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,

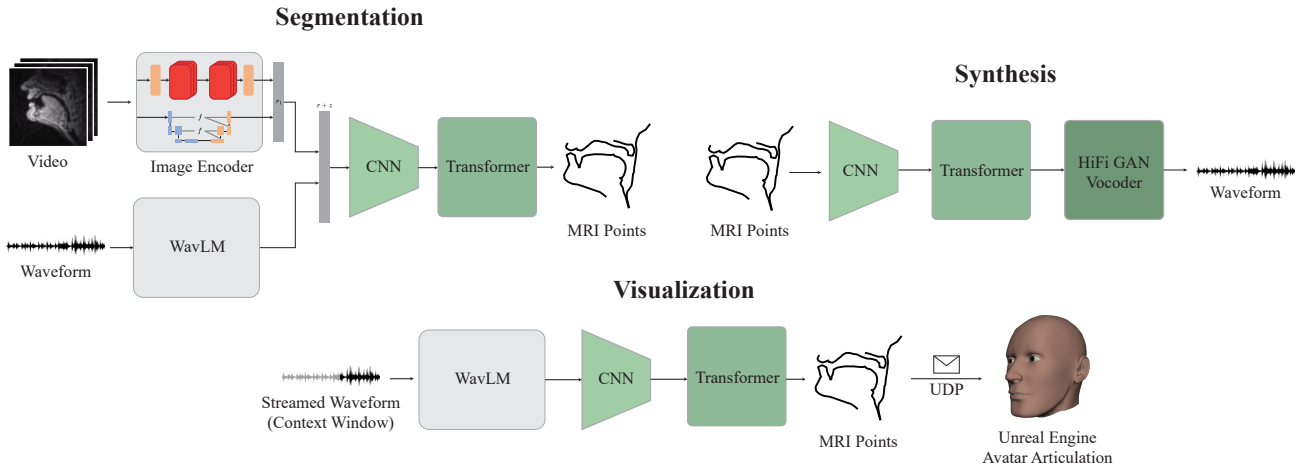

Figure 1. Summary of our approach showing three architectures for MRI analysis in segmentation, synthesis, and 3D visualization.

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.

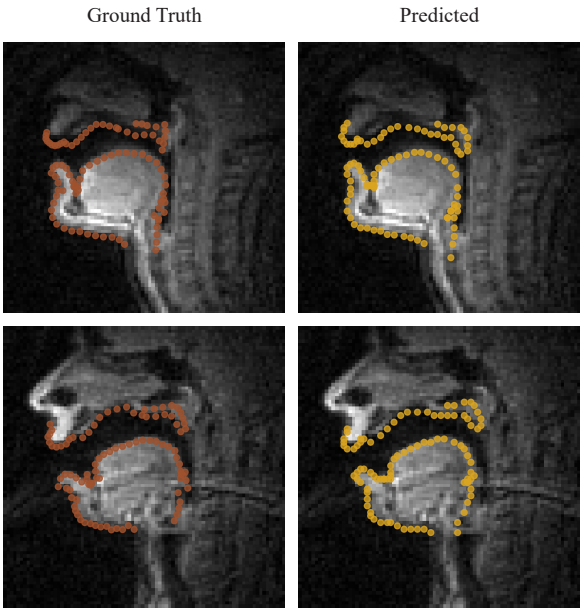

Figure 2. Two representative examples of predicted MRI points (*right*) compared to ground truth (*left*). Unseen examples spoken by unseen speakers from USC-TIMIT (*top, Female*) and Speech MRI Open Dataset (*bottom, Male*).

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.

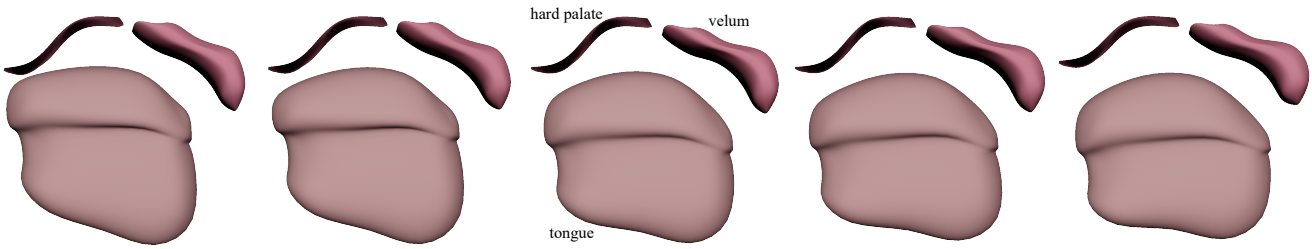

Figure 3. End of "buns"—Five frames of the tongue, hard palate, and velum generated for audio length of 60ms, corresponding to the fricative [s].

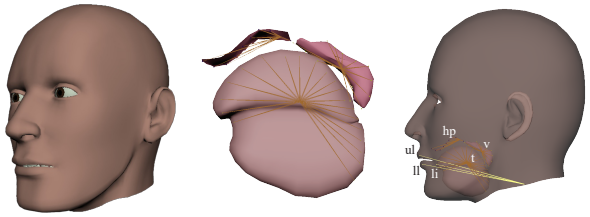

Figure 4. Three-quarter view of full 3D face model (left), midsagittal view of tongue, hard palate, and velum 3D models with joint-based rigs (middle), and midsagittal view of face, tongue, hard palate, and velum 3D models with joint-based rigs labeled (right). Key: ul - upper lip, ll - lower lip, li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

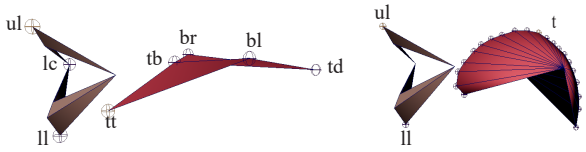

Figure 5. Comparing inner mouth model resolution of the 3D EMA-based model in [33] (*left*) and the proposed 3D MRI-based model (*right*). Tongue resolution is increased from 6 joints in [33] to 20 joints in proposed model.

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

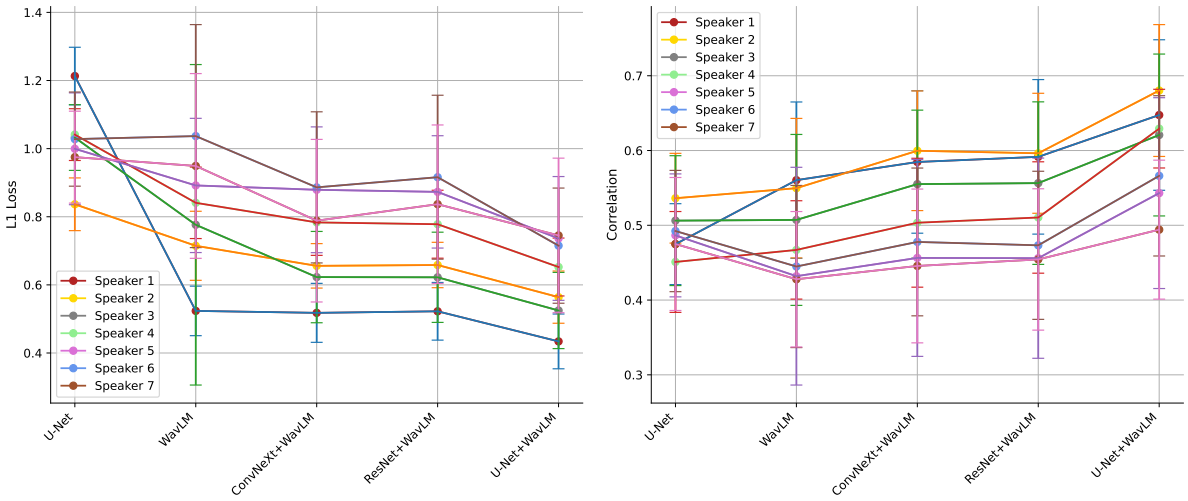

Figure 6. L1 losses [↓] (*left*) and Pearson Correlation Coefficients (PCCs) [↑] (*right*) comparing MRI trajectories of unseen examples from seen speakers of a given model with the USC-TIMIT ground truth. Varying through a subset of six representative models.

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].

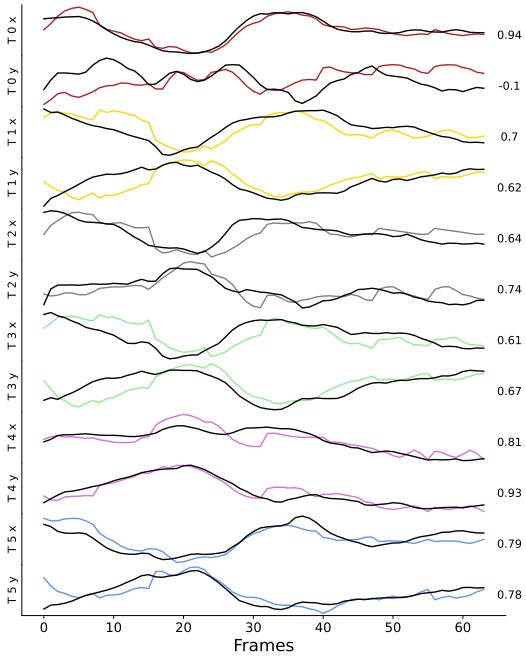

Figure 7. MRI trajectories inferred when streaming the inversion model are shown in color. The trace of the ground truth MRI data is shown in black. To the right of each plot, we present Pearson correlation coefficients (PCCs) comparing the predicted trajectories to the ground truth.

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

Multimodal Pretraining for Vocal Tract Modeling

Supplementary Material

Project Page with Video: https://speech-avatar.github.io/multimodal-mri-avatar/

1. Supplementary Video

We invite readers to view our supplementary video presented on our anonymous project page at https://speech-avatar.github.io/multimodal-mri-avatar/.

We offer examples and comparisons of our 2D and 3D avatars with the ground truth MRI data for both seen and unseen speakers in offline and real-time scenarios. Further qualitative analysis of speech-to-avatar output is conducted.

2. Appendix for Deep MRI-based Speech Synthesis

2.1. 256-channel HuBERT vectors

We follow the concurrent submission attached below to train our own 256-channel HuBERT vectors. Using the output from the 6th layer self-attention module in HuBERT, we use a linear-projection layer to project the 1024-dimensional hidden vector into 256 dimensions. During training, we linearly project the 256-dimensional features back to 1024 dimensions and compute an MSE loss between the resulting features and the ground truth ones. We freeze HuBERT during training and inference and only train the two linear layers. During inference, we only use the trained 256-dimensional features and discard the second linear projection from 256 channels to 1024 channels. We trained the model using the VCTK dataset and held out 10% of speakers as test speakers. For the remaining 90% speakers, we used a 90-10 train-validation split to train and validate our model. We used Adamax as the optimizer with beta values (0.99, 0.999). The learning rate starts at 1e-4 and is regulated by ReduceLROnPlateau with default settings.

2.2. MRI-to-256

In addition to the transformer model in Section 5.1, we adapted the discriminator from HifiGAN to discriminate between synthesized 256-channel HuBERT vectors and the ground truth. We train the transformer generator and the HifiGAN-discriminator in the same GAN setting as HifiGAN. We used Adam optimizer with beta values (0.5, 0.9) for both the generator and the discriminator. During inference, we discard the discriminator and only retain the transformer generator.

We breakdown the training into two steps: pretraining and finetuning. We tested three pretraining steps:

1. Pretrain the model using 75-speaker MRI dataset and 8-speaker HPRC EMA dataset. Within a batch, we append 12 channels of zeros for MRI samples, and we prepend 190 channels of zeros for EMA samples. This allows the 1D-convolution layer afterwards to distinguish between EMA and MRI samples.
2. Do step 1 and then finetune on 75-speaker MRI only.
3. Pretrain on single-speaker MRI and 8-speaker HPRC, using the same approach as in step 1.

We name the three pretraining approaches "Pretrain MRI-EMA", "Pretrain MRI", and "Pretrain Multi" respectively. During pretraining, we started with a learning rate of 1e-4 and halved the learning rate for every 20000 steps.

After pretraining, we finetune the model on single-speaker MRI train-data only, using a 85-5-10 train-validation-test split. The pretraining step will not use any validation or test data from the single-speaker MRI.

In both Tables 2 and 3, a speech-synthesis model trained on the multi-modal U-Net + WavLM segmentation output always outperforms models trained on the U-Net segmentation output, regardless of the pretraining approach used. In Table 1, we report synthesis models with the lowest WER obtained from the pretraining methods.

Model	Mean CER	Mean WER
U-Net + WavLM (Pretrain MRI-EMA)	0.252 ± 0.254	0.333 ± 0.263
U-Net (Pretrain MRI)	0.253 ± 0.355	0.352 ± 0.333
U-Net + WavLM (Pretrain MRI)	0.326 ± 0.264	0.456 ± 0.292
U-Net WavLM (Pretrain Multi)	0.246 ± 0.193	0.501 ± 0.391
U-Net (Pretrain MRI-EMA)	0.302 ± 0.219	0.531 ± 0.417
U-Net (Pretrain Multi)	0.381 ± 0.264	0.588 ± 0.342

Table 2. Model ablations for unseen examples from unseen speaker

Model	Mean CER	Mean WER
U-Net + WavLM (Pretrain MRI-EMA)	0.238 ± 0.328	0.313 ± 0.356
U-Net + WavLM (Pretrain Multi)	0.302 ± 0.401	0.349 ± 0.332
U-Net (Pretrain MRI-EMA)	0.248 ± 0.315	0.357 ± 0.325
U-Net (Pretrain MRI)	0.281 ± 0.306	0.364 ± 0.334
U-Net + WavLM (Pretrain MRI)	0.333 ± 0.423	0.375 ± 0.373
U-Net (Pretrain Multi)	0.390 ± 0.383	0.476 ± 0.379

Table 3. Model ablations for unseen examples from seen speaker

TOWARDS STREAMING SPEECH-TO-AVATAR SYNTHESIS

Anonymous

Anonymous Institution

ABSTRACT

Streaming speech-to-avatar synthesis creates real-time animations for a virtual character from audio data. Accurate avatar representations of speech are important for the visualization of sound in linguistics, phonetics, and phonology, visual feedback to assist second language acquisition, and virtual embodiment for paralyzed patients. Previous works have highlighted the capability of deep articulatory inversion to perform high-quality avatar animation using electromagnetic articulography (EMA) features. However, these models focus on offline avatar synthesis with recordings rather than real-time audio, which is necessary for live avatar visualization or embodiment. To address this issue, we propose a method using articulatory inversion for streaming high quality facial and inner-mouth avatar animation from real-time audio. Our approach achieves 130ms average streaming latency for every 0.1 seconds of audio with a 0.792 correlation with ground truth articulations. Finally, we show generated mouth and tongue animations to demonstrate the efficacy of our methodology.

***Index Terms*—** articulatory inversion, streaming, speech-to-avatar

1. INTRODUCTION

Speech-driven avatar animation is useful for many applications in speech and linguistics. Specifically, it can facilitate second language (L2) pronunciation learning via visual feedback [1, 2, 3] and aid hearing-impaired individuals to lip-read when only an audio signal is available during communication. In addition, accurate facial and tongue animation has been shown to help with virtual embodiment for paralyzed patients [4]. Solutions to the speech-driven avatar task date back to [5, 6], which proposed predicting phonemes using a combination of Qualisys optical motion tracking and electromagnetic articulography (EMA) data.

Key tasks within the development of automated avatars include both real-time and offline speech-driven animation of the face and inner mouth, as needed in interactive systems or multimedia like video games [7]. Previous works focusing on offline synthesis have achieved success using face scans [8, 9] as well as using various input modalities like MRI [10] and/or EMA [11] to model the movement of articulators. More recently, deep articulatory inversion techniques have shown to produce high-quality speech-to-EMA models and subsequent avatar animations with the additional help of a 3D rig optimizer model [12].

However, these advances in offline animation methodologies have not been extended to streaming solutions yet. Current real-time speech-driven facial animations have employed deep neural networks (DNNs) to prove an acoustic-visual mapping is possible [13], but do not offer avatar visualizations via a 3D facial mesh or latency analysis for streaming purposes.

In this work, we aim to connect the recent advances in deep articulatory inversion to improving real-time speech-driven facial and tongue animation. We propose a low latency streaming synthesis approach to predict batches of EMA data from speech using acoustic-to-articulatory inversion, animate a joint-based 3D model in Autodesk Maya, and evaluate predicted animations by comparing the generated motion capture to ground truth labels. Specifically, we achieve average latencies of 130ms per 0.1 seconds of streamed audio using shared memory buffers and demonstrate average EMA correlations of 0.792 during real-time prediction.

2. METHODS

For the proposed articulatory streaming architecture, we use an acoustic-to-articulatory inversion process (AAI) followed by a mapping between each EMA feature and a corresponding joint or curve on the 3D face model.

2.1. Acoustic-to-Articulatory Inversion

We first utilize articulatory inversion to generate corresponding EMA data. This data consists of 12-dimensional features, which provide the midsagittal x and y coordinates of the tongue tip, body, and dorsum, the upper and lower lips, and the lower incisor.

We use two models for inversion—BiGRU-based and Transformer-based. We first follow [14] which uses a BiGRU architecture with chunked autoregression and adversarial training. Specifically, we use the model with an MLP to help with coarticulation, a CNN as the discriminator for realistic outputs, and the final layer of HuBERT [15] to map speech into a compressed yet generalizable representation.

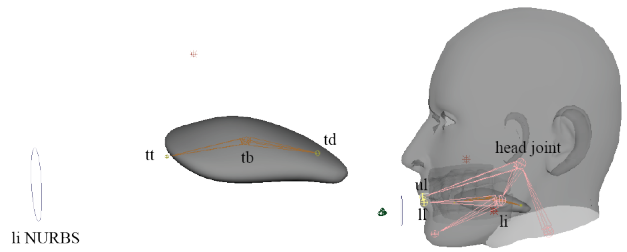

Fig. 1: Midsagittal view of the 3D face avatar used for articulatory streaming; tongue and NURBS curve on the left are zoomed in from full face model on the right.

Additionally, we use a state-of-the-art six-layer Transformer model prepended with three residual convolutional blocks following [16]. The model uses the tenth layer of WavLM for speech representations of audio inputs and outputs EMA, tract variables (TVs), phonemes, and pitch simultaneously [17].

The outputs we use from the inversion include EMA feature position (x, y) data independently normalized to a $[-1, 1]$ range. Using the M02 speaker from the Haskins Production Rate Comparison (HPRC) dataset's minimum and maximum feature values, we denormalize the predicted EMA data into a 2D space where the covariances between features are preserved.

2.2. 3D Face Model

The facial model in Figure 1 was constructed in Autodesk Maya using teeth from [18], the face from [19], and a custom tongue. A joint-based rig was made for each of the tongue tip, tongue body, and tongue dorsum EMA features.

For the head, we built a similar rig with appropriate skin weights to map areas of the face to the corresponding joint. To approximate the hinge-based movements of the lower incisor, we constrain the rotation of the lower incisor joint to the y-translation of a NURBS curve as controlled by the li EMA feature. The remaining upper and lower lip features were rigged using joints connected to the head and lower incisor joints, respectively.

When translating the y-dimension of the lower incisor NURBS handle, the translation values of the lower lip feature joint remain constant as a result of Maya's handling of joint rotations. To realign the EMA lower lip feature data, we calculate the global position of the lower lip joint when the lower incisor joint rotates:

$$\mathrm{ll}_x = r\cos\theta, \quad \mathrm{ll}_y = r\sin\theta$$

where r is the radius of the lower incisor joint given by the distance between the lower lip joint and the lower incisor joint and θ is the angle the lower incisor joint rotates by. Streamed EMA data then translates the calculated lower lip global position.

2.3. Input Stream Processing

During the streaming task, we use an audio input stream from a WAV file or from microphone input that provides waveform data in batches of 1600 samples, which is 0.1 seconds of data for 16 kHz sampling rate audio. Since the AAI model is not trained to infer correct EMA data for silence, we use Google's WebRTC Voice Activity Detector (VAD) to detect whether the current batch contains speech. If not, we show the previous frame for the next 0.1 seconds to emulate a period of silence.

If the batch contains speech, we deploy articulatory inversion to generate the relevant EMA data. However, 0.1 seconds of audio data provides insufficient context for accurate EMA inference, and if streamed in this manner, creates noisy animations. To remedy this issue, we employ a rolling context window for every batch of audio from the input stream. When we receive 1600 samples of audio data, we prepend that audio with the last $16000n - 3200$ samples and append the next 1600 samples to the current batch to construct a window of n seconds. This method creates an intentional 100ms delay as we initially wait for two batches of audio to be recorded before beginning inversion to include forward context in our window. We call the batch of data we send the "working" batch and the data we just sent for animation the "sent" batch.

While this provides sufficient context to the model, we may not always have n seconds of audio data to create a window when streaming. For example, when streaming begins, we only have access to small amounts of audio data being recorded. This lack of preliminary data means the model has limited context to draw from for the first n seconds of audio streamed. We try to address this by testing four sources of initial artificial context until we have n seconds of recorded context: silence, a recording of an articulated vowel, a random utterance from the HPRC dataset, and an n-second looped buffer of what data we have read so far.

The full context window is processed by the AAI model which outputs up to 100 frames of EMA data for the 12-dimensional features, corresponding to an overall 100 frames per second frequency. We discard the first 80 frames and last 10 frames of EMA and keep only the 10 frames corresponding to the 0.1 seconds of working batch audio.

Finally, even with rolling context, the BiGRU architecture only has n seconds of total audio data, and is unaware of the previous EMA outputs it has generated. The independence from batch to batch can create inconsistencies in the EMA data, where the end of one batch's EMA may not correspond exactly to the beginning of the next batch's data. To smooth out these discrepancies, we interpolate from the last frame of

the previous batch across the first four frames of the current batch using a cubic Bézier curve. After evaluating this curve on the first four frame steps of the current batch, we return the four interpolated EMA frames alongside the remaining original six EMA frames.

2.4. Streaming to Avatar

To stream EMA data from the AAI model to the facial model in Maya, we utilize concurrent processes and a shared memory buffer to facilitate fast data transfer. The first process initializes a shared memory buffer of 5 MB and begins processing the incoming audio data. Simultaneously, we use Maya's Python wrapper to begin a process that connects to the same memory buffer. Since Maya has low threading support, in this way we avoid the use of a separate thread or process to hold a queue for incoming data. Rather, we continuously check if the buffer has new data, and only continue animation if so. This also protects against potential packet loss issues because the shared memory buffer keeps all cumulative EMA data, so Maya will always have access to any data it has not animated yet.

After receiving a batch of EMA data in Maya, we transform every relevant joint then refresh the Viewport 2.0, Maya's real-time hardware renderer. Thus, at every time step, all of the EMA features will concurrently update to show their next locations, and we enable real-time speech-to-avatar streaming.

3. RESULTS

	Latency (ms)	
Streaming Portion	**1-sec window**	**2-sec window**
Model	76.7	83.3
Send	4.19	4.05
Animate	56.3	71.8
Overall	**133**	**166**

Table 1: Average latency for each portion of streaming; "Model": articulatory inversion using BiGRU, "Send": reading/writing the shared memory buffer, "Animate": transforming avatar rig and refreshing the Viewport 2.0.

3.1. Streaming Latency

Figure 3 provides an example of the latency contributions of each part of the streaming process. These are further summarized in Table 1, measuring the average latency for a 3.6 second length audio. When speech is detected, the largest latency bottleneck comes from the AAI model. The context window inversion forces the model to convert $16000n$ samples to EMA rather than just the working 1600 samples, where n is the length of the context window. We observe the second largest latency contribution is from the animation of the avatar in Maya, as the hardware renderer requires time to refresh the viewport.

We can also see that the shared memory buffer is an effective way to stream data when transferring data between processes on the same device, since its added latency is very minimal even at higher size data transfers.

	Inversion Model PCCs	
Artificial Context	**BiGRU**	**Transformer**
None	0.612	0.774
Silence	0.704	0.771
Vowel	0.705	0.762
HPRC Utterance	0.720	**0.792**
Looped Buffer	0.704	0.769

Table 2: Average Pearson correlation coefficients for predictions made by each model compared to the ground truth EMA for every method of providing initial artificial context.

3.2. Qualitative Streaming Analysis

We visually observe real-time tongue movements accurately portraying how a real tongue would articulate the streamed utterances. For example, Figure 2 highlights the movement of the tongue during the "a-ta" portion of a "pa-ta-ka" utterance. We qualitatively determine the tongue's position near the front of the teeth matches how one would articulate the same utterance. Additionally, we observe that the avatar correctly portrays tongue positions during prolonged vowel sounds. For example, the tongue tip nearing the lower incisor during an "ah" sound or the tongue receding further towards the jaw during an "ooh" sound support the model's ability to make accurate predictions.

3.3. Streaming Articulatory Inversion Analysis

Figure 4 highlights visual similarities between predicted and ground truth articulator traces for each EMA feature and dimension. We observe high Pearson correlation coefficients for each feature, markedly improving after the initial 50-100 frames. Despite best efforts to mask the lack of source data when streaming begins, the first 0.5-1 second of EMA data may be noisier than latter predictions. Generally in the streaming task, the Transformer model achieves higher correlations with the reference EMA data compared to the BiGRU model across all artificial contexts as seen in Table 2.

4. CONCLUSION

In this work, we present a real-time speech-to-avatar approach using acoustic-to-articulatory inversion, reaching an average of 130ms latency for every batch of 0.1 seconds of audio data. Our results enable low-latency speech-driven

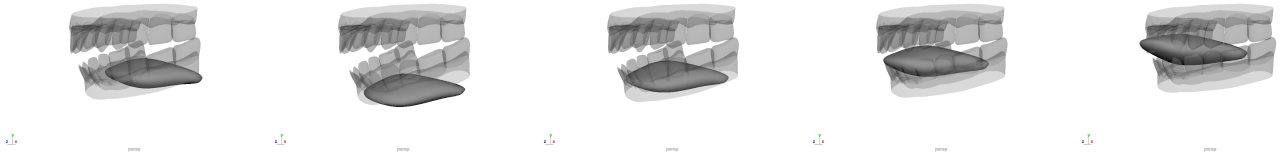

Fig. 2: “pa-ta-ka” - Series of five frames generated for audio length of 0.1 seconds

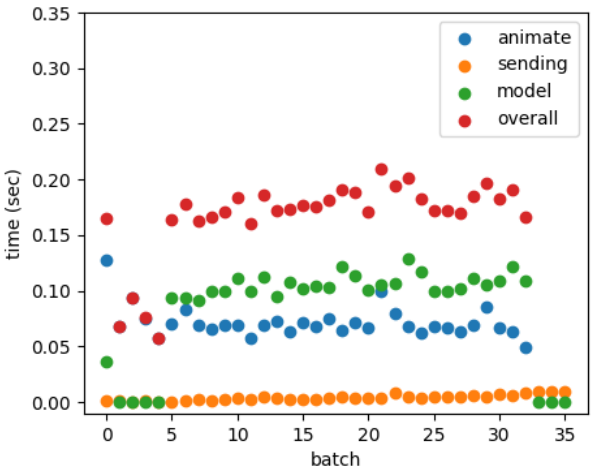

Fig. 3: 2-sec window profiling test for real-time performance at 100fps; each batch represents 0.1 seconds of audio data, and time in seconds represents the time required for a given subprocess during streaming

streaming of true-to-life mouth and tongue animation. We also demonstrate the efficacy of a shared memory buffer for streaming over a single device. To the best of our knowledge, this approach is the first to facilitate real-time avatar tongue and face animations from speech using deep learning models. We show compelling visual demonstrations of real-time microphone-based streaming for practical use. In the future, we plan to improve the accuracy of predicting real-time EMA data using transduction inversion techniques and eliminate the need for manually provided context.

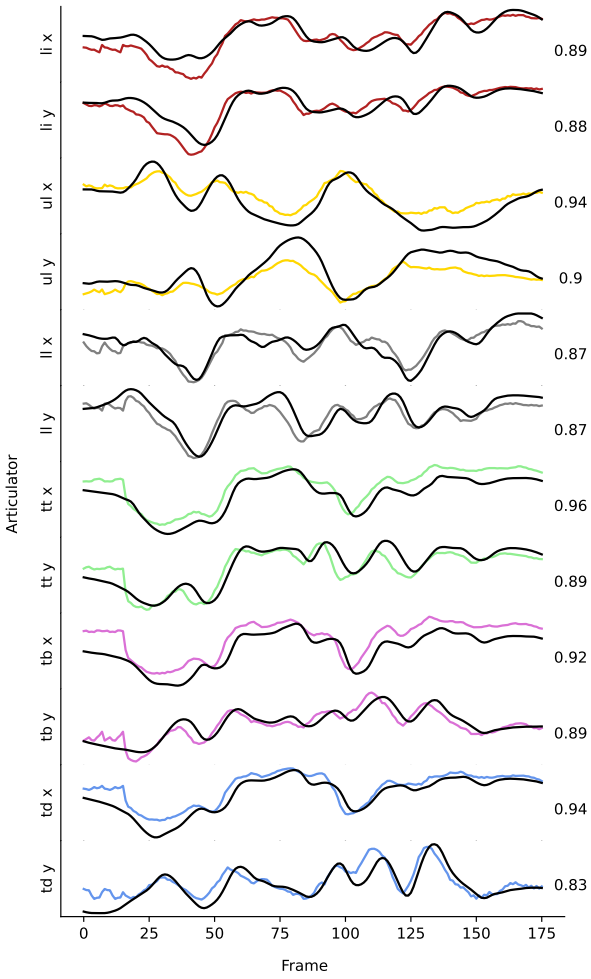

Fig. 4: Midsagittal articulator movements inferred from streamed audio data using the Transformer-based model (in color). The trace of the reference EMA data also shown (in black). Pearson correlation coefficients (PCCs) comparing predicted trajectories to ground truth are shown to the right of each feature’s plot.

5. REFERENCES

[1] 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.

[2] 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*, 2010, pp. 1862–1865.

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

[4] 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*, vol. 620, no. 7976, pp. 1037–1046, Aug. 2023.

[5] Jonas Beskow, Inger Karlsson, Jo Kewley, and Giampiero Salvi, "Synface - a talking head telephone for the hearing-impaired," *Lecture Notes in Computer Science*, 2003.

[6] Giampiero Salvi, Jonas Beskow, Samer Al Moubayed, and Björn Granström, "SynFace—speech-driven facial animation for virtual speech-reading support," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, pp. 1–10, 2009.

[7] Mauricio Radovan and Laurette Pretorius, "Facial animation in a nutshell," in *Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing couuntries - SAICSIT '06*. 2006, ACM Press.

[8] 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)*, 2019, pp. 10101–10111.

[9] Monica Villanueva Aylagas, Hector Anadon Leon, Mattias Teye, and Konrad Tollmar, "Voice2face: Audio-driven facial and tongue rig animations with cVAEs," *Computer Graphics Forum*, vol. 41, no. 8, pp. 255–265, Dec. 2022.

[10] Pierre Badin, Pascal Borel, Gérard Bailly, Lionel Revéret, Monica Baciu, and Christoph Segebarth, "Towards an audiovisual virtual talking head: 3d articulatory modeling of tongue, lips and face based on mri and video images," 1998.

[11] Rui Li and Jun Yu, "An audio-visual 3d virtual articulation system for visual speech synthesis," in *2017 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE)*. Oct. 2017, IEEE.

[12] 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)*. June 2022, IEEE.

[13] Kai Zhao, Zhiyong Wu, and Lianhong Cai, "A real-time speech driven talking avatar based on deep neural network," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–4.

[14] Peter Wu, Li-Wei Chen, Cheol Jun Cho, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala K. Anumanchipalli, "Speaker-independent acoustic-to-articulatory speech inversion," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, IEEE.

[15] 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," *TASLP*, 2021.

[16] David Gaddy and Dan Klein, "Digital voicing of silent speech," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 5521–5530, Association for Computational Linguistics.

[17] 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*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[18] JL Penkoff, "Upper and lower 3d model," `https://www.turbosquid.com/3d-models/teeth-3d-model-1636970`, 2020.

[19] Mad Mouse Design, "Male head," `https://www.turbosquid.com/3d-models/male-head-obj/346686`, 2007.

DEEP ARTICULATORY MRI FEATURE EXTRACTION AND SPEECH SYNTHESIS

Anonymous

Anonymous Institution

ABSTRACT

Recent developments in articulatory synthesis have shown that representations of physical vocal tract features contain vital information for speech generation and can be used as an interpretable and generalizable representation for a variety of tasks in speech modeling and linguistics. Previous works analyzing MRI features show that this modality encapsulates important speech-related features that other articulatory spaces cannot capture. However, despite the increasing abundance of articulatory MRI recordings, the extraction of an interpretable feature set for speech tasks remains computationally expensive and requires human supervision. In this paper, we introduce a unified multi-speaker MRI feature extraction model which efficiently generates an articulatory feature set from real-time MRI video for speakers across age, gender, and accents. We further show that features extracted by our model allow for articulatory speech synthesis that outperforms current state-of-the-art MRI synthesis benchmarks, and enable large-scale speech research using vocal tract MRIs.

***Index Terms*—** Speech, Articulatory synthesis, Magnetic Resonance Imaging (MRI), Speech representation, Medical image detection

1. INTRODUCTION

Articulatory understanding of speech is a promising method of interpreting speech signals on a large scale across modalities. As it involves directly modeling speech in terms of the physical movements of articulators, it is both generalizable and interpretable. Representations of speech using physical vocal tract intermediates can be used to synthesize speech across speaker and dialect variations [1, 2, 3], with popular self-supervised speech representations already proving to be strongly correlated with articulatory features [4]. In addition, articulatory representations provide a vital bridge between the physiological processes of speech production and the resulting phonemes, linking everything from neural speech and soft tissue movement to vocal waveforms [5]. While other modalities such as electromagnetic articulography (EMA) and electromyography (EMG) have successfully been used to recreate certain aspects of speech [6, 7], magnetic resonance imaging (MRI) allows for continuous global imaging of the vocal tract, providing unique insight into phonological theory and speaker modelling [8]. Already, data from real-time MRI (rtMRI) of the vocal tract has shown success in synthesizing speech and understanding causes of speech disfluency [9, 10, 11].

Current work using real-time articulatory MRI falls into two broad categories: (1) those which rely on previously extracted articulator labels from raw rtMRI videos [9, 11], or (2) models which directly work with the videos, but do not contain an interpretable intermediate representation [10, 12].

We solve this by introducing a multi-speaker feature extractor, which efficiently extracts articulatory features from raw rtMRI and enables downstream use in speech-related tasks. The existing method for extraction of articulatory boundaries requires detailed human annotation of a frame and subsequently takes 20 minutes to label each remaining frame on a typical desktop computer [13]. In contrast, our model uses a deep fully-convolutional network which takes just seconds to label an entire video and is less prone to mislabelling without any human supervision during the labelling process. We further show that the features extracted by this model from unseen speakers can be used to synthesize higher fidelity speech than the current state-of-the-art models while also maintaining an interpretable intermediate representation.

2. ARTICULATOR RECOGNITION MODEL

Our model uses an attention-gated fully convolutional network to efficiently map MRI images to articulatory keypoints. The model is multi-speaker and generalizes across age, gender, and accents.

2.1. Data Collection

We use the real-time articulatory MRI recordings of 7 speakers in the TIMIT database [14] speaking phonetically diverse sentences. Subjects read sentences off a projection screen as they lay supine in an MRI scanner with their heads steady. Images were reproduced from the two anterior coils, and the extraction of the midsagittal plane results in a spatial resolution of 84 x 84 pixels sampled at 83 frames per second. Each frame of the recording includes 170 annotated points aligned with anatomical features, generated by the original feature extraction method [13], which we use as our ground truth labelling, and will subsequently be referred to as the "ground truth" points. Each of these videos have aligned audio recorded at 20kHz concurrent to the MRI which we subsequently processed using the Adobe Podcast toolkit to reduce reverberation from the MRI recording environment, as per the process described in Wu et al. [9].

Using these annotations, we utilize 6 of the 7 speakers for training (3 male, 3 female). Because many of the 170 points were introduced to fit the inductive biases of the original annotation algorithm, we reduce the feature set to the 115 points found to be most relevant to speech by Wu et al. [9]. During training, we applied random affine transformations to frames and the corresponding annotations to promote generalization to unseen speakers.

2.2. Model Design

The residual fully convolutional architecture (FCN), also known as the U-Net [15], has historically performed well on low resolution medical images, especially with reduced amount of training data. Because labelled data was only originally available from seven speakers, this architecture provided the best fit while also generalizing to held out speakers. We modify the traditional format to additionally include normalization layers. Input MRI images were

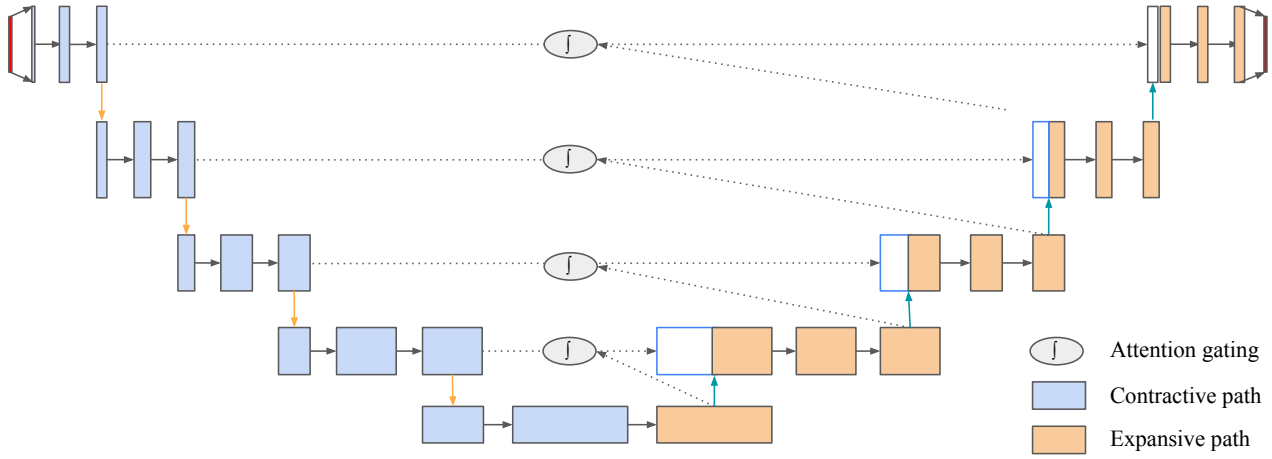

Fig. 1. Our fully convolutional model. Dotted lines represent the paths of the attention gating in each contractive/expansive layer.

padded to a spatial dimension of 96 by 96 and subsequently reduced in the spatial dimension by a factor of two in each layer of the contracting path before expansion, visualized in Figure 1.

Of the spatial features, the key articulators only occupy a subset of the space. For this reason, we learn a spatial weighting map on the residual connection to effectively suppress the components of the signal which are not important for speech. To do this, we introduce an attention gating mechanism, similar to the Attention U-Net [16]. This adds minimal additional complexity, but filters the the target spatial area from the residual activations of the corresponding contraction block before appending to the expansion block. The spatial weighting map is produced by information from both the residual activations of the contraction block (pixel vector x) and the previous expansion block (gating vector g), allowing it to take into account low-level spatial information and downstream high-level semantic context. Specifically, we pass x and g through 1x1 convolutional layers to match the spatial dimensions and use additive attention with an additional convolution to reduce to one channel. This is our learned spatial weighting which is passed through a sigmoid activation and upsampled back to the spatial dimension of our pixel vector x and multiplied to produce the final weighted activations that we then concatenate as part of the residual connection.

2.3. Training

We train this model on approximately 90 minutes of labelled midsaggital rtMRI video for a total of 6 epochs. The model outputs an 84 by 84 grid for each of the 115 articulatory points, as seen in Figure 2. Each of the 115 target keypoints are modeled as 2-dimensional Gaussian distributions over the 84 by 84 spatial grid with a standard deviation of 2 pixels. For generating keypoint locations from the output heatmaps, we took a weighted average of the k pixels with the highest output values, where the best k was found experimentally to be 25.

Typically, the pixelwise mean squared error loss, also known as L2 loss, is used for heatmap regression tasks, but we also introduce using the Kullback–Leibler (KL) divergence between the output and target grids in which each output grid is restricted to a 2-dimensional probability distributions using a softmax nonlinearity. To our knowledge, this training objective has not been used for heatmap regression in this context in the past, but appears to guide the model into producing an output that also appears Gaussian in nature, and is a natural fit for measuring the difference in the two probability distributions.

In addition, articulators in our 115 point set have varying degrees of movement and importance in speech synthesis, as determined in Wu et al. [9]. To explore this, we also try using keypoint standard deviation σ from the 6 training speakers and each keypoint's importance for speech synthesis ω measured in Wu et al. [9] to determine the weight W of a point p using the following formula, where α, β, and γ are tuned hyperparameters:

$$W(p) = \alpha\sigma(p) \times \beta(\omega(p) - \gamma)$$

This scientifically-informed weighting emphasizes the importance of articulators that show significant movement and are important to speech production over those which show minimal movement or have been found to be less essential.

3. RESULTS

We evaluate various components of our feature extraction FCN independent of downstream use with the explicit goal of labelling articulatory features that are informative for speech.

3.1. Loss Function Baseline Comparison

Heatmap regression typically utilizes on L2 loss for training. This however, does not take advantage of the fact that the target output is a probability distribution. In order to address this, we also introduce the KL-divergence between output and target probability distributions. We first train both models on the loss as is, without any articulatory weighting.

L2 Baseline Loss: The baseline L2 loss for the untrained speaker had a root mean squared error (RMSE) of 7.33 pixels. On further inspecting the outputs of the trained model, we noted that they did not visually represent a continuous Gaussian distribution.

KL-Divergence Loss: To determine whether adding the inductive bias of a probability distribution helped the model train more

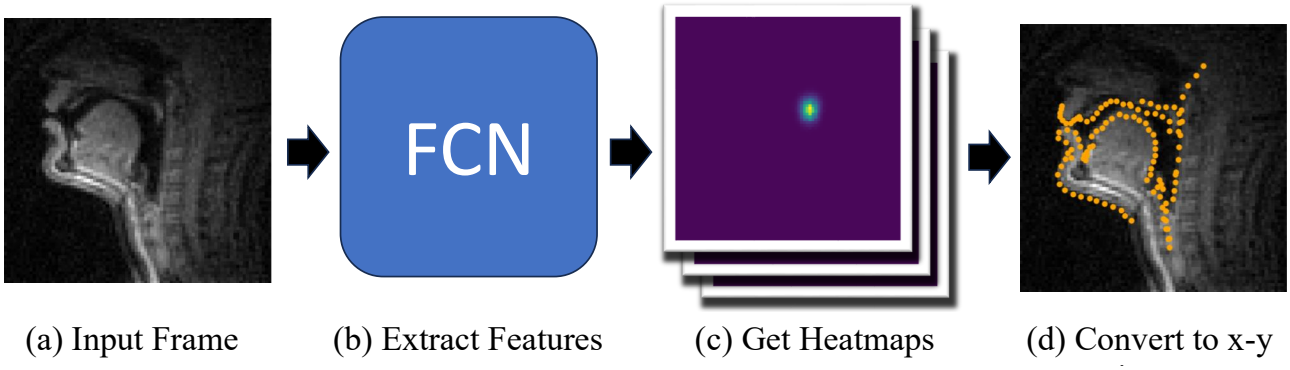

Fig. 2. Pipeline for extracting articulator x-y points from a sample MRI frame.

accurately, we applied a softmax to each 84 by 84 output grid, and used the KL-Divergence between this distribution and the target distribution as the loss for that particular keypoint. In this case, the RMSE was reduced to 3.74 pixels, with outputs resembling pointwise Gaussian distributions, similar to the target. We hypothesize that this method showed better results because it takes advantage of the fact that the output is already a probability distribution. In addition, the loss is specifically tailored towards matching distributions, and thus is well suited for heatmaps. Given that rtMRI is a lower-resolution input than many other heatmap regression tasks, the 47% improvement in RMSE results in significantly improved keypoint detection.

3.2. Loss Weighting Baseline Comparison

While the model trained using the KL-divergence exhibited better RMSE, the smaller-scale articulators which are also important for speech, such as the lips, epiglottis, and velum, were not well detected given their size and the amount they differ between frames. On the other hand, points along larger articulators, such as the pharyngeal wall, were well represented, despite minimal movement and importance in speech production. Therefore, we introduce the scientifically-informed weighting metric described in Section 2.3 to weigh the loss on the individual keypoints by their relative role in speech production.

Baseline comparison: In comparing these results to the baseline RMSE of 3.74 pixels from the equally-weighted KL-divergence loss, we see that the RMSE increased slightly to 3.92 pixels. This increase was to be expected, as the RMSE is calculated using a balanced weighting of pixels, similar to the baseline loss. However, upon further inspection of the keypoint outputs, most of this error can be attributed to slight shifts in less phonologically important articulators such as the pharyngeal wall, with significant improvement on the more important articulators. This is visualized in Figure 3. In addition, because shifts and spacing in the "ground truth" keypoints are more or less arbitrary for large articulators, the boundaries of the weighted model are still visually accurate. The results are also summarized in Table 1.

3.3. Interesting Results

When evaluating the multi-speaker model on the held out speaker, we noticed that in some recordings, the "ground truth" labellings from the original algorithm were actually blatant mislabels, an example of which is visible in Figure 4. Even so, our algorithm which used these labels as supervised targets learned to correctly fit these frames, ignoring the mislabels as outliers. This phenomenon underscores the importance of a deep learning approach to this problem, as it enables the learning of high-level semantic information for feature extraction, and thus is not as prone to mislabelling as the original pixel countour detection system.

With the increasingly large corpus of vocal tract rtMRI video available publicly [17], our efficient and accurate feature extractor is a vital step forward in allowing speech researchers to use this data effectively.

Loss	RMSE
MSE (L2)	7.33
KL-div	3.74
KL-div + Weighting	3.92

Table 1. Comparison of the root mean squared error of the models trained using L2 loss, KL-divergence loss, and KL-divergence loss with articulatory weighting. More details are available in Section 3

4. FURTHER EXPERIMENTS

To demonstrate the viability of downstream tasks using the outputs of the feature extraction model, we used generated features to run two additional experiments: (1) Test the performance of the model on rtMRI of other accents outside of the training set, and (2) Train an articulatory synthesis model on the extracted points from an unseen speaker to determine the accuracy of the predicted articulators in speech-related tasks.

4.1. Cross-accent predictions

Investigation into vocal tract shaping for English second language speakers, who produce accented speech, shows that the articulatory posturing differs from native English speakers [18]. Using the same pretrained MRI feature extraction model, which was trained using only native English speakers, we tested the zero-shot capabilities of the model to accented speech. For Indian-accented English speech, we compared predicted features to the ground truth, and found that the model performed well at extracting articulators despite vocal tract shaping differences, with a RMSE of 3.88.

This lends credence to the idea that articulatory representations of speech are universally representative and can be used as interme-

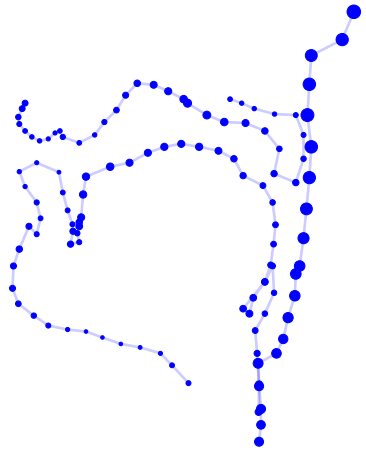

Fig. 3. Visualization of the feature extraction model's error for each keypoint when articulatory weighting is used. In the image, each keypoint's width indicates the proportion of the RMSE that can be attributed to that keypoint. We see that most of the error comes from the less important, and more arbitrarily labelled, pharyngeal wall rather than the smaller and more important articulators.

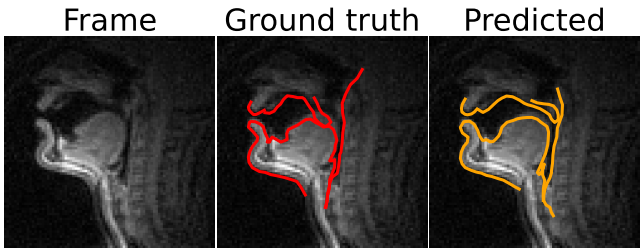

Fig. 4. One example of a frame where the "ground truth" algorithm mislabels an articulator location, in this case the velum. Here, our predicted boundaries are much more accurate, despite having been trained using the original "ground truth" algorithm as the supervised training targets.

diates for cross-accent and cross-language modeling.

4.2. Articulatory Synthesis

While RMSE and visual inspection can verify the validity of the extracted articulators to a degree, the true test of the extracted features comes with use in a downstream speech task. To do this, we use the extracted features to train an articulatory synthesis model, and compare the result to current state-of-the-art rtMRI speech synthesis. For the vocoder, we use HiFi-CAR in accordance with the current top performing MRI feature-to-speech model [9], which was trained on the "ground truth" features. We will refer to this model as "HiFi-CAR (GT)" for readibility. Using the same speaker and training data as HiFi-CAR (GT), we trained a synthesis model on our predicted features, which were centered in the same way, as described in Wu et al [9]. We will refer to this model as "HiFi-CAR (Pred)".

For an objective comparison of synthesized speech, we calculate the mel-cepstral distortion (MCD) between the synthesized speech and the original cleaned speech recording. We also use the character error rate (CER) of the synthesized speech transcribed using Whisper's automatic speech recognition (ASR) model [19]. The mean and standard deviation of the MCD across utterances are summarized in Table 2. The synthesized speech from our FCN predicted features is closer to the true speech than the current state-of-the-art MRI feature-to-speech method [9]. The only difference in these two pipelines is that the features were extracted by different algorithms, indicating that our multi-speaker feature extractor learns to do better than even the features it was trained on. We hypothesize that this is because the original algorithm is more susceptible to random perturbations between frames and does not learn semantically meaningful information. In contrast, our feature extractor is able to learn high-level feature information from multiple speakers. This result is emphasized by the improvement in CER, which is quantified in Table 2, indicating that the speech is also more interpretable when based on our extracted features.

Model	MCD	CER
HiFi-CAR (GT) [9]	6.64 ± 0.64	$69.2\% \pm 28.1\%$
HiFi-CAR (Pred)	$\mathbf{5.93 \pm 0.92}$	$\mathbf{57.3\% \pm 29.8\%}$

Table 2. MCD and ASR CER of articulatory synthesis models. Synthesis using our extracted features are shown to improve quality of synthesized speech.

The result is also comparable to the current state-of-the-art MRI video-to-speech model from Yu et al. [10], with only marginally worse performance in single-speaker MCD. The advantage of our two-step pipeline of having an independent feature extractor and vocoder, however, is that we have an interpretable feature space before synthesis, allowing for future research and understanding into speech production.

5. CONCLUSION

In this work, we looked at developing a multi-speaker feature extractor from rtMRI videos of the vocal tract. We used the limited existing articulatory labelling to train a model which efficiently and accurately extracts physiological features from MRI videos of unseen speakers indiscriminate of gender, age, and accent. In this process, we also introduced a new method for keypoint detection training. Subsequent experiments show the viability of these features in downstream speech tasks, setting a new benchmark for articulatory synthesis from MRI features. As MRI provides unprecedented insight into the physiology of speech production, this feature extraction method opens up a large repository of previously unlabelled rtMRI data for future speech research from universal speech modeling to speech pathology and everything in between.

6. ACKNOWLEDGEMENTS

Hidden for anonymity

7. REFERENCES

[1] L. Goldstein et al. P. Wu, S. Watanabe, "Deep speech synthesis from articulatory representations," in *Interspeech*, 2022.

[2] C. Scully, *Speech Production and speech modeling*, chapter Articulatory synthesis, Springer, 1990.

[3] G. Fant, "What can basic research contribute to speech synthesis?," *Journal of Phonetics*, vol. 19, pp. 75–90, 1991.

[4] Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K. Anumanchipalli, "Evidence of vocal tract articulation in self-supervised learning of speech," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[5] Littlejohn K.T. Silva A.B. et al. Metzger, S.L., "A high-performance neuroprosthesis for speech decoding and avatar control.," vol. 620, pp. 1037–1046, 2023.

[6] Peter Wu, Li-Wei Chen, Cheol Jun Cho, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala K. Anumanchipalli, "Speaker-independent acoustic-to-articulatory speech inversion," 2023.

[7] David Gaddy and Dan Klein, "Digital voicing of silent speech," 2020.

[8] Shrikanth S. Narayanan, Erik Bresch, Prasanta Kumar Ghosh, Louis M. Goldstein, Athanasios Katsamanis, Yoon-Chul Kim, Adam C. Lammert, Michael I. Proctor, Vikram Ramanarayanan, and Yinghua Zhu, "A multimodal real-time mri articulatory corpus for speech research," in *Interspeech*, 2011.

[9] 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," 2023.

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

[11] 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*, vol. 97, 2022.

[12] Yuto Otani, Shun Sawada, Hidefumi Ohmura, and Kouichi Katsurada, "Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," in *Proc. INTERSPEECH 2023*, 2023, pp. 127–131.

[13] 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*, vol. 28, no. 3, pp. 323–338, 2009.

[14] 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*, vol. 136, pp. 1307, 09 2014.

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

[16] 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.

[17] Yongwan Lim, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Lingala, Colin Vaz, Tanner Sorensen, Miran Oh, Sarah Harper, Weiyi Chen, Yoonjeong Lee, Johannes Töger, Mairym Llorens Monteserin, Caitlin Smith, Bianca Godinez, Louis Goldstein, Dani Byrd, Krishna Nayak, and Shrikanth Narayanan, "A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images," *Scientific Data*, vol. 8, 07 2021.

[18] A. Benítez, Vikram Ramanarayanan, L. Goldstein, and Shrikanth Narayanan, "A real-time mri study of articulatory setting in second language speech," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 701–705, 01 2014.

[19] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

ACOUSTIC-TO-ARTICULATORY INVERSION FOR MULTI-LINGUAL AND DOWNSTREAM TASKS

Anonymous
Anonymous Institution

ABSTRACT

Deep acoustic-to-articulatory inversion models provide a direction for efficiently annotating speech with articulatory features. Current models are primarily monolingual, with cross-lingual performances around 0.5 Pearson correlation. To help bridge this gap, we propose an inversion model with cross-lingual correlations of up to 0.78. We also show how EMA features estimated with our model can improve performance on downstream speech tasks, including speech perceptual quality estimation, sentiment analysis, and pronunciation scoring.

***Index Terms*—** articulatory inversion, articulatory speech processing

1. INTRODUCTION

Articulatory representations of speech aim to model speech in terms of vocal tract features, providing an interpretable quantification of speech production [1]. Acoustic-to-articulatory inversion is an approach to automatically label speech with articulatory features, offering an alternative to expensive and potentially invasive data collection methods [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25].

In this work, we focus on estimating electromagnetic articulography (EMA) data from speech. EMA data is comprised of locations of parts of the mouth, specifically the upper lip, lower lip, lower incisor, tongue tip, tongue blade, and tongue dorsum, via electromagnetic fields [26, 2, 27]. Many researchers have investigated ways to build an invertible mapping from speech to EMA [3, 2, 9, 6, 5, 8, 21, 22, 23, 24, 25]. While models can achieve cross-speaker generalization [2, 3], to our knowledge, cross-language generalization remains challenging, with current methods generally yielding around 0.5 Pearson correlation between estimated and ground truth EMA [24, 25]. We note that for speakers who have native languages that share a language family, inversion models may be able to more readily generalize across these speakers [23]. This paper focuses on estimating EMA for target languages that are fairly different from source languages seen during training. Like prior work [2, 3], we also estimate tract variables (TV) derived from EMA and analyze results in the TV space.

We propose a speech-to-EMA model, detailed in Section 3, that improves cross-speaker TV estimation performance from the previous 0.782 Pearson correlation to 0.808. Our inversion model also yields unseen-language EMA estimation correlations of up to 0.784, which is higher than the previous correlation reported when training using that language, 0.778 [22]. Finally, we improve performance on multiple downstream speech tasks through utilizing EMA features estimated with our model. These tasks include speech perceptual quality estimation, sentiment analysis, and pronunciation scoring, detailed in Sections 6.1 through 6.3.

2. EMA DATASETS

2.1. HPRC

We train our inversion model on the Haskins Production Rate Comparison database (HPRC) [28]. HPRC is an 8-speaker dataset containing 7.9 hours of 44.1 kHz speech and 100 Hz EMA. In order to match sampling rate used in our input feature extractor, we downsample the audio to 16 kHz. Like prior work [2, 3], we only consider information along the midsagittal plane and do not use the provided mouth left and jaw left data. We use the same train-val-test split as [2], holding out 1 male and 1 female speaker for our test set and having a 90-10 train-val split on the remaining 6 speakers.

2.2. DKU-JNU-EMA

The DKU-JNU-EMA dataset has EMA and speech for Cantonese, Hakka, and Teochew, with two to seven native speakers for each language [21]. We divide the utterances into 137 minutes, 8 minutes, and 16 minutes of speech, using a 85-5-10 train-dev-test split. We then train our inversion model on the train set and report the test correlation in Table 2, denoted as "monolingual." EMA data here has a sampling rate of 200 Hz and speech waveforms 16 kHz.

2.3. MSPKA

The MSPKA dataset has EMA and speech for one male and two female Italian speakers, totaling two hours of speech [22]. Similarly to DKU-JNU-EMA, we divide the utterances into 113 minutes, 7 minutes, and 14 minutes following 85-5-10 train-dev-test split. We train our inversion model on

the train set and report the test correlation in Table 3, denoted as "monolingual." All EMAs recorded have 200 Hz sampling rate. All speech waveforms use 22050 Hz sampling rate, which we downsample to 16 kHz before any further processing.

3. INVERSION MODEL

3.1. Model Architecture

We explore 4 different model architectures for articulatory inversion: (1) bidirectional GRU (BiGRU) [29, 30, 2], (2) bidirectional LSTM (BiLSTM) [31], (3) temporal convolutional network (TCN) [3], and Transformer [32, 33]. Our BiGRU architecture followed prior speech-to-EMA work [30, 2], and our BiLSTM architecture replaces each GRU with an LSTM. Our TCN follows [3]. We used 128 and 64 channels for the 4th and 5th convolution layer. For the 4th, 5th, 6th convolution layer, we used kernel sizes of [1, 3, 3], a stride of 1, and paddings of [0, 1, 1] respectively. For our Transformer-based model, we follow [34], with 6 layers and 1024 hidden dimensions. We used 1024 channels for all prepended convolutional layers, with kernel size of 3, stride of 1, and padding of 1.

3.2. Input Feature Extractor

For our input feature extractor, we use the tenth layer of WavLM, a Transformer-based model trained using self-supervised learning [35, 36, 33]. [1] We observed these features to outperform the HuBERT features used in [37, 2]. Since WavLM outputs 50 Hz features and our EMA is 100 Hz, we add a 1D convolution layer with kernel size of 3, stride of 2 and padding of 1 before the last linear projection layer to halve the time dimension.

3.3. Multi-Task Learning

Prior work has shown that multi-task learning can improve articulatory inversion performance [30, 2]. Building on methods jointly training EMA and phoneme prediction, we predict EMA, tract variables, phonemes, and pitch at the same time. Our loss function includes weighted summation of L1 TV and EMA prediction losses, a cross-entropy phoneme prediction loss, and a L1 pitch prediction loss. We used the weights [1.0, 0.5, 0.1] for each component respectively.

We also experimented with predicting periodicity, the distance between EMA points, and HuBERT-Soft [38]. We note that including both periodicity and distance is generally worse than including each feature alone. In Table 1, we denote model with periodicity concatenated to the variables above as "Periodicity", and the model with distance "Distance." We denote adding HuBERT-Soft as "HuB.-Soft". We estimate pitch and periodicity from waveforms using CREPE [39].

[1] We used the wavlm-large model in https://github.com/s3prl/s3prl.

Table 1: Pearson correlations on HPRC test set speakers.

model	EMA ↑	TV ↑
Transformer Norm. HuB.-Soft	**0.800**	**0.808**
Transformer Norm.	**0.798**	**0.803**
Transformer Norm. Periodicity	0.795	0.802
BiLSTM Norm. Periodicity	0.795	0.802
BiLSTM Norm.	0.794	0.801
Transformer Norm. Distance	0.793	0.805
BiLSTM Norm. Distance	0.791	0.799
BiGRU Norm.	0.786	0.793
Transformer [34]	0.771	0.791
TCN [3]	0.768	0.698
BiGRU [2]	0.757	0.782
BiLSTM	0.753	0.780

3.4. Normalizing EMA

To further improve model performance, we ablated three different re-scaling methods for EMA and TV. As a baseline, we scale EMA and TV at the speaker level to [-1, 1], as in prior work [30, 2]. We then experimented normalizing EMA and TV at speaker level or at sample level to have 0 mean and unit variance. We note that normalizing at the speaker level achieved the best performance. In Table 1, we denote this normalization approach as "norm" and all entries without "norm" follow the [-1, 1] re-scaling approach.

4. MULTI-SPEAKER SPEECH-TO-EMA

Table 1 summarizes our inversion results with the HPRC dataset [28]. Our best proposed models noticeably outperform the previous model, "BiGRU" [2], increasing the EMA and TV correlations to 0.8. We note our best TV performance, 0.808, is noticeably larger than 0.784, the best TV performance from prior work [2]. While our best model in Table 1 utilizes HuB.-Soft, we found that the second-best model not using HuB.-Soft and otherwise the same performed better in initial cross-lingual experiments. We hypothesize that this is due to HuBERT only being trained on English. Thus, we use the second-best model, a Transformer with normalized EMA, for our multilingual experiments in Section 5. We also observed that combinations of HuB.-Soft, periodicity, and distance were worse than not combining them, so we have omitted these results for readability.

5. MULTILINGUAL SPEECH-TO-EMA

Tables 2 and 3 summarize our cross-lingual EMA prediction results. Specifically, we use our model trained only on English data and estimate EMA for Chinese languages and Italian, upsampling EMA estimations to 200 Hz in order to match the sampling rate of the ground truth. We compare this

Table 2: Mean test set EMA correlations for DKU-JNU-EMA. Monolingual refers to the model trained on the test speaker's data. HPRC-Test and HPRC-All use our model trained on HPRC, with the former evaluated on the speaker's test set and the latter evaluated on all of the speaker's utterances.

speaker	Monolingual	HPRC-Test	HPRC-All
canfxy	0.317	0.308	0.332
cangjc	0.291	0.269	0.275
mandxy	0.733	0.552	0.561
mandyb	0.599	0.526	0.505
manlww	0.815	0.656	0.666
manlxz	0.548	0.479	0.465
manly	0.841	0.716	0.710
manzy	0.591	0.530	0.535
manzzj	0.737	0.671	0.657
Average	0.684	0.585	0.578

Table 3: Mean test set EMA correlation for MSPKA.

speaker	Monolingual	HPRC-Test	HPRC-All
cnz	0.913	0.784	0.782
lls	0.780	0.603	0.607
olm	0.766	0.636	0.670
Average	0.825	0.679	0.689

performance with single-speaker, monolingual baselines that use our proposed model and each train on one test speaker, using the splits described in Sections 2.2 and 2.3. Our inversion model yields unseen-language EMA estimation correlations of up to 0.784, which is higher than the previous correlation reported when training using that language, 0.778 [22]. While monolingual models perform slightly better, we observe strong cross-lingual inversion performance with both datasets, as visualized in Figures 1 and 2.

6. DOWNSTREAM TASK RESULTS

We utilize our inversion model in three downstream tasks: (1) predicting perceptual voice qualities, (2) sentiment analysis, and (3) pronunciation Scoring. In order to emphasize the usefulness of the input features rather than the model, we use linear regression and classification models for these tasks. We use ElasticNet, a combination of ridge and lasso regression, for linear regression. Specifically, we map 20 openSmile functional features extracted from EMA trajectories to non-categorical features, such as perceptual qualities of the voice. We used Linear Discriminant Analysis for tasks with categorical variables, i.e., emotion classification As a comparison, we applied the full set 2016 openSmile features on speech waveforms, extracting over 6300 features, and followed the same regression and classification step above. We denote features extracted from EMA as "EMA" and those extracted from speech as "Speech" in Tables 4 and 5. We also concatenated speech features and EMA features and performed the same regression analysis, denoted as "EMA + Speech" in the tables. For regression, we used alpha = 0.01 and set the maximum iteration to be 100k. For each feature, we report the best result from four regression algorithms in Table 4. All tasks have a 90-10 train-test split.

Table 4: Downstream regression and classification results.

Task	Speech	EMA	EMA + Speech
PVQD ↓	0.480	**0.250**	0.430
Speechocean762 ↓	0.617	0.647	**0.607**
ICAP (Accuracy) ↑	0.203	0.255	**0.367**

Table 5: PVQD regression results.

PVQD feature	Speech	EMA	EMA + Speech
breathiness ↓	0.470	**0.220**	0.420
loudness ↓	0.430	**0.210**	0.380
pitch ↓	0.510	**0.240**	0.430
roughness ↓	0.470	**0.290**	0.450
severity ↓	0.510	**0.300**	0.510
strain ↓	0.490	**0.260**	0.370

6.1. Predicting Perceptual Voice Qualities

The Perceptual Voice Qualities Database (PVQD) is comprised of clinical voice recordings assessed using the CAPE-V protocol [40]. The following qualities are rated on a 100-point scale: Overall severity, Roughness, Breathiness, Strain, Pitch, and Loudness [40]. From tables 4 and 5, we observe that EMA inputs and combined EMA-speech inputs both outperform speech-only inputs, suggesting that our EMA estimates are useful for improving downstream perceptual quality assessment.

6.2. Sentiment Analysis

The interactive emotional dyadic motion capture database (IEMOCAP) is a dataset for understanding expressive human communication, containing speech annotated with sentiment [41]. This dataset contains data for ten actors recorded during both scripted and spontaneous spoken communication scenarios [41]. We predict sentiment labels, with results in Table 4. Like our PVQD results, we observe that inputs containing EMA-based features noticeably outperform speech-only features.

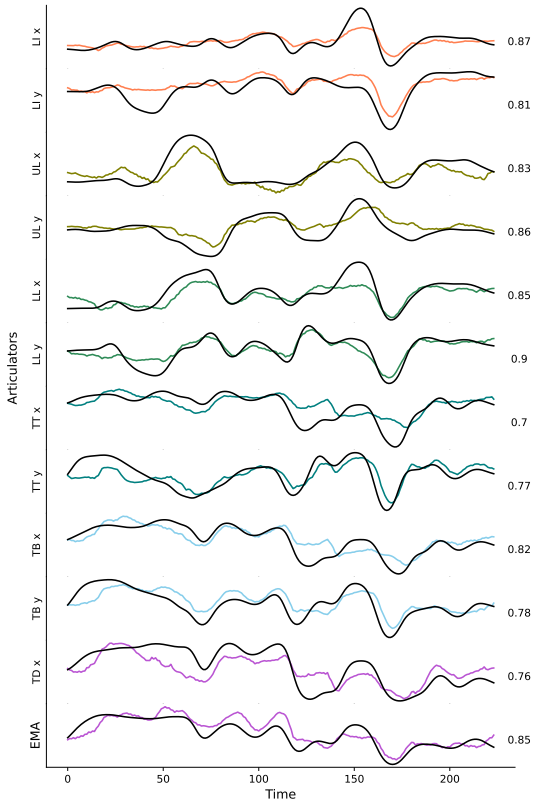

Fig. 1: Predicted midsagittal articulator movements for an Italian utterance (in color), using Transformer Norm trained on only English. The ground truth trace is also shown (in black).

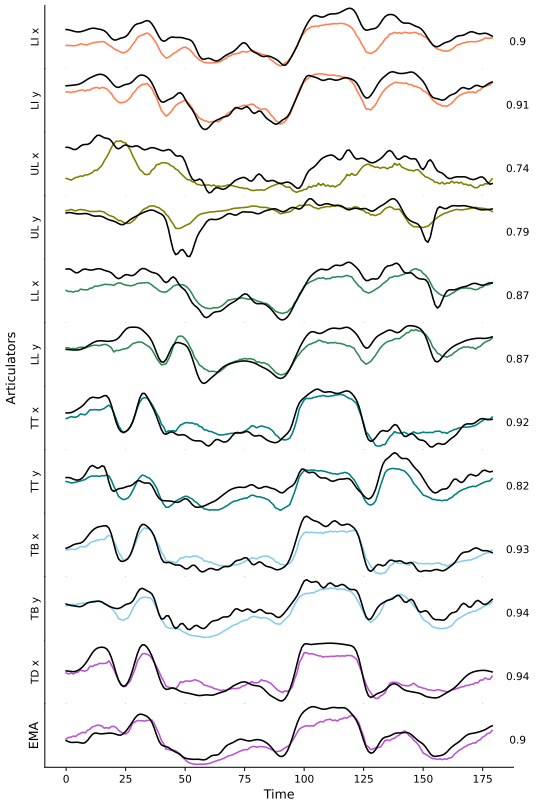

Fig. 2: Predicted midsagittal articulator movements for a Chinese utterance (in color), using Transformer Norm trained on only English. The ground truth trace is also shown (in black).

6.3. Pronunciation Scoring

Speechocean762 is a dataset for pronunciation scoring tasks, conmprised of 5000 English utterances from 250 non-native speakers whose first language is Mandarin [42]. Five experts provided sentence-level annotations on pronunciation accuracy, stress, fluency, and prosody [42]. We predict these annotations, with results in Table 4. Similarly to the other downstream tasks, incorporating estimated EMA features into the input improves performance.

7. CONCLUSION

In this work, we propose an inversion model that improves cross-speaker TV estimation performance from the previous 0.782 Pearson correlation to 0.808. Our inversion model also yields unseen-language EMA estimation correlations of up to 0.784, which is higher than the previous correlation reported when training using that language, 0.778. Finally, we also show how EMA features estimated with our model can improve performance on downstream speech tasks, including speech perceptual quality estimation, sentiment analysis, and pronunciation scoring. In the future, we are interested in exploring more downstream tasks and languages.

8. REFERENCES

[1] Paul Mermelstein, "Articulatory model for the study of speech production," *JASA*, 1973.

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

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

[4] Peng Liu et al., "A deep recurrent approach for acoustic-to-articulatory inversion," in *ICASSP*, 2015.

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

[6] Tomoki Toda et al., "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *ICSLP*, 2004.

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

[8] Korin Richmond, *Estimating articulatory parameters from the acoustic speech signal*, Ph.D. thesis, University of Edinburgh, 2002.

[9] An Ji, *Speaker independent acoustic-to-articulatory inversion*, Ph.D. thesis, Marquette University, 2014.

[10] Asterios Toutios and Konstantinos Margaritis, "A rough guide to the acoustic-to-articulatory inversion of speech," in *HERCMA*, 2003.

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

[12] Jianrong Wang et al., "Acoustic-to-articulatory inversion based on speech decomposition and auxiliary feature," in *ICASSP*, 2022.

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

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

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

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

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

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

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

[20] Paul K. Krug et al., "Self-Supervised Solution to the Control Problem of Articulatory Synthesis," in *Interspeech*, 2023.

[21] 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.

[22] Claudia Canevari et al., "A new italian dataset of parallel acoustic and articulatory data," in *Interspeech*, 2015.

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

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

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

[26] Alan Wrench, "The mocha-timit articulatory database," 1999.

[27] Peter Wu et al., "Deep speech synthesis from articulatory representations," *Interspeech*, 2022.

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

[29] Junyoung Chung et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv*, 2014.

[30] Yashish M Siriwardena et al., "Acoustic-to-articulatory speech inversion with multi-task learning," *Interspeech*, 2022.

[31] Sepp Hochreiter et al., "Long short-term memory," *Neural computation*, 1997.

[32] Sathvik Udupa et al., "Streaming model for acoustic to articulatory inversion with transformer networks," in *Interspeech*, 2022.

[33] Ashish Vaswani et al., "Attention is all you need," *NeurIPS*, 2017.

[34] David Gaddy and Dan Klein, "An improved model for voicing silent speech," in *IJCNLP*, 2021.

[35] Sanyuan Chen et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *J-STSP*, 2022.

[36] Shu wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021.

[37] Wei-Ning Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, 2021.

[38] Benjamin van Niekerk et al., "A comparison of discrete and soft speech units for improved voice conversion," in *ICASSP*, 2022.

[39] Jong Wook Kim et al., "Crepe: A convolutional representation for pitch estimation," in *ICASSP*, 2018.

[40] Patrick R. Walden, "Perceptual voice qualities database (pvqd): Database characteristics," *Journal of Voice*, 2022.

[41] Carlos Busso et al., "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[42] Junbo Zhang et al., "speechocean762: An open-source non-native english speech corpus for pronunciation assessment," in *Interspeech*, 2021.

ARTICULATORY SYNTHESIS WITH MULTI-MODAL AND SELF-SUPERVISED FEATURES

Anonymous
Anonymous

ABSTRACT

Deep articulatory synthesis involves synthesizing speech with features describing the physiology of the vocal tract. Two popular articulatory modalities include MRI and EMG. In this work, we propose two approaches for improving articulatory synthesis from these modalities. First, we analyze cross-modal relationships with EMA, another popular articulatory feature, and devise a multi-modal pre-training approach based on this analysis. Second, we propose dimensionality reduction methods for making self-supervised features better-suited for articulatory synthesis. Compared to prior work, our approach combining these two techniques improves MRI-to-speech performance by 59% WER. Moreover, we observe that our methodology noticeably improves performance on an EMG-to-speech task.

***Index Terms*—** articulatory synthesis, articulatory speech processing

1. INTRODUCTION

Deep speech synthesis algorithms have produced promising results for text-to-speech [1, 2], client-side privacy [3, 4], speech translation [5, 6], and augmentative and alternative communication [7, 8, 9, 10]. However, current methods still face generalizability challenges, suggesting that these speech synthesizers are underspecified [7, 11]. Deep articulatory synthesis aims to solve this challenge by grounding deep speech synthesizers in vocal tract features [12, 13, 14, 15, 16, 17]. We focus on two popular articulatory modalities, real-time magnetic resonance imaging (MRI) and electromygraphy (EMG). While prior work showed that synthesizing speech from MRI [14, 18] and EMG [10, 9] is feasible, making these synthesizers generalizable remains challenging. To help bridge this gap, we devise multi-modal pre-training methods and modified self-supervised features that improve speech synthesis from both modalities. With less than 10 minutes of single-speaker training data, our MRI-to-speech model achieves a test-set automatic speech recognition (ASR) character error rate (CER) of 18%, compared to 69% from the previous model [14]. Our EMG-to-speech model similarly noticeably outperforms the baseline, and both ASR results match our human listening tests.

2. ARTICULATORY MODALITIES

2.1. Electromagnetic Articulography

Electromagnetic Articulography (EMA) data is comprised of the midsagittal x-y coordinates of 6 articulatory positions: lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum [19, 12]. In this work, we use the Haskins Production Rate Comparison database (HPRC), an 8-speaker dataset containing 7.9 hours of 44.1 kHz speech and 100 Hz EMA [20]. To maintain consistency with prior work [12, 17, 16], we focused only on the midsagittal plane and discarded the provided mouth left and jaw left data in HPRC. We utilize HPRC in our multi-modal pre-training approach, detailed in Section 4.

2.2. Magnetic Resonance Imaging

Real-time magnetic resonance imaging (MRI) provides a much more comprehensive feature set of the human vocal tract than EMA [21, 22, 14, 18]. Specifically, midsagittal MRI images not only contain the six locations described by EMA, but also the hard palate, pharynx, epiglottis, velum, and larynx, all of which are informative for speech synthesis [14]. In this work, we used the same 11-minute, single-speaker real-time MRI dataset as [14]. This dataset is comprised of 20 kHz speech and $83.\bar{3}$ Hz midsagittal MRI data, with 170 x-y points annotated for each MRI frame. Following [14], we applied the same speech enhancement technique to denoise target audio and used the same 200-11-25 train-dev-test split on the 236 utterances. We normalize each MRI dimension to have a range of $[-1, 1]$. Additionally, we discarded the annotated points on the back, reducing the number of points from 170 to 155, which we observed to improve MRI-to-speech performance. Figure 2 depicts these 155 points.

2.3. Electromyography

Surface electromyography (EMG) measures electrical potentials caused by nearby muscle activity using electrodes placed on top of the skin [23]. When placed near articulators, EMG provides another low-dimensional manifold of articulatory movements [23, 8, 10, 9]. In this work, we use the EMG dataset in [10], which consists of EMG data and speech for vocalized utterances. We use the 3.9-hour vocalized speech subset, denoted "Parallel Vocalized Speech" in

[10]. Our train-dev-test data split contains 195 minutes, 12 minutes, and 23 minutes of speech, respectively. Speech has a sampling rate of 16 kHz, and EMG 1000 Hz.

3. ARTICULATORY SYNTHESIS MODELS

3.1. Transformer

To map articulatory features to acoustics, we explore multiple intermediate representations, namely Mel spectrograms, HuBERT [24], and our proposed modified self-supervised representations detailed in Section 5. For our model mapping articulatory features to intermediate features, we build on EMG-to-spectrum model proposed by [10]. Namely, we map articulatory representations to the intermediate self-supervised speech representations using a six-layer Transformer [25] prepended with three residual convolution blocks. We trained the Transformer using the L1 loss function, the Adam optimizer [26] with betas [0.5, 0.9], and a batch size of 16. During training, we randomly cut a 0.5 seconds to 2 seconds window from each sample in the batch, with the window length fixed within the batch.

3.2. Hifi-CAR

To map intermediate speech representations to waveforms, we use HiFi-CAR, an auto-regressive temporal convolutional network optimized with adversarial training [12, 27, 28]. Our intermediate-to-waveform models are finetuned on the VCTK dataset [29], using a HiFi-GAN [27] trained on LibriTTS [30] as pre-training weights like in [14]. Following [14], we use also use Hifi-CAR to directly synthesize speech from MRI features. We trained HiFi-CAR with a batch size of 32, the Adam optimizer with [0.5, 0.9] for beta values [26], and a learning rate starting with 0.0001 that halved every 15k steps.

3.3. Adversarial Training

Like prior articulatory synthesis work [12, 14], we explore adversarial training with all of our models using the approach from HiFi-GAN [27]. ND in the tables below refers to models that do not use adversarial training. Like [27], we set the weights of the L1 generator loss, feature matching loss, and adversarial learning loss to [45, 2.0, 1.0]. Other training hyperparameters are the same as those in Section 3.1.

4. MULTI-MODAL PRE-TRAINING

Multi-modal pre-training involves training a model with multiple modalities jointly, with the resulting model able to perform better in downstream tasks compared to models trained with less modalities [31, 32]. We extend this strategy to articulatory synthesis through pre-training with more than one articulatory modality as input and finetuning the resulting model with only the target articulatory modality.

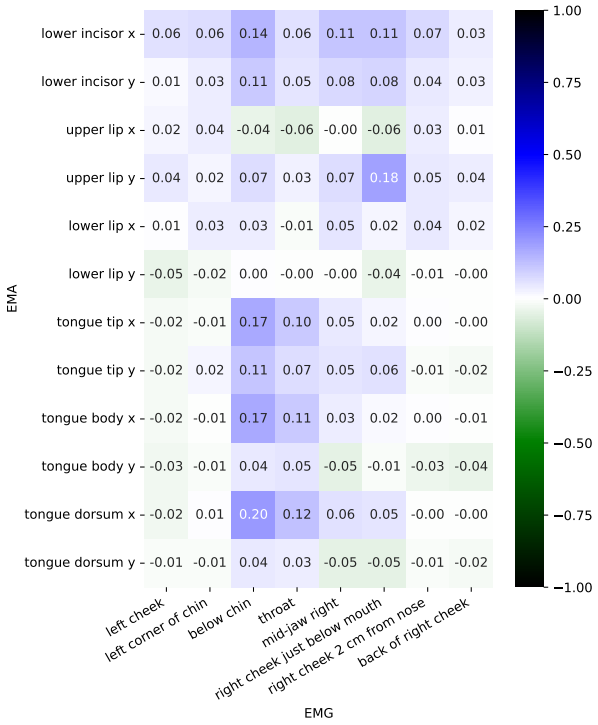

Fig. 1: Mean Correlation between EMG and inferred EMA.

First, we choose compatible pairs of articulatory modalities by analyzing their correlation across time. Specifically, for our MRI and EMG datasets described in Sections 2.2 and 2.3, we estimate aligned EMA data by predicting EMA from speech using the model in [17]. We calculate the average Pearson correlation between inferred EMA and ground-truth articulatory modalities, as shown in Figures 1 and 2. For Figure 2, we visualized correlation by coloring each MRI point in the midsagittal plane with the highest-correlation EMA point for readability, where MRI points with maximum correlation magnitude below 0.3 are omitted.

Correlations between EMA and EMG all have magnitude below 0.3, indicating a weak correlation between these modalities. On the other hand, EMA and MRI are highly correlated, with EMA points spatially aligning with MRI ones. Given these results, we focus on multi-modal pre-training with EMA and MRI in this work. Specifically, we train models with articulatory inputs using a corpus combining our EMA and MRI datasets in Sections 2.1 and 2.2. We prepend a linear layer to the model for each modality, where the output of these layers are 128-dimensional inputs to the same network. We train this multimodal model using the same hyperparameters as the models with single-modality inputs, and finetune the resulting model on the target modality dataset with the same hyperparameters. Models utilizing multi-modal pre-training contain "Multi" in the tables below.

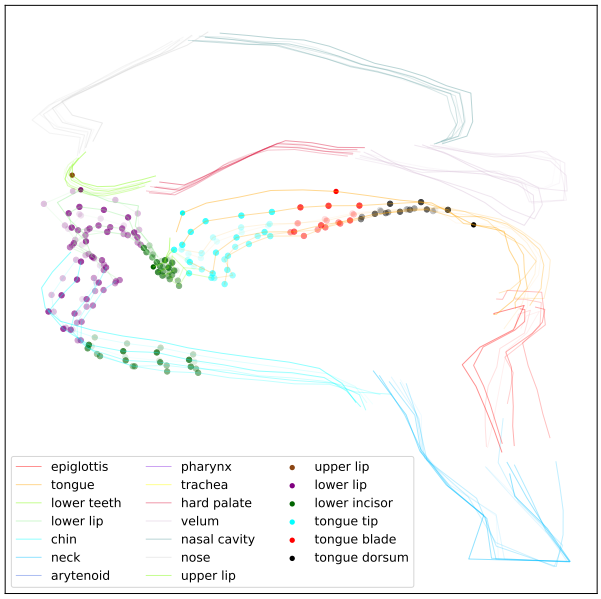

Fig. 2: Extracted MRI-features for the utterance "apa." Lighter is earlier in time. Each point is colored with the highest-correlation EMA feature. Points with maximum correlation magnitude below 0.3 are omitted.

5. SELF-SUPERVISED SPEECH REPRESENTATIONS

Self-supervised speech representations recently have been utilized to improve speech synthesis performance [7, 5]. In this work, we modify self-supervised speech representations into features features that are better-suited for articulatory synthesis. Like [7], we use the seventh layer of HuBERT, a Transformer-based model trained with self-supervision [24, 33, 25]. [1] We use this as our baseline self-supervised speech representation.

We also explore methods for reducing the dimensionality of the HuBERT manifold in order to make it easier to learn a mapping to the self-supervised feature. Specifically, we devise four methods, with the respective method involving: (1) linear projections, (2) low-pass filtering, (3) neural ordinary differential equations (ODE) [34], and (4) temporal activation regularization (TAR) [35]. All approaches linearly project HuBERT from 1024 to 256 dimensions, with the first approach only doing this. Our second method adds a differential low-pass filter with cutoff frequency of 0.4 after the linear layer. For our third method, we use a neural ODE to map each 256-dimensional frame to the next one and add a mean squared error (MSE) loss minimizing the distance between these frames. Our fourth method adds an MSE loss minimizing the distance between adjacent 256-dimensional frames without transforming them. Intuitively, methods two through four encourage the resulting feature to be smooth. To train these modified representations, we linearly project the resulting 256-dimensional features back to 1024 dimensions and compute an MSE loss between the resulting features and the ground truth ones. Thus, the final loss function is computed by adding this reconstruction loss with any additional losses mentioned for each approach. We discard the 256-to-1024 projection layer during inference and use the learnt 256-dimensional feature as an alternative to HuBERT. We used VCTK to train these modified representations [29]. Our four approaches are denoted as MLP, LP, NODE, and TAR, respectively, in the tables below. Since HuBERT accepts 16 kHz speech as input and outputs 50 Hz features, we downsample speech and linearly interpolate intermediate features to match these sampling samples. Given the success of discretized speech units as intermediates for other synthesis tasks [5, 7], we also explore using units generated from our self-supervised representations as intermediates, using k-means clustering on VCTK [29] to discretize vectors, with $k = 200$.

[1] We used the hubert_large_ll60k model in https://github.com/s3prl/s3prl.

Table 1: ASR Results for MRI-to-Speech

model	mean cer	mean wer
MLP Multi	0.20 ± 0.15	**0.31 ± 0.25**
LP Multi	**0.18 ± 0.16**	0.33 ± 0.32
HuBERT Multi	0.21 ± 0.19	0.34 ± 0.31
LP	0.28 ± 0.18	0.42 ± 0.27
TAR Multi ND	0.33 ± 0.26	0.46 ± 0.28
HuBERT	0.31 ± 0.19	0.53 ± 0.35
MLP	0.36 ± 0.26	0.57 ± 0.39
Spectrogram	0.44 ± 0.42	0.62 ± 0.51
NODE	0.46 ± 0.22	0.66 ± 0.27
HuBERT Units ND	0.66 ± 0.30	0.83 ± 0.22
Direct	0.67 ± 0.20	0.90 ± 0.26

Table 2: Human evaluation of multimodal MRI-to-speech

model	Multi Score	Non-multi Score
MLP	**1.67 ± 0.24**	0.33 ± 0.24
LP	**1.33 ± 0.24**	0.67 ± 0.24
HuBERT	**1.67 ± 0.47**	0.33 ± 0.47
TAR ND	**1.17 ± 0.24**	0.83 ± 0.24
NODE	**1.67 ± 0.47**	0.33 ± 0.47

6. RESULTS

6.1. Multi-Modal Pre-Training Results

Table 1 summarizes the automatic speech recognition (ASR) character (CER) and word error rates (WER) for the test set predictions of our MRI-to-speech models. We use Whis-

Table 3: Human evaluation of self-supervised MRI-to-speech

Model	Score
MLP Multi	**10.67 ± 1.25**
LP Multi	9.67 ± 0.47
NODE Multi	7.50 ± 1.08
TAR Multi ND	6.83 ± 0.62
HUBERT Multi	5.33 ± 1.25
Direct	2.00 ± 0.00
Spectrogram	0.00 ± 0.00

Table 4: ASR EMG-VOI-to-Intermediate-to-Speech

model	mean cer	mean wer
MLP ND	**0.13 ± 0.23**	**0.22 ± 0.23**
LP ND	0.14 ± 0.25	0.23 ± 0.24
HuBERT ND	0.16 ± 0.23	0.25 ± 0.24
NODE ND	0.18 ± 0.18	0.29 ± 0.26
TAR ND	0.19 ± 0.19	0.32 ± 0.27
Spectrogram ND	0.30 ± 0.22	0.47 ± 0.29
HuBERT Units ND	0.56 ± 0.57	0.84 ± 0.76
Direct	1.14 ± 1.49	1.45 ± 2.21

Table 5: Human evaluation of EMG-to-Speech

model	Mean mcd
LP ND	**11.33 ± 0.94**
MLP ND	7.67 ± 2.36
NODE ND	7.33 ± 1.25
HuBERT ND	7.33 ± 1.70
TAR ND	6.00 ± 0.82
Spec ND	1.67 ± 0.24
Direct	0.67 ± 0.62

per for ASR [36], and report means and standard deviations across data points. Our top three models all utilize multi-modal pre-training and perform noticeably better than then rest of the models, suggesting that multi-modal pre-training noticeably improves MRI-to-speech performance.

We also perform human evaluation, comparing with and without multi-modal pre-training for each model. 3 listeners participated, each listening to 10 samples, 2 for each model pair. Listeners can select either model or neither for their naturalness preference. For each model, we add 1 to its score if it was selected and 0.5 if it was involved in a neither choice. Table 2 summarizes these results, with means and standard deviations taken across listeners. All multimodal models received higher scores, reinforcing the benefits of multi-modal pre-training.

6.2. Self-Supervised Representation Results

As shown in our MRI-to-Speech ASR results in Table 1, nearly all of our self-supervised approaches outperform the other models. In particular, the top two approaches utilize our proposed modified representations, indicating their usefulness for MRI-to-Speech. We also do human evaluation with 3 listeners, each listening to 42 samples, composed of 2 utterances per pairwise comparison between 7 models. Table 3 summarizes these results and show that our modified representations all noticeably outperform the other approaches, highlighting the suitability of our representations for synthesizing natural speech from MRI. Our EMG-to-speech exhibit matching trends, as shown in our ASR and human evaluation results in Tables 5 and 4.

Contrary to results in other speech synthesis tasks [5, 7], models with discretized units as intermediates performed poorly, much worse than their non-unit counterparts for all self-supervised features. We include HUBERT unit results in our ASR tables for reference, with other models utilizing units performing comparably. We hypothesize this trend is due to the phonemic nature of discretized units [24], in contrast to the articulatory nature of self-supervised representations [37].

7. CONCLUSION

In this work, we devise multi-modal pre-training methods and modified self-supervised features that improve the performance of MRI-to-speech and EMG-to-speech models. With less than 10 minutes of single-speaker training data, our MRI-to-speech model achieves a test-set automatic speech recognition (ASR) character error rate (CER) of 18%, compared to 69% from the previous model [14]. Our EMG-to-speech model similarly noticeably outperforms the baseline, and both ASR results match our human listening tests. In the future, we are interested in extending these results to multi-speaker tasks.

8. REFERENCES

[1] Tomoki Hayashi et al., "Espnet2-tts: Extending the edge of tts research," *arXiv*, 2021.

[2] Yuxuan Wang et al., "Tacotron: Towards end-to-end speech synthesis," *Interspeech*, 2017.

[3] Peter Wu et al., "Understanding the tradeoffs in client-side privacy for downstream speech tasks," in *APSIPA*, 2021.

[4] Loes van Bemmel et al., "Beyond neural-on-neural approaches to speaker gender protection," in *ICASSP*, 2023.

[5] Jiatong Shi et al., "Enhancing speech-to-speech translation with multiple tts targets," in *ICASSP*, 2023.

[6] Ye Jia et al., "Direct speech-to-speech translation with a sequence-to-sequence model," *Interspeech*, 2019.

[7] Sean L Metzger et al., "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, 2023.

[8] B. Denby et al., "Silent speech interfaces," *Speech Communication*, 2010.

[9] Kevin Scheck and Tanja Schultz, "Multi-speaker speech synthesis from electromyographic signals by soft speech unit prediction," in *ICASSP*, 2023.

[10] David Gaddy and Dan Klein, "Digital voicing of silent speech," in *EMNLP*, 2020.

[11] Dan Lim et al., "Jets: Jointly training fastspeech2 and hifi-gan for end to end text to speech," *Interspeech*, 2022.

[12] Peter Wu et al., "Deep speech synthesis from articulatory representations," *Interspeech*, 2022.

[13] Ibrahim Ibrahimov et al., "Data augmentation methods on ultrasound tongue images for articulation-to-speech synthesis," in *SSW*, 2023.

[14] Peter Wu et al., "Deep speech synthesis from mri-based articulatory representations," *Interspeech*, 2023.

[15] Paul K. Krug et al., "Self-Supervised Solution to the Control Problem of Articulatory Synthesis," in *Interspeech*, 2023.

[16] Yashish M Siriwardena and Carol Espy-Wilson, "The secret source: Incorporating source features to improve acoustic-to-articulatory speech inversion," in *ICASSP*, 2023.

[17] Peter Wu et al., "Speaker-independent acoustic-to-articulatory speech inversion," in *ICASSP*, 2023.

[18] Yuto Otani et al., "Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," in *Interspeech*, 2023.

[19] Paul W Schönle et al., "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, 1987.

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

[21] T Baer et al., "Application of mri to the analysis of speech production," *Magnetic resonance imaging*, 1987.

[22] Yongwan Lim et al., "A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images," *Scientific data*, 2021.

[23] Katherine S Harris et al., "Component gestures in the production of oral and nasal labial stops," *JASA*, 1962.

[24] Wei-Ning Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, 2021.

[25] Ashish Vaswani et al., "Attention is all you need," *NeurIPS*, 2017.

[26] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

[27] Jiaqi Su et al., "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Interspeech*, 2017.

[28] Max Morrison et al., "Chunked autoregressive gan for conditional waveform synthesis," *ICLR*, 2021.

[29] Christophe Veaux et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *CSTR*, 2017.

[30] Heiga Zen et al., "Libritts: A corpus derived from librispeech for text-to-speech," *Interspeech*, 2019.

[31] Paul Pu Liang et al., "Multibench: Multiscale benchmarks for multimodal representation learning," *NeurIPS*, 2021.

[32] Zhe Wang et al., "The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition," in *ICASSP*, 2023.

[33] Shu wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021.

[34] Ricky TQ Chen et al., "Neural ordinary differential equations," *NeurIPS*, 2018.

[35] Stephen Merity et al., "Revisiting activation regularization for language rnns," *arXiv*, 2017.

[36] Alec Radford et al., "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[37] Cheol Jun Cho et al., "Evidence of vocal tract articulation in self-supervised learning of speech," in *ICASSP*, 2023.