering the mid-sagittal plane of oro-pharynx and laryngeal region of the vocal tract during speech at

* Corresponding author.
E-mail address: xuanshi@usc.edu (X. Shi).

https://doi.org/10.1016/j.csl.2025.101928
Received 7 April 2025; Received in revised form 18 December 2025; Accepted 18 December 2025
Available online 31 December 2025
0885-2308/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

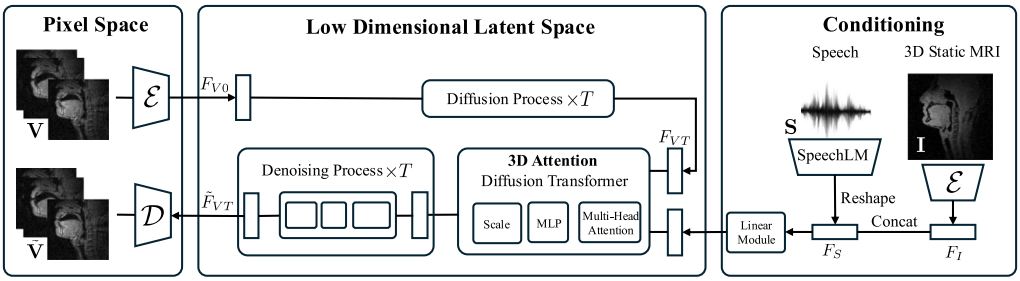


Fig. 1. Overall architecture.

high frame rates. However, many existing rt-MRI-based AAI methods suffer from poor generalization to unseen speech or speakers, and fail to accurately and objectively evaluate generation performance specific to different articulators.

In this work, we develop a novel AAI system that generates detailed articulatory vocal-tract rt-MRI image sequences using acoustic speech and 3D static MRI as input. We hypothesize that both speech and 3D static MRI provide complementary acoustic and structural (volumetric) information, which can enhance generalization to unseen reference speech and speakers. To this end, we combine the extracted speech and 3D static MRI features as conditioning, and use a latent diffusion model (LDM) to progressively generate the target rt-MRI sequences in latent space. The conditioning inputs facilitate the control of the generation from unseen reference, while the LDM allows efficient and accurate video generation.

In most existing video generation studies, the quality of the generated video is typically evaluated using subjective scores or perceptual scores, like FVD (Unterthiner et al., 2019), SSIM (Wang et al., 2004). However, in the scenario of speech articulatory generation, as the key controlled variable, the accuracy of articulatory constriction estimation is more important for further linguistic analysis and practical applications. In this work, we introduce the region of interests (ROI) pixel intensity correlation for the articulation constriction location and degree accuracy evaluation. We manually annotated a subset of the rt-MRI dataset as a reference for evaluation. The annotations are planned to be shared with the community for public use. Our major contributions are as follows.

- We develop a speech-driven articulation rt-MRI generation system, closing the gap between acoustic information and high-quality articulation data inversion.
- We incorporate linguistic domain knowledge into video generation evaluation, rather than relying solely on perceptual scores, enabling a linguistically informed interpretation of model performance. Additionally, we share our articulator annotation data to support further research.
- Our evaluation results demonstrate the effectiveness of the model in modeling and generating articulator dynamics. Moreover, the model shows strong generalization capability on unseen reference speech, indicating the potential application to low-resource AAI tasks.

2. Related works

Acoustic to articulatory inversion has a rich history in speech analysis, and continues to be actively investigated in recent years with various combinations of input and output formats and model types (Cho et al., 2024; McGhee et al., 2024; Yang et al., 2024; Csapó, 2020; Oura et al., 2024; Udupa and Ghosh, 2023). Specifically, acoustic representations, such as direct waveform, hand-crafted acoustic spectrogram, or hidden states extracted from self-supervised speech model, are leveraged to generate the corresponding articulatory feature, which is generally represented by EMA (Cho et al., 2024; McGhee et al., 2024), ultrasound (Yang et al., 2024), X-ray (McGhee et al., 2024), or (rt-)MRI signals. For example, Cho et al. (2024) uses linear mapping to recover articulatory kinematics in EMA from the representation of various SSL models. This study empirically demonstrates the existence of articulatory phonology information in self-supervised speech learning. McGhee et al. (2024) proposes to embed the AAI into an "acoustic (speech) $\mapsto$ articulatory (X-ray/EMA) $\mapsto$ acoustic (speech)" pipeline and evaluate the performance of AAI with the WER of the synthetic speech. Their AAI approach also shows the possibility of synthesizing intelligible speech from unseen languages. Targeting the ultrasound tongue imaging (UTI) generation, Yang et al. (2024) uses a diffusion model to generate articulatory UTI with textual and audio input. Compared with EMA, ultrasound, and X-ray, rt-MRI has gained increasing attention in speech research mainly owing to the detailed imaging of the whole vocal tract (Csapó, 2020; Oura et al., 2024; Udupa and Ghosh, 2023). Csapó (2020) uses LSTM and DNN-based models to generate speaker-dependent rt-MRI from the Mel-generalized cepstral linear spectral pairs (MGC-LSPs) acoustic feature. This system shows competitive generation results that are similar to the ground truth rt-MRI recordings measured by structural similarity index (SSIM). Oura et al. (2024) adapts the model and input/output features similarly to Csapó (2020) for the Japanese speech AAI task, introducing a preprocessing pipeline for normalization and filtering of rt-MRI data. Without the explicit acoustic feature, Udupa and Ghosh (2023) takes phoneme sequences as input to generate articulatory rt-MRI movies for each speaker. In particular, Udupa and Ghosh (2023) demonstrates that their AAI system is capable of synthesizing rt-MRI from audio, achieving a minimum mean opinion score of 3.3 on a 5-point scale.

3. Method

3.1. Problem formulation and hypothesis

In this work, we define the input triplet data point: [speech waveform **S**, rt-MRI sequence **V**, and 3D static MRI **I**], specifically, $\mathbf{S} \in \mathbb{R}^{t_S}$, $\mathbf{V} \in \mathbb{R}^{t_V \times h_V \times w_V \times c}$, and $\mathbf{I} \in \mathbb{R}^{h_I \times w_I \times c}$. t_S and t_V are the time scale of speech and video, respectively, while h, w, and c denote the height, width, and channel of each image. We select the mid-sagittal frame from the whole 3D static volumetric scan as our **I** input. It is worth highlighting that speech and rt-MRI video originate from the same recording but their time scales t_S and t_V are different in value due to the distinct speech and video sampling rates.

In our hypothesis, speech contains acoustic and semantic information, while 3D static MRI conveys anatomical details of the speaker's dynamic vocal tract shaping. Therefore, in this work, we aim to develop a speaker-independent model to generate articulatory rt-MRI videos **V** using the conditioning inputs of speech **S** and the speaker's 3D static MRI **I**.

3.2. Acoustic feature extraction from the SSL speech model

Self-supervised learning (SSL) speech models have been shown to capture rich acoustic and articulatory features at the token level across a variety of downstream tasks (Cho et al., 2024; Shi et al., 2024). Given a speech waveform **S**, we use a pre-trained SSL speech model to extract the acoustic feature tokens $F_S \in \mathbb{R}^{t_{F_S} \times d_{F_S}}$ from model's hidden states, where t_{F_S} and d_{F_S} are the number and dimension of the extracted acoustic feature tokens, respectively. In this work, we employ SpeechLM (Zhang et al., 2024) as our token-level acoustic feature generator because it encodes phonetic and semantic information.

3.3. Diffusion model-based articulatory rt-MRI generation

Our video generation is established on the latent diffusion model (LDM) (Rombach et al., 2022). As shown in Fig. 1, the overall articulatory rt-MRI generation system has three components: the pixel compression and reconstruction module, the latent diffusion generation module, and the conditioning module. We use the CausalVideoVAE (PKU-Yuan Lab et al., 2024), comprised of the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$, to compress video to the latent space $F_{V0} = \mathcal{E}(V)$ and to reproduce the video from the latent space $\tilde{V} = \mathcal{D}(\tilde{F}_{VT})$, where $\tilde{F}_{VT}$ is the estimated latent feature, bridging the video from pixel space to the latent space. CausalVideoVAE uses Causal3DConv (Yu et al., 2024) to conduct up (down)-sampling on both temporal and spatial scale.

In the latent space, there are two processes: diffusion process and denoising process. The diffusion process adds Gaussian noise into the data gradually F_{V0}: $q(F_{Vt}|F_{Vt-1}) = \mathcal{N}(F_{Vt}; \sqrt{1-\beta_t}F_{Vt-1}, \beta_t I)$ for T steps and finally get an isotropic Gaussian noise, where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is the noise scheduler. The denoising process iteratively predicts the distribution of the less noisy data F_{Vt-1}: $p_\theta(F_{Vt-1}|F_{Vt}, F_S, F_I) = \mathcal{N}(F_{Vt-1}; \mu_\theta(F_{Vt}, t, F_S, F_I), \beta_t I)$ with the conditionings of speech feature F_S and static MRI feature F_I, where μ_θ is the diffusion model. In the denoising process, we follow the setting from Ma et al. (2024) and used the combination of the mean square error (MSE) loss L_{MSE} and variance low bound loss L_{VLB} (Nichol and Dhariwal, 2021) as objective function, where L_{VLB} is used to decode the diffusion process, and L_{MSE} is reconstruction loss, where $\epsilon \sim \mathcal{N}(0, I)$:

$$\begin{aligned} L_{VLB} &= -\log p_\theta(F_{V0}|F_{V1}) \\ &\quad + D_{KL}(q(F_{Vt-1}|F_{Vt}, F_{V0}) \parallel p_\theta(F_{Vt01}|F_{Vt})) \\ L_{MSE} &= \mathbb{E}_{F_{V0}, \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta(F_{Vt}, t, F_S, F_I)\right\|_2^2\right] \end{aligned} \tag{1}$$

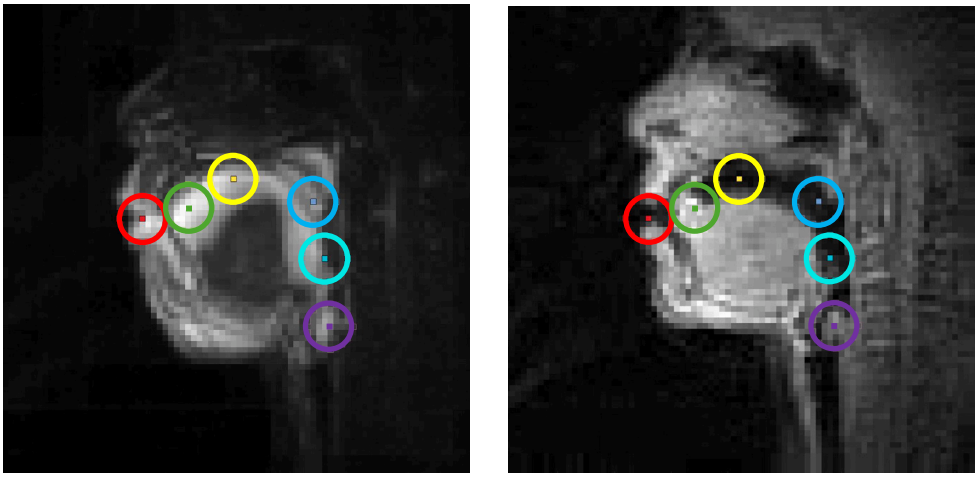
Specifically, in the conditioning of LDM, we first reshape F_S into $F'_S \in \mathbb{R}^{t_{F_V} \times w_{F_S} \times d_{F_S}}$ to match the dimensions of the video latent feature. Here, $w_{F_S} = t_{F_S}/t_{F_V}$ represents the down-sampling ratio from F_S to F_V. We then concatenate F'_S with F_I to form the conditioning input. After passing through a linear layer-based perceptual resampler, the concatenated conditioning is injected into the diffusion model via cross-attention. Notably, as certain phonemes (such as stops) involve silent constriction periods (Browman and Goldstein, 1989), we extend the input speech window w_{F_S} by a 200 ms look-ahead to fully capture the speech production gesture.

4. Dataset and annotations

4.1. Dataset

We conduct all the inversion experiments and evaluation on a large publicly available articulatory rt-MRI dataset, the USC 75-Speaker Dataset (Lim et al., 2021), which is a widely used dataset in speech science research studies. The dataset is comprised of speech from 75 individuals. Each speaker has provided 17 min of speech in English, including read consonant and vowel sequences (VCV/bVt), sentences, phonetically balanced passages, and free speech elicited through picture descriptions and topic introductions. Each acquisition is recorded with synchronized audio and rt-MRI. Furthermore, the dataset includes 3D static MRI and T2-weighted MRI, providing vocal tract volumetric information.

Considering the 75-Speaker dataset does not provide a standard data partition, we divided the dataset into training, validation, and test sets. To evaluate the model's generalization capability on unseen speakers' speech, we carefully selected 8 gender-balanced speakers to form an unseen speaker dataset. For all speakers in seen and unseen categories, we further stratified 10% of the speech acquisitions as unseen speech, covering vowel-consonant and passage reading and spontaneous speech. We formed the validation and test sets with non-overlapping combinations of unseen speakers and speech.



(a) Annotations on the standard deviation (SD) figure. (b) Annotations on one rt-MRI frame.

Fig. 2. ROI selection and illustration. 2(a): ROIs are selected to capture the region with the maximum articulatory dynamic intensity, as highlighted in the SD figure of the acquisition. 2(b): Each ROI should cover the target articulator constriction.

Table 1
Perceptual score – FVD for all systems. "VCV/bVt" are the two specific types of sequences in the consonant-vowel sequences. "VCV" denotes the vowel-consonant-vowel sequences, while "bVt" denotes /b/-vowel-/t/ sequences. "↓"/"↑" denote lower/higher is better.

Inference data type			FVD ↓		
Unseen speaker	Unseen speech	Speech type	FILM	Same-ref.	Cross-ref.
	✓	VCV/bVt	717 ± 334	1175 ± 483	1215 ± 483
	✓	Running	1014 ± 426	1218 ± 484	1252 ± 476
✓		VCV/bVt	755 ± 332	1401 ± 531	1555 ± 600
✓		Running	907 ± 323	1519 ± 533	1678 ± 571
✓	✓	VCV/bVt	726 ± 293	1458 ± 513	1593 ± 580
✓	✓	Running	921 ± 300	1533 ± 548	1778 ± 568

4.2. Annotation availability

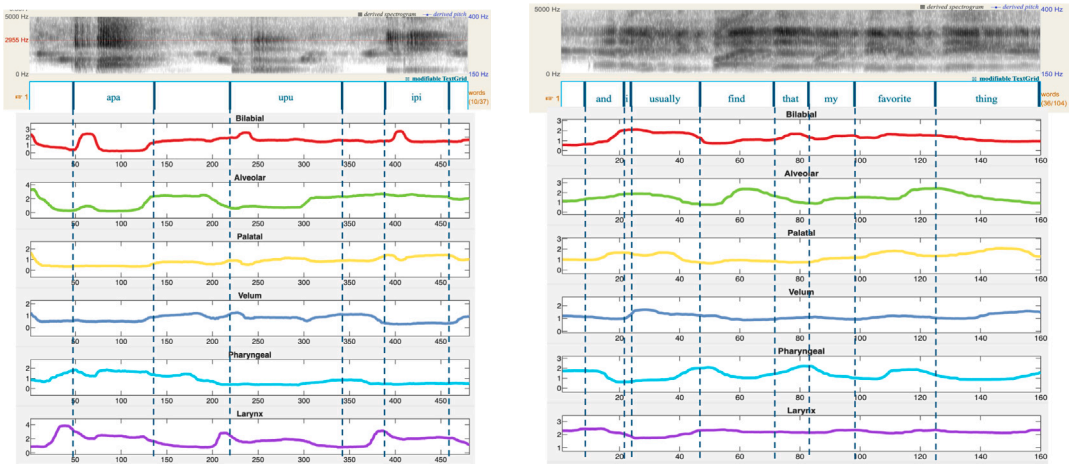
RT-MRI contains high-dimensional information in spatial and temporal, and brings unique challenges for analysis and evaluation. A widely used practice to tackle the challenge of extracting speech production information is to use the region of interests (ROI) analysis method: extract the pixel intensity sequence in each ROI and analyze their pixel intensity changing sequence. This method captures the articulatory dynamics in speech production, while avoids labor intensive articulation boundary segmentation. The pixel intensity within each ROI is represented as time-series data throughout speech production. As illustrated in Fig. 3, distinct articulatory patterns emerge between VCV/bVt and running speech. The consonant-vowel sequence (VCV/bVt) (Fig. 3(a)) exhibits precise, controlled articulator movements, whereas spontaneous and reading speech (running) (Fig. 3(b)) reveals more complex coarticulation patterns with reduced motor control. In the ROI pixel intensity extraction (Fig. 2), we select six articulation regions to cover the oral vocal tract: bilabial, alveolar, palatal, velum, pharyngeal, and laryngeal. Following previous studies (Lammert et al., 2010; Oh and Lee, 2018; Blaylock, 2021), each ROI is selected as a circular region, centered at the most active region, with a radius of 4 pixels, covering the maximum degree of articulatory constriction. Each annotation is provided by one annotator and checked by two additional annotators. Annotators are trained in phonetics and speech science. The ROI annotations are publicly available[1] as complementary material for the 75-Speaker database to support articulatory analysis.

5. Experiments

5.1. Experimental setting

We form data triplets [**S**, **V**, **I**] for training and inference, as introduced in Section 3.1. We apply the Silero VAD from The Silero-Team (2024) to remove silence in each speech sample. In VCV and /bVt/ acquisitions, speakers are asked to leave a significant

[1] https://github.com/Alexuan/75speaker_annotation_ROI



(a) VCV/bVt Speech (b) Running Speech

Fig. 3. Examples of the time-aligned speech spectrum, word-level transcription alignment, and pixel intensity time sequence in ROIs.

Table 2
Perceptual score – SSIM for all systems. Naming is consistent with the Table 1. SSIM values ranging from 0 (no similarity) to 1 (identical).

Inference data type			SSIM ↑		
Unseen speaker	Unseen speech	Speech type	FILM	Same-ref.	Cross-ref.
	✓	VCV/bVt	0.785 ± 0.067	0.560 ± 0.079	0.466 ± 0.091
	✓	Running	0.783 ± 0.061	0.560 ± 0.083	0.477 ± 0.097
✓		VCV/bVt	0.795 ± 0.059	0.462 ± 0.081	0.319 ± 0.077
✓		Running	0.797 ± 0.053	0.459 ± 0.084	0.316 ± 0.084
✓	✓	VCV/bVt	0.796 ± 0.059	0.449 ± 0.078	0.312 ± 0.073
✓	✓	Running	0.795 ± 0.051	0.453 ± 0.081	0.313 ± 0.090

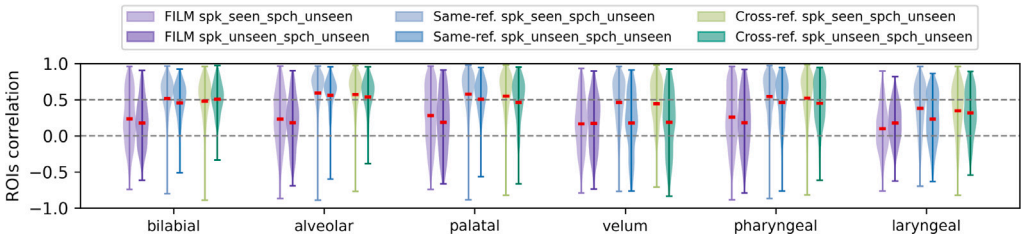
interval between each two phrases, shown in Lim et al. (2021), therefore we use VAD to segment out each consonant-vowel phrase. In passage reading and running speech acquisitions, as speakers are expected to speak continuously, we only apply VAD to remove the silence at the beginning and the very end of the audio. Considering the articulatory dynamics in speech planning, we preserved extra 200 ms ahead of each audio segment, aligning with the 200 ms look-ahead in Section 3.3. Due to computational limitations, we further segment each acquisition into video segments with 240 frames to facilitate training and inference.

We use SpeechLM (Zhang et al., 2024)[2] to encode speech into phoneme-based codec and Open-Sora-Plan (PKU-Yuan Lab et al., 2024) to implement the CausalVideoVAE and LDM for video generation. We first load the pre-trained parameters for all modules which are provided from their official releases,[3] then we fix the parameters of the SpeechLM and CausalVideoVAE while only fine-tuning the LDM on the training data.

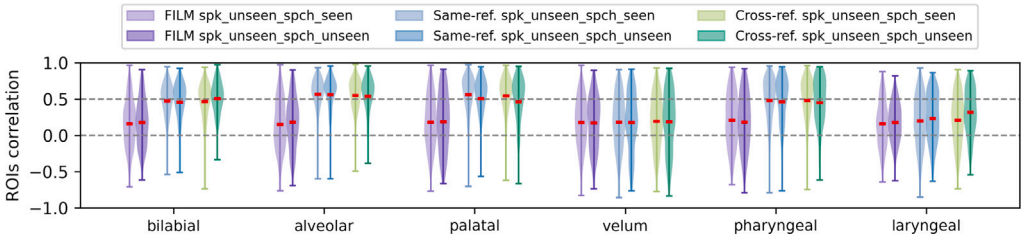
5.2. Video generation baselines

We select two types of baselines for model performance comparison. In the first baseline, we generate the rt-MRI video with the input of speech and the reference image taken from the rt-MRI instead of the 3D static MRI. This approach is referred to as **"same-reference"** (same-ref.) video generation, while our proposed approach represents **"cross-reference"** (cross-ref.) video generation. This first baseline is designed to assess its theoretical performance upper bound. The second baseline is the FILM model (Reda et al., 2022), selected from the field of frame interpolation-based video generation. This choice was necessitated by the lack of prior open-source works in speaker-independent, audio-driven rt-MRI generation. Although FILM is designed for synthesizing frames between two given inputs, we evaluate it by providing only the first and last frames of each test sequence. This setup serves as the closest possible approximation to our proposed method, which generates video from a single image input.

[2] https://github.com/microsoft/unilm/tree/master/speechlm
[3] https://huggingface.co/LanguageBind/Open-Sora-Plan-v1.1.0



(a) Inference on seen and unseen speakers speaking unseen speech.



(b) Inference on unseen speakers speaking seen and unseen speech.

Fig. 4. Violin plot of the correlation distributions between generative systems and groundtruth under different ablation inference conditions. Each system is represented by a distinct color theme, with different color tones indicating the ablation conditions. "Same-ref." and "Cross-ref." indicate same-reference and cross-reference systems, respectively. "spk_(un)seen_spch_(un)seen" denotes the model inference on (un)seen speaker and (un)seen speech.

5.3. Evaluation metrics

5.3.1. Pixel intensity correlation within ROIs

This evaluation method is devised to evaluate the dynamic constriction similarity between ground truth reference and synthetic articulation, which is an essential indicator in phonetic and clinical applications. We calculate pixel intensity in each linguistic knowledge-guided circular ROI for both ground truth and generated videos and calculate the correlation between them to evaluate their similarity in constriction location and degree.

5.3.2. Perception scores

Perception Scores are widely used in objective video quality assessment. We use the Fréchet Video Distance (FVD) (Unterthiner et al., 2019) and the Structural Similarity Index (SSIM) (Wang et al., 2004) to measure the similarity between ground truth and generated video in feature distribution and structure, respectively.

5.4. Experimental results

5.4.1. Pixel intensity correlation within ROIs

Figs. 4(a) and 4(b) show the ROI pixel intensity correlation distributions in different "unseen speaker" and "unseen speech" conditions.

As can be seen in Figs. 4(a) and 4(b), both same-ref. and cross-ref. systems show superior pixel intensity correlation within ROIs than the FILM in the pharyngeal region and front articulators: bilabial, alveolar, and palatal regions, even though they are based on different reference image input resource. In the "unseen speaker" condition, however, we can observe that the velar and laryngeal regions demonstrate significant correlation degradation. From the perspective of linguistic phonetics, constrictions formed by the tongue dorsum against the velum are known to vary substantially in location as a function of the flanking vowels. Additionally, unlike each of the other constriction locations examined (i.e., bilabial, alveolar, palatal, velar, pharyngeal), no English phonemes actively target mid-sagittal constrictions in the laryngeal region. These two factors which give rise to inherently increased variability in the velar and laryngeal regions likely result in the observed generalization difficulty on "unseen speaker". Under the "unseen speech" evaluation setting, the video generation model maintains decent pixel intensity correlations within ROIs, indicating that our model has the capability to generate reasonable rt-RMI under the low speech resource setting.

5.4.2. Perception score

Surprisingly, the baseline substantially outperforms the proposed systems in both FVD (Table 1) and SSIM (Table 2), which contrasts with the results of pixel intensity correlation within ROIs. While the perceptual scores and ROIs pixel intensity correlations

show differing performance trends, pixel intensity correlation is particularly relevant for our applications: It directly measures gesture constriction accuracy, making it more practical and meaningful for linguistic and clinical purposes.

Overall, the same-ref. system yields better performance than the cross-ref. system with a small margin in all types of experimental conditions. Both same-ref. and cross-ref. systems have better perceptual score performance on VCV and /bVt/ sequences than running speech, which may be because coarticulation is more pervasive in running speech than in these tightly controlled (VCV and /bVt/) targets. Moreover, compared with the video generation under the ‘‘unseen speaker and unseen speech’’ condition, both same-ref. and cross-ref. systems show greater improvement in perception scores when generating rt-MRI from ‘‘seen speaker’’ data compared to ‘‘seen speech’’ data. This aligns with the observations made in the ROI pixel intensity correlation (Fig. 4). However, this pattern is not observed in the FILM baseline system, likely due to their differing video generation mechanisms.

6. Conclusion

Acoustic-to-Articulatory inversion remains a significant open problem with broad implications across multiple applications. In this work, we leverage a latent diffusion model to generate rt-MRI from both speech and 3D static MRI inputs. The model is evaluated using both linguistic-based ROI pixel intensity correlations and perceptual scores. Our results demonstrate promising accuracy in generating articulatory constrictions, particularly for front articulators and pharyngeal region, and show robustness in generalizing to unseen speech data. Future work will focus on more controlled analyses to further assess the model's performance across individual articulators. We will also continue to refine the model's generation capabilities for improved accuracy and robustness.

CRediT authorship contribution statement

Xuan Shi: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Tiantian Feng:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis, Data curation. **Jay Park:** Writing – review & editing, Software, Resources, Data curation. **Christina Hagedorn:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis. **Louis Goldstein:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Shrikanth Narayanan:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used Gemini 3 in order to proofread. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Shrikanth Narayanan reports financial support was provided by National Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by The U.S. National Science Foundation under Grant NSF (IIS-2311676, BCS-2240349, RI-2106930). The authors alone are responsible for the content and conclusions. We also thank Dr. Yubin Zhang and Dr. Yijing Lu for the knowledgeable discussions.

References

Blaylock, R., 2021. reedblaylock/VocalTract-ROI-Toolbox. http://dx.doi.org/10.5281/zenodo.5130639.

Browman, C.P., Goldstein, L., 1989. Articulatory gestures as phonological units. Phonology 6 (2), 201–251.

Cho, C.J., Mohamed, A., Black, A.W., Anumanchipalli, G.K., 2024. Self-supervised models of speech infer universal articulatory kinematics. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 12061–12065.

Csapó, T.G., 2020. Speaker dependent acoustic-to-articulatory inversion using real-time MRI of the vocal tract. In: 21st Annual Conference of the International Speech Communication Association. Interspeech 2020, ISCA, pp. 3720–3724.

Ghosh, P.K., Narayanan, S.S., 2010. A generalized smoothness criterion for acoustic-to-articulatory inversion. J. Acoust. Soc. Am. 128 (4), 2162–2172. http://dx.doi.org/10.1121/1.3455847.

Ghosh, P.K., Narayanan, S.S., 2011. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. J. Acoust. Soc. Am. Express Lett. 130 (4), EL251–El257. http://dx.doi.org/10.1121/1.3634122.

Ghosh, P.K., Narayanan, S.S., 2013. On smoothing articulatory trajectories obtained from Gaussian mixture model based acoustic-to-articulatory inversion. J. Acoust. Soc. Am. Express Lett. 134 (2), EL258–EL264. http://dx.doi.org/10.1121/1.4813590.

Lammert, A.C., Proctor, M.I., Narayanan, S.S., et al., 2010. Data-driven analysis of realtime vocal tract MRI using correlated image regions. In: Interspeech. pp. 1572–1575.

Li, M., Kim, J., Lammert, A., Ghosh, P.K., Ramanarayanan, V., Narayanan, S.S., 2016. Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. 36, 196–211. http://dx.doi.org/10.1016/j.csl.2015.05.003.

Lim, Y., Toutios, A., Bliesener, Y., Tian, Y., Lingala, S.G., Vaz, C., Sorensen, T., Oh, M., Harper, S., Chen, W., et al., 2021. A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images. Sci. Data 8 (1), 187.

Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.-F., Chen, C., Qiao, Y., 2024. Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048.

McGhee, C., Knill, K., Gales, M., 2024. Highly intelligible speaker-independent articulatory synthesis. In: 25th Annual Conference of the International Speech Communication Association. Interspeech 2024, ISCA.

Nguyen, H., Foley, S., Huang, K., Shi, X., Feng, T., Narayanan, S., 2024. Speech2rtMRI: Speech-guided diffusion model for real-time MRI video of the vocal tract during speech. arXiv preprint arXiv:2409.15525.

Nichol, A.Q., Dhariwal, P., 2021. Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. PMLR, pp. 8162–8171.

Oh, M., Lee, Y., 2018. ACT: An automatic centroid tracking tool for analyzing vocal tract actions in real-time magnetic resonance imaging speech production data. J. Acoust. Soc. Am. 144 (4), EL290–EL296.

Oura, A., Kikuchi, H., Kobayashi, T., 2024. Preprocessing for acoustic-to-articulatory inversion using real-time MRI movies of Japanese speech. In: 25th Annual Conference of the International Speech Communication Association. Interspeech 2024, ISCA.

PKU-Yuan Lab, Tuzhan, A., et al., 2024. Open-sora-plan. http://dx.doi.org/10.5281/zenodo.10948109.

Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C., Curless, B., 2022. Film: Frame interpolation for large motion. In: European Conference on Computer Vision. Springer, pp. 250–266.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.

Shi, X., Feng, T., Huang, K., Kadiri, S., Lee, J., Lu, Y., Zhang, Y., Goldstein, L., Narayanan, S.S., 2024. Direct articulatory observation reveals phoneme recognition performance characteristics of a self-supervised speech model.

The Silero-Team, 2024. Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. https://github.com/snakers4/silero-vad.

Udupa, S., Ghosh, P.K., 2023. 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, IEEE, pp. 1–5.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S., 2019. FVD: a new metric for video generation. In: Deep Generative Models for Highly Structured Data. ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019, OpenReview.net.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Yang, Y., Su, R., Liu, X., Yan, N., Wang, L., 2024. An audio-textual diffusion model for converting speech signals into ultrasound tongue imaging data. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2170–2174.

Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., Gong, B., Yang, M., Essa, I., Ross, D.A., Jiang, L., 2024. Language Model Beats Diffusion - Tokenizer is key to visual generation. In: The Twelfth International Conference on Learning Representations. ICLR.

Zhang, Z., Chen, S., Zhou, L., Wu, Y., Ren, S., Liu, S., Yao, Z., Gong, X., Dai, L., Li, J., et al., 2024. Speechlm: Enhanced speech pre-training with unpaired textual data. IEEE/ACM Trans. Audio Speech Lang. Process.