Multimodal Pretraining for Vocal Tract Modeling

Anonymous CVPR submission

Paper ID 14145

Abstract

Accurate modeling of the vocal tract is necessary for naturalistic facial animation, avatar rendering for virtual environments, and pronunciation tutoring. However, vocal tract modeling is challenging because internal articulators like the tongue and velum are occluded from external motion capture technologies. Real-time magnetic resonance imaging (RT-MRI) offer a direct way to measure precise movements of internal articulators during naturalistic speaking, offering a possible solution to modeling. However, segmented and annotated datasets of MRI are limited in size due to time-consuming and computationally expensive labeling methods. We first present a labeling strategy for the Speech MRI Open Dataset comprising over 20 hours of audio-aligned RT-MRI video using a vision-only segmentation approach. We then apply a multimodal pretraining algorithm to: (1) improve segmentation of vocal articulators, (2) synthesize intelligible speech from inferred segments, and (3) animate 2D and 3D facial avatars that capture the complex articulator patterns underlying naturalistic speech production. We also extend these animation techniques to a high performance streamable facial avatar driven directly from speech, achieving 107ms average latency. Together, we set a new benchmark for vocal tract modeling in MRI image segmentation, intelligible MRI-to-speech synthesis, and real-time speech-to-avatar rendering.

1. Introduction

Vocal tract modeling is an essential technology in many applications including facial animation, naturalistic speaking avatars, and second language pronunciation learning [21, 29, 34, 50]. Modeling is also necessary in healthcare applications such as Brain-Computer Interfaces for communication [5] and diagnosing and treating speech disfluencies [32, 40].

Methods of external motion capture cannot record precise and accurate vocal tract movements for occluded articulators. Hence, the inner mouth is often poorly lit or neglected in multimedia approaches to motion capture-based facial animation [33].

Popular approaches to solving the issue of inner mouth occlusion include electromagnetic articulography (EMA) and electromyography (EMG) as models for the vocal tract. However, these methods only contain a small subset of articulatory features.

A more comprehensive approach uses Real-Time Magnetic Resonance Imaging (RT-MRI) of the vocal tract. This technology offers audio-aligned videos of internal and external articulators that are not measurable by other articulatory representations. When tested against downstream speech-related tasks, RT-MRI has been shown to more reliably and completely model the vocal tract in comparison to EMA [59]. However, current state-of-the-art labeling methods for extracting interpretable features from these videos are time-consuming, computationally expensive, and prone to errors [8]. Therefore, only a small amount of vocal tract RT-MRI data is labeled [35].

In this paper, we propose a comprehensive application of pretraining with both video and audio modalities for modeling the vocal tract. We first present a high-performance vision-based and multimodal RT-MRI feature extraction approach. Using these results, we label the Speech MRI Open Dataset [30] containing over 20 hours of vocal tract RT-MRI data for 75 speakers diverse in age, gender, and accent. To our knowledge, this dataset increases the amount of labeled public RT-MRI data of the vocal tract by over a factor of **9**.

Using this newly labeled dataset, we deploy an MRI-EMA pretraining method to further evaluate our feature extraction models using MRI-based deep speech synthesis. We achieve significant improvements in intelligibility compared to synthesis using the ground truth MRI tracks.

Another downstream application of our multimodal pretraining is 2D and 3D facial avatar visualization. We propose a deep articulatory inversion technique for speech-to-MRI prediction with a direct mapping to the 3D avatar, enabling us to visualize complex naturalistic speech production. We employ this method in both offline and real-time scenarios, achieving an average streaming latency of 107ms/batch of 100ms audio data. This result is a **20%** improvement over the previous baseline despite using a 3D

model with more than 6× the vertices.

To briefly summarize, our contributions in this work are: (1) a labeled version of the 75-speaker Speech MRI Open Dataset, (2) the application of a multimodal pretraining method for RT-MRI analysis, and (3) low latency facial avatar visualizations of MRI-based vocal tracts in offline and real-time contexts. These architectures are outlined in Figure 1.

2. Related Work

2.1. MRI Feature Extraction

Most published work using vocal tract RT-MRI feature extractions either do not use high-dimensional interpretable vocal tract representations, or only use previously extracted features for downstream tasks without generalizing well to unseen speakers [22, 32, 59, 66]. The existing algorithm for speaker segmentation traces airway contours using hand-drawn reference boundaries. It extracts 170 points of MRI coordinates per frame, taking up to 20 minutes to converge for a single frame [35]. From this point forward, we refer to the outputs of this algorithm as the "ground truth." More recent work in [2] uses a deep attention-gated U-Net for multi-speaker deep RT-MRI feature extraction, which is more efficient and generalizes well to unseen speakers.

2.2. Deep MRI-Based Speech Synthesis

Our concurrent submission attached in supplemental materials trains single-speaker MRI to speech. We detail our approach in Section 5.2. In this work, we explore new MRI-to-speech methodologies using the newly labelled 75-speaker dataset.

2.3. Speech-Driven Avatar Animation

Within the domain of automated facial animation, there have been many different approaches to driving an avatar from speech. Linguistic methods aim to map phonemes to visemes on the avatar [16, 44, 52, 63]. However, these systems require manually defined complex rules without a streamlined approach for inner mouth animation. Other approaches to speech-driven animation include deep learning models that are trained on audio-mesh paired data to perform mesh deformations [6, 15, 62]. However, these methods typically do not accurately model the tongue or suffer from oversmoothing when directly regressing to facial mesh movements [62].

Inner mouth animation from medical imaging representations of the vocal tract has been explored in [11], which provides real-time (21 Hz) 3D tongue animation using a streaming ultrasound snake contour extraction algorithm. Due to the tongue tip not being captured well in ultrasound, this tongue model is incomplete. Similarly, in [49], a kinematic tongue model extracted from a single MRI volumetric scan was directly animated using offline EMA data. These works demonstrate the benefits of being grounded within physiology but remain inaccessible to users without access to equipment for directly capturing the vocal tract.

Advances in deep articulatory inversion models have demonstrated the ability to approximate the physiology-grounded representations of the vocal tract from solely speech inputs [7, 9, 10, 20, 24–26, 28, 31, 37, 41, 43, 45, 47, 48, 51, 54, 55, 57, 58, 60, 61, 64, 65]. These approaches avoid potentially invasive recording equipment while retaining the valuable vocal tract information from speech production.

Our concurrent submission included in supplementary materials [3] builds on these works. We use WavLM [12] as feature extractors for input speech with multitask learning to predict tract variables, phonemes, pitch, and EMA simultaneously. In this work, we use a similar multi-task learning approach but with a pretraining RT-MRI and EMA inversion model on 75 RT-MRI speakers and 8 HPRC EMA speakers as described in Section 4.2 [53].

More recently, Medina *et al.* [33] combines the two methods of deep articulatory inversion models and vocal tract image representations. The result is a high-performance offline solution to facial animation with an emphasis on the inner mouth using EMA as an intermediate physiological representation. However, this approach has a low-dimensional tongue rig and loses physiological information when optimizing for the Metahuman FACS rig [17, 19].

Inspired by this speech-driven system, we developed a streaming system for speech-to-avatar synthesis as a concurrent paper submission (included in supplementary materials) [1]. This work builds on the approach from our previous submission in the following ways: (1) introducing RT-MRI as a replacement intermediary feature for EMA to increase inner mouth mesh resolution, (2) grounding the 3D facial model within human physiology according to RT-MRI of the vocal tract, and (3) optimizing streaming using a new system in Unreal Engine as opposed to Autodesk Maya.

3. Datasets

3.1. USC-TIMIT Dataset

We use the labeled 8-speaker RT-MRI USC-TIMIT dataset of the vocal tract for training and as the ground truth feature extractions for all baseline and experimental approaches described in [35]. Given sentences on a projection screen, subjects were instructed to read each out at a natural speaking rate while laying supine in an MRI scanner. A four-channel upper airway receiver coil array was used for recepting signals, which were processed to reproduce 84×84 midsaggital MR videos capturing lingual, labial, and jaw motion,

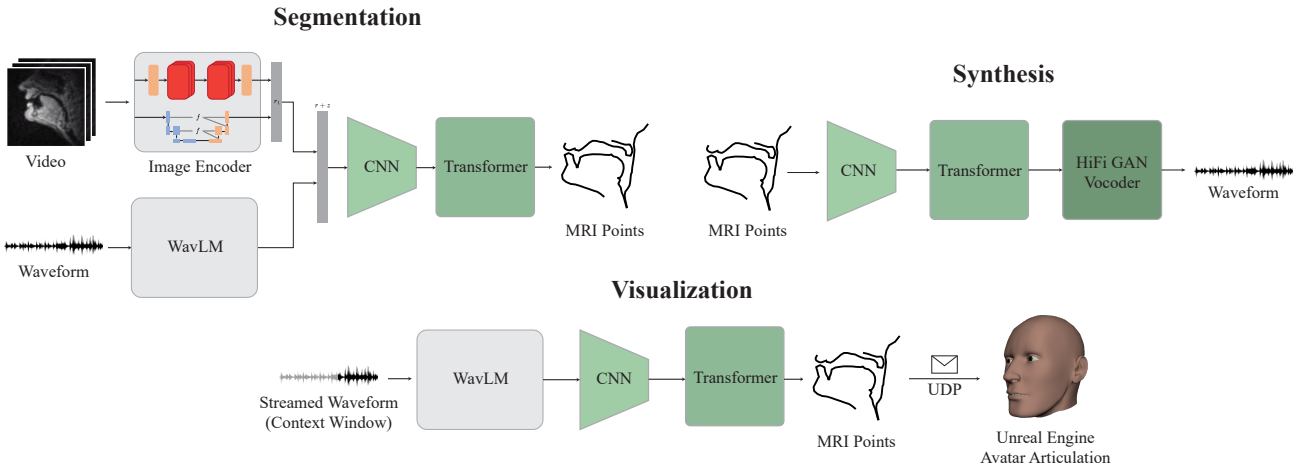

Figure 1. Summary of our approach showing three architectures for MRI analysis in segmentation, synthesis, and 3D visualization.

and velum, pharynx, and larynx articulations. These videos are collected at 83.33 Hz, and the algorithm described in Section 2.1 is used to extract the 170 representative points of the vocal tract. Of these 170 points, we take the subset of 95 points (190 x and y coordinates) that has been determined to be most vital for speech tasks in [59].

Paired with these trajectories is the 16kHz speech data (resampled from original 20kHz) corresponding to the read sentence during any RT-MRI scan. We further enhanced this audio using Adobe Podcast to reduce reverbation, as done in [59].

3.2. Speech MRI Open Dataset

The Speech MRI Open Dataset [30] is a multispeaker dataset from USC that provides synchronized speech of 75 diverse speakers with raw multi-coil RT-MRI videos of the vocal tract during articulation. Such a large, rich dataset can help solve many open problems in fields related to phonetics, spoken language, and vocal articulation. However, unlike the USC-TIMIT dataset, the data does not include labeled MRI feature points tracked over time, except for 6 speakers. This labeled data is not yet available to the public.

4. Multimodal Feature Extraction

4.1. Image-Based Feature Extraction

We follow [2], a U-Net style model [42] with an attention gating mechanism [36], for a baseline in MRI feature extraction using the frames alone as the input modality. We learn a spatial weighting map for each predicted point trained using Kullback-Leibler (KL) divergence loss between the weighting map and a 2D Gaussian heatmap.

This model is trained on 66 minutes of ground truth RT-MRI data from 7 of the 8 USC-TIMIT speakers. Given an MRI image, the U-Net predicts 95 84×84 spatial weighting maps for each of the 95 MRI points, and converts to points using a weighted average of the 25 highest pixels.

Our concurrent submission's [2] U-Net (in supplementary materials) is being used in our work as a baseline labeling model that informs future models.

An additional challenge with the ground truth data is the presence of jitter, or random high-frequency perturbations across consecutive frames, leading to noisy point tracking. While the trained U-Net outputs smoother point tracks, we additionally apply a temporal Gaussian low-pass filter independently for each point. This results in far smoother tracks, while maintaining accurate point detection per frame.

Using the U-Net model as a pretrained convolutional input, we further explore joint point tracking using a convolutional LSTM as in [22] (CLSTM) and a Transformer. The CLSTM, previously used in MRI video segmentation [66], applies a 2-layer LSTM to the predicted U-Net outputs, trained on speech from the same 7 USC-TIMIT speakers. The Transformer similarly uses the U-Net points from each timestep, with an additional positional encoding. Traditionally, multi-frame point tracking is done using optical flow [14] or by extension, Kalman filtering [13]. Recent joint point tracking results have found that explicitly appending optical flow to the Transformer input has resulted in better resulting tracks [46]. Thus, we further input predicted single frame optical flow, averaged over articulators using the Lucas-Kanade assumption that points in close proximity have similar flow in subsequent timesteps. Both the CRNN and the Transformer methods did not achieve equal or better performance than smoothed U-Net tracks on MRI videos of unseen speakers, reinforcing the fact that articulatory MRI tracking is fundamentally different than other

traditional video tracking problems. To address this, we explore additional modalities in the following sections.

4.2. Speech-Based Feature Extraction

For a secondary unimodal baseline in MRI feature extraction, we use the speech audio waveforms corresponding to the USC-TIMIT MRI trajectories as the input modality. Using the 10th layer of WavLM, we derive speech representations from the audio as the input to a Transformer prepended with three residual convolutional blocks.

Additionally, the Transformer model is trained on the speech data from 7 of the 8 USC-TIMIT speakers for multitask learning and outputs MRI trajectories and pitch from the speech representations simultaneously.

4.3. Multimodal Feature Extraction

We experiment with multiple multimodal models for feature extraction, using representations from video frames and from speech waveforms. We concatenate the two representations as input to a Transformer. Following 4.2, we train each of the multimodal models on the same 7 of 8 USC-TIMIT speakers with weighted L1 loss on outputted MRI and pitch.

4.4. Labeling Speech MRI Open Dataset

We deploy the previously described U-Net model trained on data from 7 USC-TIMIT speakers and 5 of the 6 labeled Speech MRI Open Dataset speakers, with temporal Gaussian low-pass filtering, to fully label video and audio aligned MRI point trajectories for the entire Speech MRI Open Dataset.

In Figure 2, we highlight the efficacy and generalizing qualities of the U-Net model on unseen speakers, allowing us to expand the amount of labeled MRI to over 20 hours across 83 total speakers. Qualitatively, the predicted segmentations closely follow the ground truth to trace the desired MRI segments, achieving a high quality labeling for unseen speakers. Further results are discussed in Section 7. As part of this paper, we also present this labeling for use in future downstream speech tasks, increasing the amount of labeled articulatory RT-MRI data available by over a factor of 9.

5. Deep MRI-Based Speech Synthesis

Another noteworthy challenge using labeled MRI data is speech synthesis from the MRI articulatory space. This is important for synchronized between avatar and speech rendering. Due to the higher resolution of MRI compared to the six points of EMA data used in prior works, a more detailed study of the vocal tract during speech production can be conducted. We explore single-speaker deep speech synthesis using the newly-labeled Speech MRI Open Dataset as an additional evaluation metric for potential future applications of this data.

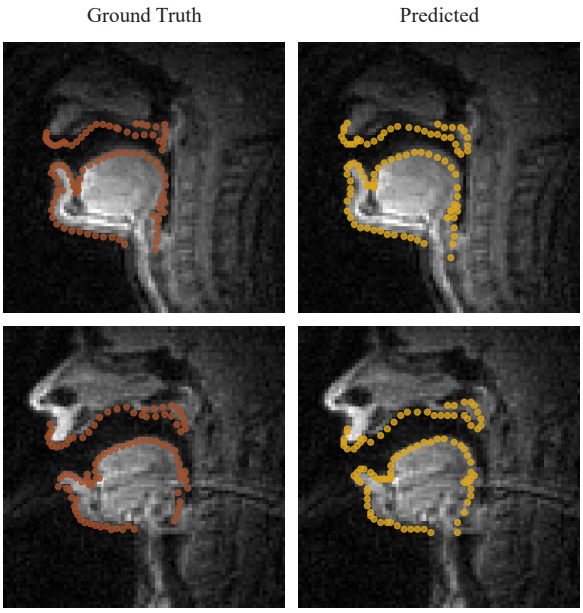

Figure 2. Two representative examples of predicted MRI points (*right*) compared to ground truth (*left*). Unseen examples spoken by unseen speakers from USC-TIMIT (*top, Female*) and Speech MRI Open Dataset (*bottom, Male*).

5.1. Models

For our main speech synthesis architecture, we build on [4] as illustrated in Figure 1. Given a set of 95 MRI point trajectories, we follow [59] to first concatenate and flatten the 95 x-y pairs into a 190-length vector and center the data around a point located on the hard palate with the lowest standard deviation.

With this preprocessed MRI, we first use 3 1D convolutional layers to encode the MRI points into an input for a 6-layer Transformer [18, 56]. The output of the Transformer is 256-channel HuBERT vectors [23] finetuned on VCTK with additional regularization loss. We then use a HiFi-GAN vocoder [27] to directly synthesize speech from these learned intermediate features. Full training details are provided in the appendix.

5.2. Experiments

For evaluation, we pretrain the Transformer on a version of the 75-speaker dataset labeled using a U-Net trained on 7 USC-TIMIT speakers. We combine the MRI dataset with HPRC EMA dataset [53]. Our concurrent submission attached in supplementary materials shows that pretraining with both MRI and EMA modality helps the model gener-

alize to unseen examples. However, concurrent submission only pretrained on single MRI speaker. With the 75-speaker dataset, we also finetune this multi-modal pretrained transformer on 75-speaker MRI only. Given these two types of pretrained weights, we finetune each of them on single speaker data based on the chosen segmentation of a USC-TIMIT speaker.

We choose the ground truth and the U-Net predicted MRI trajectories of USC-TIMIT as two baseline metrics to compare to the performance of the previously described multimodal segmentation models. Additionally, we independently finetune the model on two separate speakers: one seen speaker and one unseen speaker (by the labeling U-Net model). In this manner, we can assess the quality of a given segmentation model on the important downstream task of recovering intelligibility from the information-rich MRI segmentations. Full model ablation is included in the appendix.

6. Visualization

We further investigate the vocal tract during natural articulation using offline and real-time 3D visualizations of the face and mouth in combination with deep articulatory inversion. We aim to provide an anatomically-coherent visual representation of inferred MRI trajectories in speech production. Building on [38], our animated 3D model provides a faster, higher performance system for generating speech-driven avatar movements.

6.1. Deep Articulatory Inversion

For the proposed vocal animation architecture, the first step is to use an acoustic-to-articulatory inversion technique to predict MRI trajectories that will feed a custom 3D facial model. We use the newly labeled data of 75 speakers to learn to predict midsagittal x and y coordinates of the tongue (20 points), hard palate, velum (15 points), lips (ℓ points), lower incisor, and epiglottis.

We follow the same unimodal segmentation architecture from Section 4.2 as a speech-to-MRI inversion model. Combining the newly labeled 75-speaker MRI dataset with the HPRC dataset, we first pretrain the model with two heads, one outputting EMA, tract variables, phonemes, and pitch, and the other head outputting MRI and pitch. We then finetune the model on the 8-speaker TIMIT dataset with 1 speaker held out as test speaker.

6.2. Face and Vocal Tract Model

We construct a custom 3D face and mouth model in Autodesk Maya. We define point 88 of the 95-point MRI trajectory subset on the hard palate to be the origin of the 3D visualization space.

Relative to this point, we trace a tongue mesh according to a frame from the USC-TIMIT Napa speaker. For each tongue point in the MRI trajectories, we bind a joint to the tongue model in Maya. We also bind a tongue centroid joint for positional control of the 3D tongue and preservation of the overall topology of the mesh during deformation.

We follow a similar procedure for the hard palate, lips, and epiglottis to preserve movement trends and minimize noise. Due to noise from the inversion model being unpredictably amplified in high-resolution visualization of the velum and lips, we infer a low-joint mapping of these features from the original points from the MRI tracks. This approach retains the general trend of the vocal tract movements while disregarding the noise of individual MRI points trajectories. Finally, we model the movement of the lower incisor as a proxy for jaw movements through a hinge-based approximation following [1].

The full face model is illustrated and labeled with important features in Figure 4. Additionally, in Figure 5, we highlight a major benefit of using RT-MRI over EMA-based representations of past works: we can support higher granularity in tongue movements through 20 joints instead of 3.

6.3. Offline Animation

For animating the vocal tract when given a full utterance, we can offline the approach reported by [38] within Autodesk Maya as a qualitative baseline. The deep articulatory inversion model is deployed and provided the entire waveform as input. The MRI trajectory outputs from this model are then post-processed to correspond one-to-one with the rig joints of the facial model in Maya. For each timestep of the MRI trajectories, we transform each joint according to the MRI track and either set a keyframe or refresh the Viewport 2.0 to simulate animation. In the keyframe case, we can simply then play the animation from the start frame to the end frame at 83.33 Hz after all timesteps are keyed to achieve offline vocal animation.

However, since scripting in Maya is a blocking process, this method effectively freezes the program until animation is complete. A user is unable to interact with the 3D model and visually analyze speech production from multiple angles. Additionally, using a keyframe-based approach entails a delay before any animation is played while using a viewport-based approach has high refresh latency as reported by [38].

Alternatively, the real-time steps outlined in the following sections can be adapted to resolve the aforementioned issues. Deploying a network-based approach in an optimized game engine alleviates both problems of interactivity and latency.

6.4. Real-time Speech-to-MRI Processing

We build on and modify the general strategy proposed in [1] to accommodate streamed audio for vocal animation. The specific streaming system is briefly outlined below.

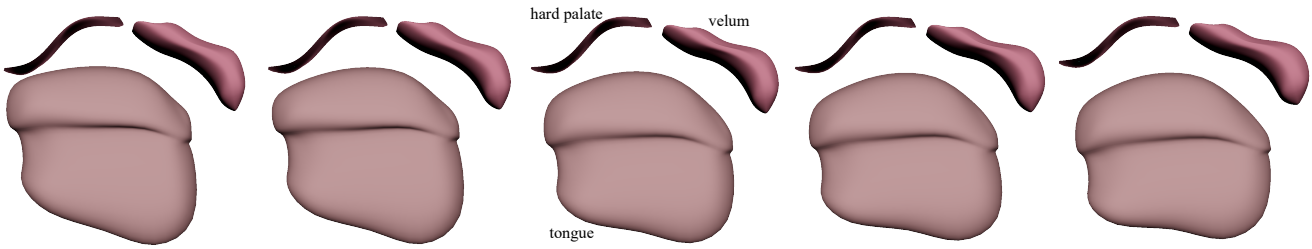

Figure 3. End of "buns"—Five frames of the tongue, hard palate, and velum generated for audio length of 60ms, corresponding to the fricative [s].

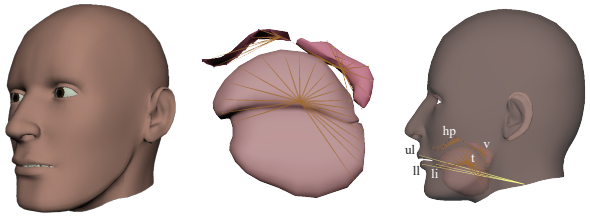

Figure 4. Three-quarter view of full 3D face model (left), midsagittal view of tongue, hard palate, and velum 3D models with joint-based rigs (middle), and midsagittal view of face, tongue, hard palate, and velum 3D models with joint-based rigs labeled (right). Key: ul - upper lip, ll - lower lip, li - lower incisor, hp - hard palate, v - velum, t - tongue.

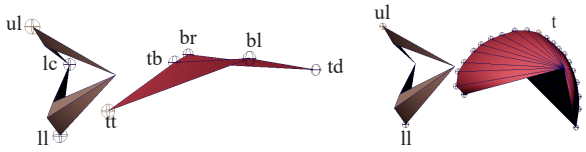

Figure 5. Comparing inner mouth model resolution of the 3D EMA-based model in [33] (*left*) and the proposed 3D MRI-based model (*right*). Tongue resolution is increased from 6 joints in [33] to 20 joints in proposed model.

Rather than receiving the full waveform as in the offline case, we instead use an audio input stream to collect batches of 1600 audio samples in sequential order from a WAV file or from a live microphone. Given that our inversion model predicts noisy MRI trajectories from silence, we employ Google's WebRTC Voice Activity Detector (VAD) to classify if a batch contains speech or not. If not, we do not move the facial model for 100ms as silence simulation.

If the VAD detects speech, we use a sliding context window of n seconds. This window consists of three parts: the *seen* batch, the *current* batch, and the *forward* batch. The current batch contains the 1600 samples that we have just received from the audio input stream. The forward batch is 1600 samples from a look-ahead window of 100 ms duration. This audio is "future audio" from an intentional initial 100ms delay, where we wait for two batches to be streamed in before starting speech processing. Using the forward batch, we add future context that helps the inversion model by providing coarticulation information from the next 100ms. Since the current and forward batches together have 3200 samples of speech data, we prepend the $16000n-3200$ samples of audio that we have just processed as past context in the seen batch. Together, the seen, current, and forward batches make an n-second long context window for the inversion model.

One other issue we encounter is the lack of context when streaming first begins. Since we only have access to a small number of 100ms audio batches as input, the inversion model is prone to output noise. To mitigate these initial fluctuations, we employ a source of artificial context—we prepend either a random utterance from the USC-TIMIT dataset, an articulated vowel, or silence to our streaming audio until our sliding window has a full n seconds of context to draw from.

This context window is translated to the corresponding MRI trajectories using deep articulatory inversion. Of the $83n$ frames returned by the model, we only retain the approximately 8 frames representing the current batch of audio.

6.5. Real-time MRI-to-Avatar

We import the fully rigged 3D facial model designed for vocal visualization into Unreal Engine to fully utilize its optimizations in real-time animation.

To dynamically transform the joints of our model without the use of a `Control Rig`, we define custom `Actor` and `AnimInstance` classes for each MRI segment. Using imported Maya rigs as `SkeletalMeshComponent`s, we can programmatically share our desired feature location with a corresponding `Animation Blueprint` and control the joint when the game engine is live. Additionally, each joint's orientation is set relative to the `World Space` to ensure we preserve covariances between MRI features for physiological accuracy.

To stream the processed MRI trajectory batches to Un-

real Engine, we implement a simple User Datagram Protocol (UDP) between two simultaneous processes. We first perform all speech-to-MRI preprocessing in an async-based Python process then pass information across a socket to the concurrent Unreal Engine process. However, since `Ticks` in Unreal Engine are constrained to 83.33 Hz to match the MRI frequency, packet loss will severely degrade the animation quality when the server send rate is high. To remedy this timing mismatch, we halve the server send rate to 42 Hz and utilize an asynchronous `TQueue` within Unreal Engine to collect every frame. We are able to perfectly match the rate of streaming audio in this manner.

With this implementation of real-time MRI-to-avatar in Unreal Engine, we can interact with a low-latency streaming vocal animation in a game-like 3D environment.

7. Results

Examining both the quantitative and qualitative parts of our MRI vocal tract analysis provides valuable insights into both the efficacy of our proposed methods and MRI as a representation for naturalistic speech production. We explore these topics in brief in this section.

7.1. Feature Extraction

When analyzing our various feature extraction methods, we first evaluate performance within the context of seen speakers but unseen examples.

Figure 6 highlights quantitative results in L1 losses and Pearson Correlation Coefficients (PCCs) when evaluating models on unseen examples from seen speakers. We observe a significant trend where multimodal models perform consistently better than the purely video-based U-Net. In fact, the best model in terms of both metrics includes the outputs of the U-Net as one of the input modalities alongside WavLM vectors. These results suggest the inclusion of speech within segmentation provides additional speaker-specific information related to the anatomy of the vocal tract. Since the shape of different parts of the vocal tract can greatly vary from speaker to speaker, this inclusion is crucial to a better in-domain modeling of speech production. With the image modality alone, the fully pixel value-based U-Net generalizes better to unseen speakers since contour pixel values have less dependence on the speaker compared to WavLM features in the speech modality.

For visualization of these results, we invite you to watch our demo video in supplementary materials.

7.2. Deep Speech Synthesis

Similarly, we evaluate our segmentation methods using the MRI-based speech synthesis downstream task within seen and unseen speaker contexts.

To summarize 5.2, the synthesis model is pretrained on the newly-labeled 75-speaker dataset. To then evaluate the performance of a given feature extraction model (e.g. U-Net, multimodal), we finetune this model on the predicted MRI trajectories of a USC-TIMIT speaker.

Model	Mean WER [↓] Seen Speaker	Unseen Speaker
U-Net + WavLM	**0.31** $\pm$ 0.36	**0.33** $\pm$ 0.26
U-Net	0.36 $\pm$ 0.33	0.35 $\pm$ 0.33
Ground Truth	0.34 $\pm$ 0.35	0.50 $\pm$ 0.27
U-Net + WavLM (S)	0.35 $\pm$ 0.33	0.50 $\pm$ 0.39

Table 1. USC-TIMIT speaker finetuning for seen and unseen speakers: Mean WER for speech synthesis pretrained on 75-speaker dataset. (S) denotes synthesis model pretrained using single MRI speaker. All other models are pretrained with 75-speaker MRI.

To evaluate the intelligibility of synthesized speech, we compute the word error rate (WER) on test unseen examples from the same training speaker using Whisper [39], a state-of-the-art automatic speech recognition (ASR) model. For seen speakers of segmentation models, the multimodal UNet-WavLM based synthesizer outperforms both the ground truth baseline as well as the U-Net, suggesting that the addition of the speech modality helps preserves more speech-related information within the predicted MRI point trajectories compared to a purely image-based approach. Table 1 summarizes these results.

Similarly, Table 1 highlights that the UNet-WavLM based model has the lowest WER when testing against an unseen USC-TIMIT speaker, compared to the ground truth segmentations and the U-Net. This demonstrates that the outputs from multi-modal on unseen speaker still capture representative articulatory kinematics for naturalistic speech.

7.3. Visualization

We examine the visualization results in both quantitative and qualitative manners. To evaluate streaming performance, we explore both the latency and the accuracy of the streaming system.

For our baseline, we use the 1-second sliding context window results from [38], where they achieve an average latency of 133ms per 100ms batch of streamed audio. We conduct similar profiling tests with the same 1-second sliding context window and find an average latency of **107ms** per 100ms batch of streamed audio, outperforming the baseline by a 20% margin while having 6$\times$ more vertices. We attribute this largely to the animation times being reduced from 56.3ms/batch when streaming in Maya to a constant 12ms/batch using the optimized game engine.

Additionally, we evaluate the streaming performance of

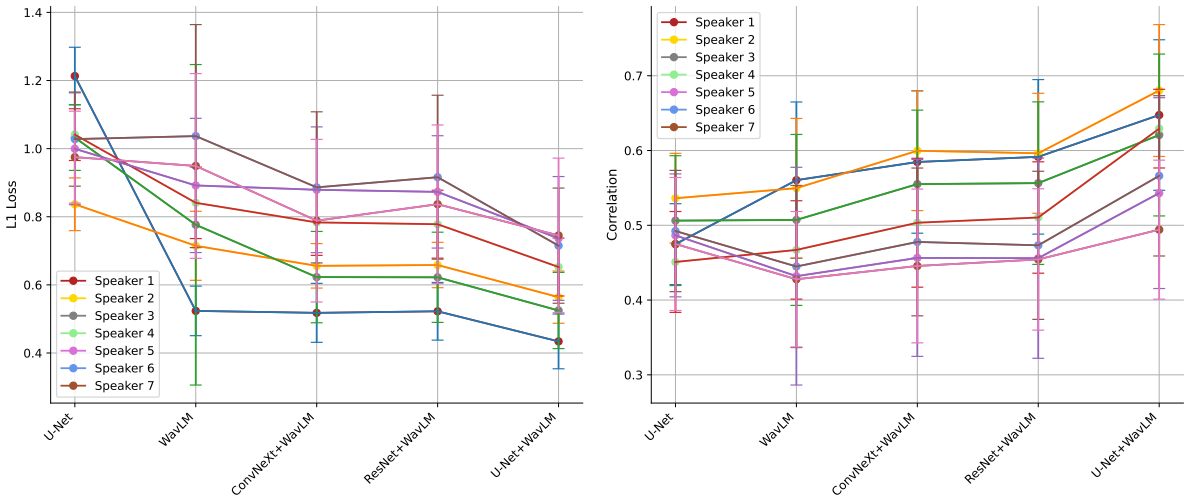

Figure 6. L1 losses [↓] (*left*) and Pearson Correlation Coefficients (PCCs) [↑] (*right*) comparing MRI trajectories of unseen examples from seen speakers of a given model with the USC-TIMIT ground truth. Varying through a subset of six representative models.

the inversion model for a quantitative metric on animation accuracy. When streaming an unseen example from a seen USC-TIMIT speaker, Figure 7 highlights the visual similarities between predicted MRI trajectories and predictions from UNet-WavLM. We observe high Pearson correlation coefficients for the displayed MRI features, which were chosen based on their importance in speech synthesis following [59].

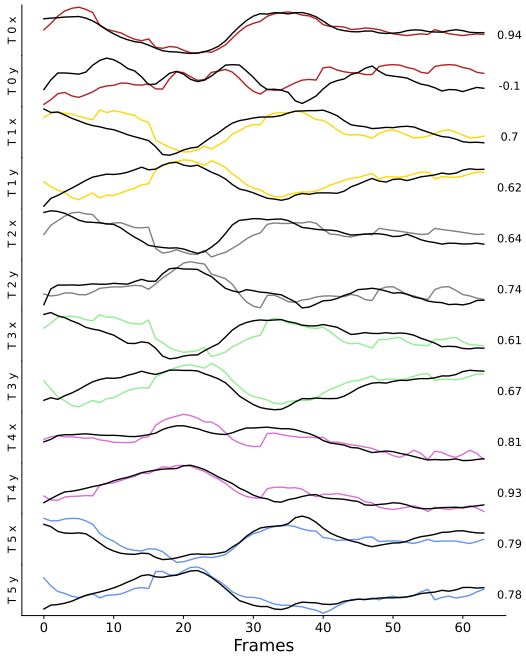

Figure 7. MRI trajectories inferred when streaming the inversion model are shown in color. The trace of the ground truth MRI data is shown in black. To the right of each plot, we present Pearson correlation coefficients (PCCs) comparing the predicted trajectories to the ground truth.

Qualitatively, we visually inspect animation results when streaming and compare these to the ground truth vocal tract movements from the MRI videos. We observe articulator movements matching the U-Net during speech production. For example, Figure 3 demonstrates accurate tongue movement corresponding to the end of the articulation of the word "buns", where the tip of the tongue first touches the hard palate during the consonant "n" then extends forward during the consonant "s" before receding. We summarize these results in further detail with additional examples in the supplementary video.

8. Conclusion

Pretraining with audio and image modalities using the newly labeled 75-speaker RT-MRI dataset establishes new MRI benchmarks in vocal tract feature extraction, deep speech synthesis, and real-time speech-driven avatars. While we achieve high quality seen speaker visualization using inversion, current models struggle to disentangle speaker-specific information from speech representations. Future work may use the labeled 75-speaker dataset for speaker-independent speech-driven facial animation.

References

[1] Anonymized. Towards streaming speech-to-avatar synthesis, 2023. 2, 5

[2] Anonymized. Deep articulatory mri feature extraction and speech synthesis, 2023. 2, 3

[3] Anonymized. Acoustic-to-articulatory inversion for multilingual and downstream tasks, 2023. 2

[4] Anonymized. Articulatory synthesis with multi-modal and self-supervised features, 2023. 4

[5] Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019. 1

[6] Monica Villanueva Aylagas, Hector Anadon Leon, Mattias Teye, and Konrad Tollmar. Voice2face: Audio-driven facial and tongue rig animations with cVAEs. *Computer Graphics Forum*, 41(8):255–265, 2022. 2

[7] Narjes Bozorg and Michael T Johnson. Acoustic-to-articulatory inversion with deep autoregressive articulatory-wavenet. *Networks (CNNs)*, 2020. 2

[8] Erik Bresch and Shrikanth Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging*, 28(3):323–338, 2009. 1

[9] Zexin Cai et al. The dku-jnu-ema electromagnetic articulography database on mandarin and chinese dialects with tandem feature based acoustic-to-articulatory inversion. In *ISCSLP*, 2018. 2

[10] Claudia Canevari et al. A new italian dataset of parallel acoustic and articulatory data. In *Interspeech*, 2015. 2

[11] Shicheng Chen, Yifeng Zheng, Chengrui Wu, Guorui Sheng, Pierre Roussel, and Bruce Denby. Direct, near real time animation of a 3d tongue model using non-invasive ultrasound images. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4994–4998, 2018. 2

[12] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. 2

[13] Yaran Chen, Dongbin Zhao, and Haoran Li. Deep kalman filter with optical flow for multiple object tracking. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3036–3041, 2019. 3

[14] TM Chin, WC Karl, and AS Willsky. Probabilistic and sequential computation of optical flow using temporal coherence. IEEE Trans Image Process, 1994. 3

[15] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 2

[16] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics*, 35 (4):1–11, 2016. 2

[17] Paul Ekman and Wallace V. Friesen. Facial action coding system, 1978. 2

[18] David Gaddy and Dan Klein. An improved model for voicing silent speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 175–181, Online, 2021. Association for Computational Linguistics. 4

[19] Epic Games. https://www.unrealengine.com/en-US/metahuman-creator. 2

[20] Prasanta Kumar Ghosh et al. A generalized smoothness criterion for acoustic-to-articulatory inversion. *JASA*, 2010. 2

[21] Bryan Gick, Barbara May Bernhardt, Penelope Bacsfalvi, and Ian Wilson. 11. ultrasound imaging applications in second language acquisition. 2008. 1

[22] S Ashwin Hebbar, Rahul Sharma, Krishna Somandepalli, Asterios Toutios, and Shrikanth Narayanan. Vocal tract articulatory contour detection in real-time magnetic resonance images using spatio-temporal context. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7354–7358, 2020. 2, 3

[23] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, 2021. 4

[24] Thomas Hueber et al. Speaker adaptation of an acoustic-to-articulatory inversion model using cascaded gaussian mixture regressions. In *Interspeech*, 2013. 2

[25] Aravind Illa et al. The impact of cross language on acoustic-to-articulatory inversion and its influence on articulatory speech synthesis. In *ICASSP*, 2022.

[26] An Ji. *Speaker independent acoustic-to-articulatory inversion*. PhD thesis, Marquette University, 2014. 2

[27] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, pages 17022–17033. Curran Associates, Inc., 2020. 4

[28] Paul K. Krug et al. Self-Supervised Solution to the Control Problem of Articulatory Synthesis. In *Interspeech*, 2023. 2

[29] June S. Levitt and William F. Katz. The effects of EMA-based augmented visual feedback on the English speakers' acquisition of the Japanese flap: a perceptual study. In *Proc. Interspeech 2010*, pages 1862–1865, 2010. 1

[30] Yongwan Lim, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Goud Lingala, Colin Vaz, Tanner Sorensen, Miran Oh, Sarah Harper, Weiyi Chen, Yoonjeong Lee, Johannes Töger, Mairym Lloréns Monteserin, Caitlin Smith, Bianca Godinez, Louis Goldstein, Dani Byrd, Krishna S. Nayak, and Shrikanth S. Narayanan. A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images. *Scientific Data*, 8(1), 2021. 1, 3

[31] Peng Liu et al. A deep recurrent approach for acoustic-to-articulatory inversion. In *ICASSP*, 2015. 2

[32] Yijing Lu, Charlotte E.E. Wiltshire, Kate E. Watkins, Mark Chiew, and Louis Goldstein. Characteristics of articulatory

gestures in stuttered speech: A case study using real-time magnetic resonance imaging. *Journal of Communication Disorders*, 97, 2022. 1, 2

[33] Salvador Medina, Denis Tome, Carsten Stoll, Mark Tiede, Kevin Munhall, Alex Hauptmann, and Iain Matthews. Speech driven tongue animation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 1, 2, 6

[34] Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023. 1

[35] Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, Dani Byrd, Erik Bresch, Prasanta Ghosh, Athanasios Katsamanis, and Michael Proctor. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *The Journal of the Acoustical Society of America*, 136(3):1307–1311, 2014. 1, 2

[36] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018. 3

[37] Slim Ouni et al. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *JASA*, 2005. 2

[38] Tejas S. Prabhune, Peter Wu, Bohan Yu, and Gopala K. Anumanchipalli. Towards streaming speech-to-avatar synthesis, 2023. 5, 7

[39] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 7

[40] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando de la Torre, and Yaser Sheikh. Audio- and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 41–50, 2021. 1

[41] Korin Richmond. *Estimating articulatory parameters from the acoustic speech signal*. PhD thesis, University of Edinburgh, 2002. 2

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[43] Kevin Scheck and Tanja Schultz. STE-GAN: Speech-to-Electromyography Signal Conversion using Generative Adversarial Networks. In *Interspeech*, 2023. 2

[44] Endang Setyati, Surya Sumpeno, Mauridhi Hery Purnomo, Koji Mikami, Masanori Kakimoto, and Kunio Kondo. Phoneme-viseme mapping for indonesian language based on blend shape animation. In *IAENG International Journal of Computer Science*, 2015. 2

[45] Abdolreza Sabzi Shahrebabaki et al. Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models. *TASLP*, 2021. 2

[46] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation, 2023. 3

[47] Hayato Shibata et al. Unsupervised acoustic-to-articulatory inversion neural network learning based on deterministic policy gradient. In *SLT*, 2021. 2

[48] Yashish M Siriwardena et al. The secret source: Incorporating source features to improve acoustic-to-articulatory speech inversion. In *ICASSP*, 2023. 2

[49] Ingmar Steiner and Slim Ouni. Progress in animation of an ema-controlled tongue model for acoustic-visual speech synthesis, 2012. 2

[50] Atsuo Suemitsu and Jianwu Dang. A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning. *The Journal of the Acoustical Society of America*, 2015. 1

[51] Guolun Sun et al. Temporal convolution network based joint optimization of acoustic-to-articulatory inversion. *Applied Sciences*, 2021. 2

[52] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 275–284, Goslar, DEU, 2012. Eurographics Association. 2

[53] Mark Kenneth Tiede et al. Quantifying kinematic aspects of reduction in a contrasting rate production task. *JASA*, 2017. 2, 4

[54] Tomoki Toda et al. Acoustic-to-articulatory inversion mapping with gaussian mixture model. In *ICSLP*, 2004. 2

[55] Asterios Toutios and Konstantinos Margaritis. A rough guide to the acoustic-to-articulatory inversion of speech. In *HERCMA*, 2003. 2

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 4

[57] Jianrong Wang et al. Acoustic-to-articulatory inversion based on speech decomposition and auxiliary feature. In *ICASSP*, 2022. 2

[58] Martijn Wieling et al. Analysis of acoustic-to-articulatory speech inversion across different accents and languages. In *Interspeech*, 2017. 2

[59] Peter Wu, Tingle Li, Yijing Lu, Yubin Zhang, Jiachen Lian, Alan W Black, Louis Goldstein, Shinji Watanabe, and Gopala K. Anumanchipalli. Deep speech synthesis from mri-based articulatory representations. In *INTERSPEECH 2023*. ISCA, 2023. 1, 2, 3, 4, 8

[60] Peter Wu et al. Speaker-independent acoustic-to-articulatory speech inversion. In *ICASSP*, 2023. 2

[61] Xurong Xie et al. Deep neural network based acoustic-to-articulatory inversion using phone sequence information. In *Interspeech*, 2016. 2

[62] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. *arXiv preprint arXiv:2301.02379*, 2023. 2

[63] Yuyu Xu, Andrew W. Feng, Stacy Marsella, and Ari Shapiro. A practical and configurable lip sync method for games. In *Proceedings of Motion on Games*, page 131–140, New York, NY, USA, 2013. Association for Computing Machinery. 2

[64] Tianfang Yan et al. Combining language corpora in a Japanese electromagnetic articulography database for acoustic-to-articulatory inversion. In *Interspeech*, 2023. 2

[65] Atef Ben Youssef et al. Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden markov models. In *Interspeech*, 2009. 2

[66] Yide Yu, Amin Honarmandi Shandiz, and László Tóth. Reconstructing speech from real-time articulatory mri using neural vocoders, 2021. 2, 3