DEEP ARTICULATORY SEGMENTATION AND SPEECH SYNTHESIS USING RT-MRI

Bohan Yu
Peter Wu
Rishi Jain
Tejas Prabhune
Gopala Krishna Anumanchipalli

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-117
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-117.html

May 17, 2024

DEEP ARTICULATORY SEGMENTATION AND SPEECH SYNTHESIS USING RT-MRI

by Bohan Yu

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Gopala Krishna Anumanchipalli
Research Advisor

05/09/2024

(Date)

* * * * * * *

Professor Steven Conolly
Second Reader

May 16, 2024

(Date)

DEEP ARTICULATORY SEGMENTATION AND SPEECH SYNTHESIS USING RT-MRI

Bohan Yu[1], *Peter Wu*[1], *Rishi Jain*[1], *Tejas Prabhune*[1], *Gopala Anumanchipalli*[1]
[1]University of California, Berkeley,

ABSTRACT

Accurate modeling of the vocal tract is necessary to construct articulatory representations for interpretable speech processing and linguistics. However, vocal tract modeling is challenging because many internal articulators are occluded from external motion capture technologies. Real-time magnetic resonance imaging (RT-MRI) allows measuring precise movements of internal articulators during speech, but annotated datasets of MRI are limited in size due to time-consuming and computationally expensive labeling methods. We first present a deep labeling strategy for the RT-MRI video using a vision-only segmentation approach. We then introduce a multimodal algorithm using audio to improve segmentation of vocal articulators. Lastly, we propose three transfer learning techniques to noticeably improve MRI-based articulatory-to-acoustic synthesis performance and extend the technique to EMG and EMA-to-speech tasks: (1) pre-trained weight initialization, (2) pre-training part of the model, and (3) multimodal pre-training. We also release labels for a 75-speaker RT-MRI dataset, increasing the amount of labeled public RT-MRI data of the vocal tract by over a factor of 9. The code and dataset labels can be found at demo page.

Index Terms— articulatory synthesis, audio-visual perception

1. INTRODUCTION

Vocal tract modeling is an essential technology in many applications including facial animation, naturalistic speaking avatars, speaker modeling, and second language pronunciation learning [1, 2, 3, 4, 5, 6]. In fact, popular self-supervised speech representations learn features correlated with articulators [7]. Modeling is also necessary in healthcare applications such as brain-computer interfaces for communication [8, 4] and treating speech disfluencies [9, 10]. Methods of external motion capture cannot record precise and accurate vocal tract movements for occluded articulators. Thus, the inner mouth is often poorly represented or neglected in multimedia approaches to motion capture-based facial animation [11]. 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 [12, 13].

A more comprehensive approach uses Real-Time Magnetic Resonance Imaging (RT-MRI) of the vocal tract [14]. This technology offers audio-aligned videos of internal and external articulators that are not measurable by other articulatory representations. When tested on downstream speech-related tasks, RT-MRI has been shown to more reliably and completely model the vocal tract in comparison to EMA [15]. For example, MRI representations distinguish between oral vowels (lowered velum) and nasal vowels (raised velum), while EMA does not track the velum at all. However, current state-of-the-art labeling methods for extracting interpretable features from these videos are time-consuming, computationally expensive, and prone to errors [16]. Therefore, only a small amount of vocal tract RT-MRI data is labeled [17] and existing MRI-to-speech synthesis models have low intelligibility [15]. As a result, current work using real-time articulatory MRI falls into two broad categories: (1) those which rely on the previously extracted articulator segmentations [15, 9], or (2) models which directly work with RT-MRI videos but do not contain an interpretable intermediate representation [18, 19]. To address the scarcity of publicly-available articulatory segmentations for RT-MRI and improve the fidelity of MRI-to-speech synthesis, we propose:

- A vision-based fully-convolutional neural network [20] for speaker-independent vocal tract boundary segmentation.
- A multimodal Transformer model which additionally includes the speech waveform to set a new benchmark for vocal tract RT-MRI segmentation.
- Labels for the 75-speaker Speech MRI Open Dataset [21] containing over 20 hours of vocal tract RT-MRI data for 75 speakers diverse in age, gender, and accent.
- Three transfer learning approaches that noticeably improve articulatory-to-acoustic synthesis performance in error-prune settings
- A deep speech representation that outperforms self-supervised learning features and spectrums as an intermediate for articulatory synthesis.

2. ARTICULATORY DATASETS

2.1. USC-TIMIT Dataset

The USC-TIMIT dataset contains labeled 8-speaker RT-MRI of the vocal tract described in [17]. Subjects were instructed to read phonetically-diverse sentences out loud at a natural speaking rate while laying supine in an MRI scanner. A four-channel upper airway receiver coil array was used for signal reception, which was processed to reproduce 84×84 pixel midsaggital MRI videos capturing lingual, labial, and jaw motion, and velum, pharynx, and larynx articulations. These videos are collected at 83.33 Hz. We start with the 170 representative points from [17] to represent vocal tract air-tissue boundary segmentations. 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 Wu *et al.* [15]. All RT-MRI video in the USC-TIMIT dataset is accompanied by existing articulator points extracted using the baseline algorithm described further in Section 3.1. We use these point labels as training targets for the other segmentation methods described in Section 3. Paired with these trajectories is the 16kHz speech data (resampled from original 20kHz) corresponding to the read sentence during any RT-MRI scan. Following previous articulatory MRI work, we further enhance this audio using Adobe Podcast to reduce reverbation [15]. For training segmentation models, we use 7 of the 8 speakers (roughly 66 minutes of RT-MRI video) and leave out the remaining speaker "Napa" as "unseen". To compare with prior works [15], we train MRI-to-speech synthesis model using the ground truth "Napa" speaker data with 155 data points.

2.2. Speech MRI Open Dataset

The Speech MRI Open Dataset [21] is a diverse 75-speaker dataset that provides 20 hours of raw multi-coil RT-MRI videos of the vocal tract during articulation, aligned with corresponding speech. 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.

2.3. EMA Dataset

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 [22, 23]. We use MNGU0, a single-speaker dataset containing 67 minutes of 16 kHz speech and 200 Hz EMA [24]. Another dataset we use is the Haskins Production Rate Comparison database (HPRC), an 8-speaker dataset containing 7.9 hours of 44.1 kHz speech and 100 Hz EMA [25]. To maintain consistency with prior work [23, 26], 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 4.3. For all of our EMA data, we concatenate the 6 x-y coordinates to form a 12-dimensional vector at each time step.

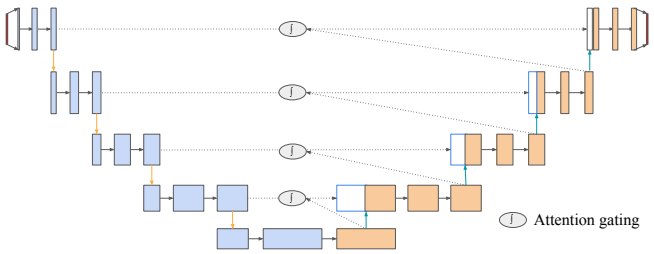

Fig. 1: The attention U-Net model. Dotted lines represent the paths of attention gating in contracting/expanding layers.

2.4. Electromyography (EMG)

Surface electromyography (EMG) measures electrical potentials caused by nearby muscle activity using electrodes placed on top of the skin [27]. When placed near articulators, EMG provides another low-dimensional manifold of articulatory movements [27, 28, 29, 30]. In this work, we use the EMG dataset in [29], which consists of EMG data and speech for vocalized utterances. We use the 3.9-hour vocalized speech subset, denoted "Parallel Vocalized Speech" in [29]. Our train-dev-test data split contains 195 minutes, 12 minutes, and 23 minutes of speech, respectively. Speech waveforms have a sampling rate of 16 kHz, and EMG 1000 Hz.

3. TRAINING SEGMENTATION MODELS

3.1. Frequency-domain Gradient Descent Baseline

The existing algorithm for articulatory RT-MRI segmentation [17] relies on hand-traced air-tissue boundaries for the first frame of every video. It subsequently performs nonlinear optimization in the frequency space of subsequent frames, requiring 20 minutes to converge for a single frame using gradient descent. This procedure is also prone to mislabeling and requires human supervision, making it expensive to run. Because each frame is optimized independently, it often results in jitter, or high-frequency perturbations, for individual articulator points across consecutive frames. As this is the only existing algorithm for articulatory RT-MRI labeling, the outputs of this model are used as the "ground truth" training targets for the following models, and the algorithm will be referred to as the "baseline" algorithm.

3.2. Heatmap U-Net

The U-Net [20], a residual fully-convolutional neural network, has historically performed well on low resolution

medical images, especially when training data is limited. Because labeled data was only originally available from eight speakers, this architecture provided the best fit while also generalizing to held out speakers. Input MRI frames were 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 expanding. Of the spatial features, the key articulators only occupy a subset of the space. For this reason, we learned a spatial weighting map on the residual connection to effectively suppress the components of the signal which are not important using an attention gating mechanism, and introduced normalization layers similar to the Attention U-Net [31] with the modification of using additive attention as opposed to multiplicative. Adding attention gating minimally increased model complexity. The architecture is visualized in Figure 1.

We trained this model on approximately 90 minutes of labeled midsaggital RT-MRI video from 7 speakers for a total of 6 epochs. The model outputs a 96 by 96 grid for each of the 95 articulatory points. Each of the target keypoints were modeled as 2-dimensional isotropic Gaussian distributions over the 96 by 96 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. During training, we also applied random affine transformations to frames and the corresponding annotations to promote generalization to unseen speakers.

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 distribution using a softmax nonlinearity. To our knowledge, this training objective has not been used for heatmap regression in medical imaging 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, articulator points have varying degrees of movement (standard deviation) and importance in speech production. To explore this, we also experimented with using the standard deviation and importance determined by Wu *et al.* [15] from the 7 training speakers in the training objective. Specifically, we multiplied the standard deviation and importance of each point to determine its weighting in the combined loss. This articulatory weighting emphasizes the importance of points that show significant movement and are important to speech production over those which show minimal movement or have been found to be less essential.

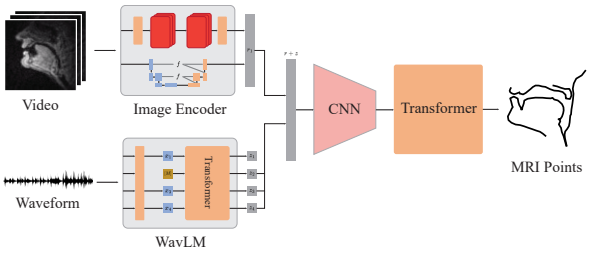

Fig. 2: Architecture of the multimodal segmentation model.

3.3. Multimodal Audio-Visual Transformer

Using the U-Net model as a pretrained convolutional input, we further explored joint point tracking methods. To ensure tracks remain smooth, we applied a temporal Gaussian low-pass filter independently for each point of the U-Net output. We also tried using a convolutional LSTM as in [32] (CLSTM) and a Transformer. The CLSTM, previously used in MRI video segmentation [18], applies a 2-layer LSTM to the predicted U-Net outputs, trained on speech from the same 7 USC-TIMIT speakers. The Transformer similarly used the U-Net points from each timestep, with an additional positional encoding. Additionally, we experimented with adding optical flow, Kalman filtering, and Lucas Kanade to improve temporal point tracking [33, 34]. 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-only tracking problems.

We subsequently experimented with multimodal models for feature extraction, using representations from video frames and speech waveforms. For video frames, we used the output of the frozen U-Net model described in Section 3.2 and also experimented with other image representation models including ResNet [35] and ConvNeXt [36]. To represent audio, we used the 10th layer of WavLM [37] to derive speech representations. The two representations were then concatenated as input to a Transformer prepended with three residual convolutional blocks as seen in Figure 2. Additionally, we experimented with an audio-only segmentation model (articulatory inversion) using the same WavLM and Transformer methods. The Transformer models were trained on the speech data from the same 7 of 8 USC-TIMIT speakers as in Section 3.2. Using multi-task learning, the Transformer experiments output MRI trajectories and pitch simultaneously, optimized using weighted L1 loss.

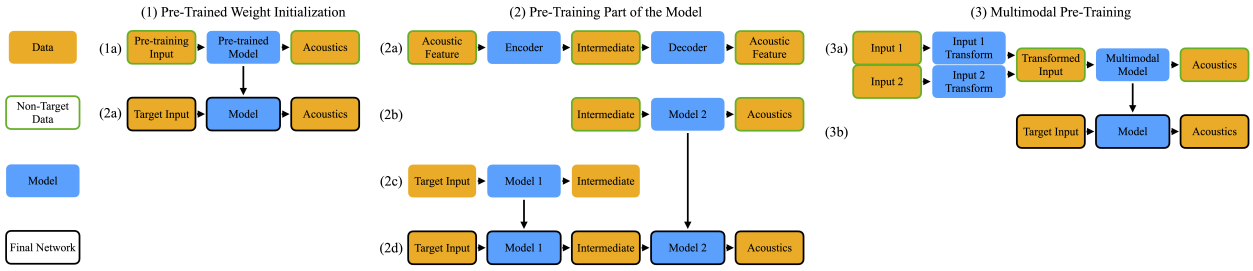

Fig. 3: Three transfer learning approaches for articulatory synthesis.

4. TRANSFER LEARNING

4.1. Pre-Trained Weight Initialization

Initializing model weights with those of a pre-trained model is an effective method to improve fine-tuning performance in limited-data settings [15], visualized in Figure 3 approach (1). We demonstrate that this method can improve intelligibility by 5% absolute WER compared to prior EMA-to-speech models, measured with an automatic speech recognizer. Moreover, this approach noticeably improves data efficiency, with details in Section 6.1.

4.2. Pre-Training Part of the Model

Pre-training part of the model is another effective method for improving performance in low-resource settings. For example, many text-to-speech (TTS) models pre-train their vocoder [38], and many classifiers pre-train their encoder [39]. Popular vocoder input representations include spectrums, high-dimensional self-supervised features, learnt representations, and units [38, 40, 41, 42]. This pre-training method has also shown success with ultrasound-speech tasks [43]. We extend these results to MRI and EMG datasets that contain significantly less and noisier data. Additionally, we propose a vocoder input dimensionality reduction approach that noticeably improves MRI- and EMG-to-speech performance.

Specifically, we reduce the dimensionality of the HuBERT [44] self-supervised representation in order to reduce the complexity of mapping to this intermediate feature, visualized as (2a) in Figure 3. We choose HuBERT given its success with other synthesis tasks [45, 46], and note our dimensionality reduction methodology can be applied to any representation. We experiment with three methods: (1) linear projection, (2) low-pass filtering, and (3) neural ordinary differential equations (ODE) [47]. Intuitively, methods 2 and 3 encourage the resulting feature to be smoother across time than the original feature. All three approaches linearly project HuBERT from 1024 to 256 dimensions. Our second method adds a differentiable low-pass filter along the time dimension with an arbitarily chosen cutoff frequency of 0.4 after the linear layer[1]. 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 mapped and original frames. We use a linear layer as our ODE function. This encourages each next frame to equal the output of iteratively applying a fixed linear transformation to the current frame, reducing the complexity of the representation space. Our three approaches are denoted as **MLP**, **Low-Pass**, and **NODE**, respectively, in the result section below.

To train each of these three representations, we linearly project the 256-dimensional vector outputs back to 1024 dimensions and compute an MSE loss between this final output and the ground truth HuBERT features (step 2a in Figure 3). 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. Then, we train an intermediate-to-acoustic HiFi-CAR (Section 3), visualized as step (2b) in Figure 3. Thirdly, in step (2c), we train an articulatory-to-intermediate Transformer (Section 3), Finally, we prepend this model to HiFi-CAR to form our articulatory-to-intermediate-to-acoustic model (step 2d). Steps (2a) and (2b) do not require articulatory data, allowing us to train these steps on a large speech corpus. Since HuBERT accepts 16 kHz speech as input, we downsample waveforms to match this sampling rate. We find pre-training part of the model to noticeably improve speech synthesis quality for MRI-to-Speech and voiced EMG-to-Speech tasks, detailed in Section 6.2.

4.3. Multimodal Articulatory 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 fewer modalities [48, 49]. We extend this strategy to articulatory synthesis by pre-training with more than one articulatory modality as input and fine-tuning the resulting model with only the target articulatory modality, visualized in Figure 3 approach (3).

Specifically, we pre-train our MRI-to-speech model with both EMA and MRI, where EMA is inferred from the ground truth speech data using a fixed speech-to-EMA model (Wu et

[1] https://github.com/adefossez/julius

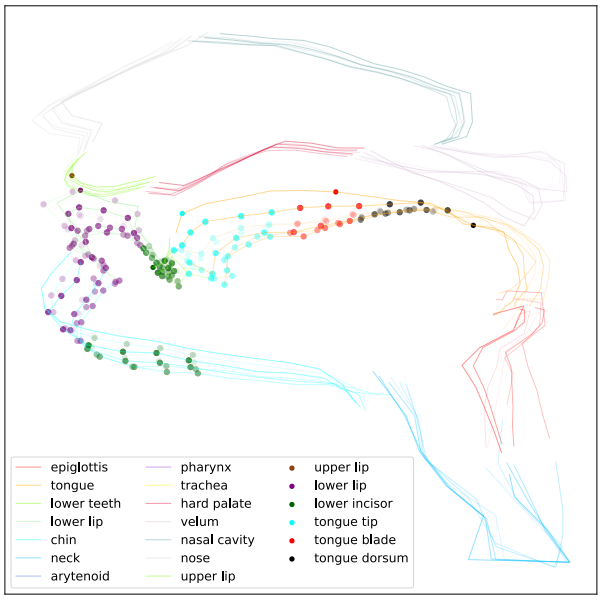

Fig. 4: Extracted MRI-features for the utterance "apa." Lighter is earlier in time. Each point is colored with the highest-correlation EMA feature. Points with maximum correlation magnitude below 0.3 are omitted for readability.

al., 2023) [23]. We linearly interpolate the estimated EMA to match the sampling rate of the MRI data. 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 fine-tune the resulting model on the target modality dataset with the same hyperparameters. Models utilizing multi-modal pre-training contain "Multi" in the tables below, and detailed optimization choices and results are in Section 6.3.

To provide more intuition on multimodal pre-training, Figure 4 illustrates the average Pearson correlation between inferred EMA and ground-truth MRI. We visualized correlation by coloring each MRI point in the midsagittal plane with the highest-correlation EMA point, where MRI points with maximum correlation magnitude below 0.3 are omitted for readability. The noticeable overlap between these modalities spatially suggests that information learned from one modality can be transferred to the other.

5. MRI SEGMENTATION RESULTS

We performed quantitative evaluations of both our vision-based and multimodal vocal tract segmentation approaches. The segmentations were then used to add articulatory labels to RT-MRI from 75 previously-unlabeled speakers. Using this data as a multimodal pretraining approach, the different segmentations were further used for a downstream speech task to measure how well speech features were captured by different segmentation methods. Finally, we conducted a qualitative hypothesis test using our best method.

Table 1: Comparison of the root mean squared error of the U-Net models trained using L2 loss, KL-divergence loss, and KL-divergence loss with articulatory weighting. More details are available in Section 5.1.

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

5.1. Vision-only U-Net

The first experiment compared L2 (mean squared error) loss with our new pixel-wise KL-divergence loss with and without articulatory weighting for the U-Net model. This was evaluated using the root mean squared error (RMSE) of the predicted x-y points for the 95 articulator points on an unseen speaker. The results in Table 1 demonstrate that the KL-divergence loss is better suited for low-resolution point recognition for air-tissue boundary segmentation. While articulatory weighting predictably increases the RMSE, manual inspection reveals that most of this error can be attributed to slight shifts in less phonologically important articulators such as the hard palate, with significant improvement on the more important articulators.

5.2. Comparison with Multimodal Transformer

When analyzing our various feature extraction methods, we first evaluate performance within the context of seen speakers but unseen examples. Figure 5 highlights quantitative results in L1 losses and Pearson Correlation Coefficients (PCCs) when evaluating models on unseen examples from seen speakers. We observe that 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 for better in-domain modeling of speech production. With only a single modality, the pixel value-based U-Net generalizes better to unseen speakers than the WavLM-based speech inversion model since contour pixel values capture speaker-specific anatomy better than speech waveforms alone. Utilizing this tradeoff, we use the U-Net model to label the unseen 75-speaker Speech MRI Open Dataset and we verify in section 6.4 whether using WavLM based segmentation approach will

benefit single-speaker MRI-to-speech synthesis.

5.3. Labeling the Speech MRI Open Dataset

We used the U-Net model to label RT-MRI video for 75 speakers in the Speech MRI Open Dataset [21]. Outputs from this model were subsequently run through a temporal Gaussian low-pass filter, which was applied independently for each articulator x-y point and used to provide video and audio-aligned MRI trajectories.

In Figure 6, we highlight the generalization of the U-Net model on unseen speakers, allowing us to expand the amount of labeled RT-MRI video to over 20 hours across 83 total speakers. Qualitatively, the predicted segmentations closely follow the MRI segments, achieving high quality labeling for unseen speakers. As part of this paper, we also present this labeling for use in future downstream speech tasks, increasing the amount of publically-available labeled articulatory RT-MRI data by over a factor of **9**. The labels are available at [*Add google drive link*].

5.4. Qualitative Evaluation

Despite relying on the output of the baseline segmentation algorithm as the training targets, our segmentation methods performed better than the baseline algorithm when evaluated on downstream speech synthesis. We hypothsize that this is because the baseline segmentations have high amounts of jitter and inconsistencies across frames, and are sometimes even physiologically implausible. In comparison, the estimates of the presented multimodal approach are much more consistent and plausible, possibly explaining why they are better suited for building downstream methods. To validate this hypothesis with a subjective evaluation, we ran a one-tailed perceptual test for statistical significance where participants looked at two video animations of vocal tract movements in side-by-side panels (one with original labels, and the other with outputs of our segmentation method). The participants then selected which rendering is a more accurate representation of the associated audio. Our results reveal the participants (n=21) prefer the outputs of our algorithm relative to the original segmentations ($p < 0.001$).

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

6. ARTICULATORY SYNTHESIS RESULTS

For all HiFi-CAR experiments, we trained this model with an autoregressive feature extractor hidden dimension of 256, a batch size of 32, and the Adam optimizer with $\{0.5, 0.9\}$ for beta values [50]. Transformer layers have a hidden dimension of 1024 and a dropout of 0.2. We trained the Transformer using the L1 loss function, the Adam optimizer [50] with betas $\{0.5, 0.9\}$, and a batch size of 16. During training, we randomly crop a 0.5 seconds to 2 seconds window from each sample in the batch, with the window length fixed within the batch. Since EMA datasets have much less noise than other articulatory modalities [15, 29], for EMA tasks, we do not do multimodal pre-training and find pre-training part of the model unnecessary. For MRI and EMG tasks, we use all three transfer learning methods, with pre-trained weight intialization applied to the baseline and intermediate-to-acoustic models.

Table 2: EMA-to-speech ASR results with and without pre-trained weight initialization on 5-minute and entire training set, with 95% confidence intervals in parentheses.

Model	5 Min. WER (%) ↓	All WER (%) ↓
No Pre-Train	22.6 (13.8-33.1)	9.4 (4.9-14.3)
Pre-Train	**17.7 (11.1-24.5)**	9.3 (4.9-14.6)

6.1. Pre-Trained Weight Initialization Results

To check the usefulness of pre-trained weight initialization, we train EMA-to-speech models with and without such initialization on MNGU0, described in Section 2.3. Our EMA-to-speech model here is HiFi-CAR, described in Section 3, with upsample scales [5, 4, 2, 2] to upsample the 200 Hz EMA input to the 16000 Hz waveform. For pre-trained weights, we use the LibriTTS [51] HiFi-GAN mel-spectrogram to speech vocoder weights in [52, 15]. Since these scales are different than those of the pre-trained vocoder, we only load the weights with matching dimensions. In addition to the 12-dimensional EMA data, we concatenate loudness and pitch to the input, each one-dimensional, forming a 14-dimensional vector input at each time step. Inspired by [53], We compute pitch using CREPE [54, 55] and loudness by taking the absolute maximum of an 80-frame window, both using the EMA data sampling rate and a hop size of 80. For our train-validation-set split, we match the 1069-60-60 utterance split in [56]. We also train only on a 5-minute subset randomly sampled from the train set in order to study data efficiency. To evaluate these EMA-to-speech synthesizers, we compute WER with the Whisper Large automatic speech recognition (ASR) model [57], with WER results in Table 2. WER using the entire train set is comparable betwen models, suggesting that pre-trained weight initialization yields at least as good performance compared to the default initializaiton. Notably, when training on only 5 minutes of data, the model with pre-trained weight initialization performed much better than the other one, suggesting that this initialization method improves data efficiency.

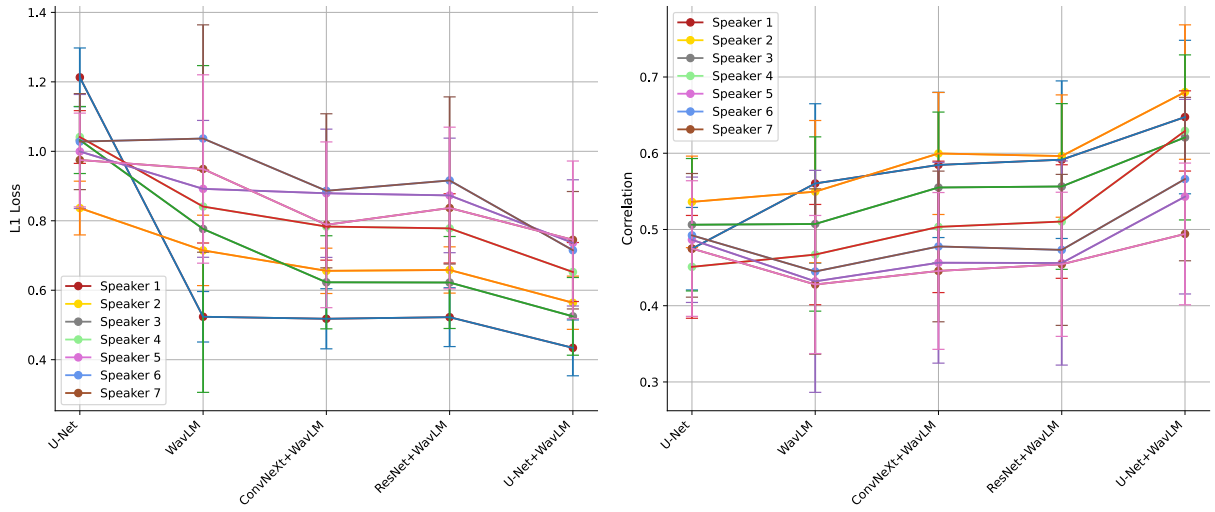

Fig. 5: L1 losses [↓] (*left*) and Pearson Correlation Coefficients (PCCs) [↑] (*right*) comparing MRI trajectories of unseen examples from seen speakers of a given model with the USC-TIMIT ground truth. Varying through a subset of five representative models.

Table 3: ASR character and word error rates on MRI-to-speech synthesis outputs, with 95% confidence intervals in parentheses. Proposed intermediates in top 3 rows (Section 4.2).

Model	CER (%) ↓	WER (%) ↓
Low-Pass	**28.2 (19.4-37.4)**	**42.4 (30.1-55.9)**
MLP	36.0 (22.7-49.5)	57.2 (36.7-78.4)
NODE	43.8 (25.-66.2)	62.0 (37.8-88.2)
HuBERT	31.1 (21.9-41.8)	53.2 (36.4-72.5)
Spectrogram	42.7 (33.3-52.5)	65.7 (52.2-80.3)
Direct	66.7 (55.4-74.3)	89.5 (74.4-100.0)

Table 4: Human evaluation scores for MRI-to-speech (mean ± standard deviation, $\in [0, 1]$). Proposed intermediates in top 3 rows (Section 4.2).

Model	MRI Score ↑	EMG Score ↑
Low-Pass	0.81 ± 0.04	**0.94 ± 0.08**
MLP	**0.89 ± 0.10**	0.64 ± 0.20
NODE	0.63 ± 0.09	0.61 ± 0.10
HuBERT	0.44 ± 0.10	0.61 ± 0.14
Spectrogram	0.00 ± 0.00	0.14 ± 0.02
Direct	0.17 ± 0.00	0.06 ± 0.05

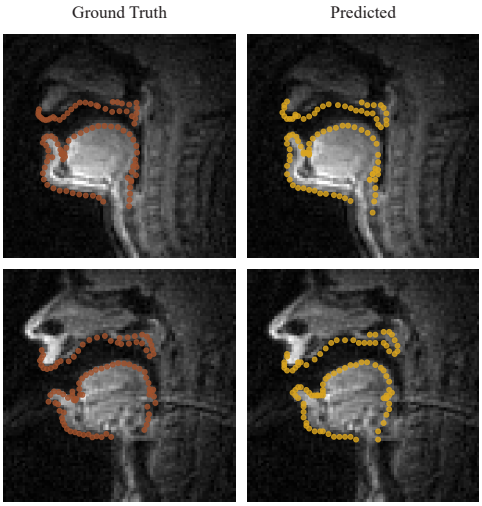

Fig. 6: Two representative examples of predicted MRI points (*right*) compared to expert hand labels (*left*). The examples are spoken by unseen Female (*bottom*) and Male (*top*) speakers in the Speech MRI Open Dataset.

6.2. Results when Pre-Training Part of the Model

We pre-train part of the model as in Section 4.2 for single-speaker MRI-to-speech and voiced EMG-to-speech tasks, with datasets described in Sections 2.1 and 2.4, respectively. Our 256-dimensional intermediate features are learnt with VCTK, which has 110 English speakers and a total of 44 hours of 44.1 kHz speech, randomly dividing speakers into

Table 5: ASR character and word error rates on voiced EMG-to-speech synthesis outputs, with 95% confidence intervals in parentheses. Proposed intermediates in top 3 rows (Section 4.2).

Model	CER (%) ↓	WER (%) ↓
Low-Pass	14.2 (10.8-18.6)	23.1 (19.5-26.8)
MLP	**13.2 (10.2-16.9)**	**22.2 (18.8-25.8)**
NODE	17.6 (15.0-20.3)	29.1 (25.4-33.3)
HuBERT	15.7 (12.5-19.7)	24.6 (20.8-28.5)
Spectrogram	30.2 (26.9-33.5)	47.3 (42.4-51.8)
Direct	113.8 (100.3-129.2)	145.1 (124.5-167.7)

an 85%-5%-10% train-validation-test split [58].

Our baseline for MRI-to-speech is [15], labeled Direct in Tables 3 and 4. Specifically, this is the HiFi-CAR model described in Section 3 with upsample scales [8, 5, 3, 2] to map $83.\bar{3}$ Hz MRI to 20 kHz acoustics. Since our voiced EMG task does not have a baseline to our knowledge [29], we also use HiFi-CAR, here with upsample scales [2, 2, 2, 2] to map 1 kHz EMG to 16 kHz acoustics. This baseline is labeled Direct in Tables 5 and 4. For partially pre-trained models, we map inputs to intermediates (Section 4.2) using the Transformer in Section 3, and intermediates to waveforms using HiFi-CARs with the same architectures as the baselines. We linearly interpolate the 50 Hz intermediate features to match the sampling rates of the inputs.

To evaluate these models, we use the ASR metric in Section 6.1 and human evaluation. As shown in Tables 3 and 5, pre-training part of the model results in much better ASR performance than the baseline for both MRI-to-speech and voiced EMG-to-speech. Also, our low-pass-filtered representation (Low-Pass) described in Section 4.2 outperforms HuBERT on both tasks. We also do human evaluation with 3 listeners, each listening to 30 samples, composed of 2 utterances per pairwise comparison between 6 models. For each pair of utterances, if one is preferred, that model receives a score of 1 and the other model 0, and otherwise both receive 0.5. Scores are averaged per model, so that each score is in [0, 1], with 1 being the highest possible score. Table 4 summarizes these results for MRI-to-speech and EMG-to-speech. All of our proposed 256-dimensional features noticeably outperform the other methods, highlighting the suitability of these features for synthesizing natural speech.

6.3. Multi-Modal Pre-Training Results

As motivated in Section 4.3, we apply our multi-modal pre-training method to single-speaker MRI-to-speech synthesis. Our MRI dataset and model architectures are the same as those in Section 6.2, with the model being modified during the multi-modal pre-training step as described in Section 4.3.

Table 6: ASR word error rates on multimodal and non-multimodal MRI-to-speech synthesis outputs, with 95% confidence intervals in parentheses. Low-Pass is a proposed intermediate (Section 4.2).

Model	Multi. WER (%) ↓	Non-multi. WER (%) ↓
Low-Pass	**33.3 (19.0-52.0)**	42.4 (30.1-55.9)
HuBERT	**34.4 (19.8-52.9)**	53.2 (36.4-72.5)

Table 7: Human evaluation scores for multimodal versus non-multimodal MRI-to-speech (mean ± standard deviation, $\in [0, 1]$). Low-Pass is a proposed intermediate (Section 4.2).

Model	Multi. Score ↑	Non-multi. Score ↑
Low-Pass	**0.714 ± 0.12**	0.34 ± 0.12
HuBERT	**0.84 ± 0.24**	0.17 ± 0.24

We pretrain our model with: (1) all of the EMA data in the HPRC dataset described in Section 2.3, and (2) the training set of our MRI dataset described in Section 2.1. The pre-training and fine-tuning steps both use the Adam optimizer with a learning rate of 10^{-4} [50]. To avoid redundancy, we report results for our best proposed representation (Low-Pass) and HuBERT. We observe similar results for all of the other models, with details and code being available in the supplementary codebase post-anonymity.

We evaluate these models with the same ASR metric as Section 6.2. Table 6 summarizes the ASR WER and character errror rates (CER) on the MRI test set. The models utilizing multi-modal pre-training all outperform their non-multi-modal counterparts, suggesting that multi-modal pre-training noticeably improves MRI-to-speech performance. We note that our best WER, 33%, is noticeably better than the 90% WER from the previous model [15]. We also perform a preliminary human evaluation study, 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. Like Section 6.2, we average scores for each model to give a number between 0 and 1, with 1 being the best possible score. Table 7 summarizes these results, with means and standard deviations taken across listeners. Matching the ASR result, the multimodal models received higher scores, reinforcing the benefits of multi-modal pre-training.

6.4. Synthesis Comparison across MRI segmentations

To evaluate our segmentation methods on speech synthesis tasks, we use the same Low-Pass feature and multimodal articulatory pretraining approach to train separate models for

Table 8: Speech synthesis ASR WER finetuning on segmentations from a seen speaker during segmentation model training, but unseen utterances. (S) denotes synthesis model pretrained using single MRI speaker. All other models are pretrained with 75-speaker MRI.

Model	WER
U-Net + WavLM	**0.313 (0.164-0.493)**
U-Net	0.364 (0.209-0.551)
Ground Truth	0.347 (0.186-0.532)
U-Net + WavLM (S)	0.349 (0.203-0.528)

Table 9: Speech synthesis ASR WER finetuning on segmentations from an unseen speaker during segmentation model training. (S) denotes synthesis model pretrained using single MRI speaker. All other models are pretrained with 75-speaker MRI.

Model	WER
U-Net + WavLM	**0.333 (0.202-0.498)**
U-Net	0.352 (0.172-0.568)
Ground Truth	0.497 (0.348-0.666)
U-Net + WavLM (S)	0.501 (0.280-0.728)

each segmentation model on one seen speaker and one unseen speaker. We also want to explore the effect of using the labeled 75 speaker MRI data during multimodal pretraining. For models pretrained with single speaker MRI data, we denote them as (S). For example, U-Net + WavLM(S) means the model is pretrained on the Napa data labeled by the U-Net + WavLM segmentation approach and then finetuned, following the approach in section 4.3. All other models without (S) are pretrained with 75 speaker MRI data labeled by the UNet.

For seen speakers of segmentation models, the multimodal U-Net + 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 8 summarizes these results.

The results in Table 9 highlight that the U-Net + WavLM based model has the lowest WER when testing on an unseen USC-TIMIT speaker "Napa", documenting that the segmentations from the multimodal model on unseen speakers still capture representative articulatory kinematics for naturalistic speech. We note that the word-error-rate for unseen speaker "Napa" using ground truth label is much worse than that of Table 6. This is because during multimodal pretraining, the MRI data we used comes from U-Net labels for 75 speakers with only 95 MRI points, posing difficulties for the model to be finetuned on the 155-point ground truth MRI data.

Table 9 also suggests that when using "U-Net + WavLM" segmentation outputs on unseen speakers, pretraining the synthesis model on 75-speaker data can achieve the level of intelligibility compared to results in Table 6 while pretraining on single speaker can't. When using "U-Net + WavLM" on seen speakers as shown in Table 8, pretraining the synthesis model on single speaker has comparable performance compared to pretraining on 75 speakers. This demonstrates (1) potential degradation of speech production information when applying the WavLM-based segmentation model on out-of-domain speakers (2) synthesis knowledge from the U-Net labeled 75-speaker Speech MRI Open Dataset can mitigate the effect of degraded vocal tract segmentations on speech synthesis. Therefore, we expect the new labels for the Speech-MRI-Open-Dataset to be beneficial for future articulatory-speech-related tasks as well.

7. CONCLUSION

In this work, we improve the accuracy of vocal-tract segmentation from RT-MRI images through a vision-only U-Net approach and used the model to label the 75-speaker Speech-MRI-Open-Dataset, increasing the amount of public labeled RT-MRI data of the vocal tract by over a factor of 9. We further improve the in-domain segmentation accuracy with a bimodal audio-vision approach combining the U-Net and WavLM. We then propose the first intelligible MRI-speech-synthesis system by utilizing regularized HuBERT and transfer learning, with extensions to EMA and EMG-to-speech synthesis. Using the developed speech synthesis system, we demonstrate that through inclusion of speech information, the segmentation output can benefit downstream MRI speech synthesis tasks. We also illustrate that pretraining the MRI speech synthesis model on the U-Net labeled 75-speaker MRI dataset can improve intelligibility. We expect future works to discover more potential usages with the dataset and improve the proposed articulatory speech synthesis systems, and we list a few below: (1) Multispeaker Speech-to-MRI inversion systems (2) Multispeaker MRI-to-Speech synthesis (3) MRI speech encodec (4) Streaming articulatory inversion and synthesis systems (5) Silent-speech synthesis (6) Transfer learning for other related tasks.

8. 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," in *Phonology and 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, pp. 1037–1046, 2023.

[5] Urvish Desai, Chiranjeevi Yarra, and Prasanta Ghosh, "Concatenative articulatory video synthesis using real-time mri data for spoken language training," in *ICASSP*, 04 2018, pp. 4999–5003.

[6] S Chandana, Chiranjeevi Yarra, Ritu Aggarwal, Sanjeev Kumar Mittal, NK Kausthubha, KT Raseena, Astha Singh, and Prasanta Kumar Ghosh, "Automatic visual augmentation for concatenation based synthesized articulatory videos from real-time mri data for spoken language training," in *Proc. Interspeech*, 2018.

[7] 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.

[8] Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019.

[9] 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.

[10] 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)*, January 2021, pp. 41–50.

[11] 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.

[12] Joseph S Perkell, Marc H Cohen, Mario A Svirsky, Melanie L Matthies, Iñaki Garabieta, and Michel TT Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3078–3096, 1992.

[13] 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, Eds., Online, Aug. 2021, pp. 175–181, Association for Computational Linguistics.

[14] Vikram Ramanarayanan, Louis Goldstein, Dani Byrd, and Shrikanth S Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, 2013.

[15] 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.

[16] 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.

[17] 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, no. 3, pp. 1307–1311, Sept. 2014.

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

[19] 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.

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

[21] 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*, vol. 8, no. 1, jul 2021.

[22] 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.

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

[24] Korin Richmond, Phil Hoole, and Simon King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Interspeech*, 08 2011, pp. 1505–1508.

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

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

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

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

[29] 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.

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

[31] 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.

[32] 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)*, 2020, pp. 7354–7358.

[33] Frank Loewenich and Frederic Maire, "A head-tracker based on the lucas-kanade optical flow algorithm," in *Proceedings of the 2006 Conference on Advances in Intelligent IT: Active Media Technology 2006*, NLD, 2006, p. 25–30, IOS Press.

[34] 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)*, 2019, pp. 3036–3041.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," 2015.

[36] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," 2022.

[37] 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.

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

[39] Robin Netzorg et al., "Towards an interpretable representation of speaker identity via perceptual voice qualities," *ICASSP*, 2024.

[40] Li-Wei Chen, Shinji Watanabe, and Alexander Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 12644–12652.

[41] Minchan Kim, Myeonghun Jeong, Byoung Jin Choi, Dongjune Lee, and Nam Kim, "Transduce and speak: Neural transducer for text-to-speech with semantic token prediction," in *ASRU*, 12 2023, pp. 1–7.

[42] Zeqian Ju et al., "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *arXiv preprint arXiv:2403.03100*, 2024.

[43] Jing-Xuan Zhang, Korin Richmond, Zhen-Hua Ling, and Lirong Dai, "Talnet: Voice reconstruction from tongue and lip articulation with transfer learning from text-to-speech synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 14402–14410.

[44] 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.*, vol. 29, pp. 3451–3460, oct 2021.

[45] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Proc. Interspeech 2021*, 2021.

[46] Rongjie Huang, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, Jinzheng He, and Zhou Zhao, "Transpeech: Speech-to-speech translation with bilateral perturbation," *ICLR*, 2023.

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

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

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

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

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

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

[53] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "Ddsp: Differentiable digital signal processing," *ICLR*, 2020.

[54] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[55] Max Morrison, Caedon Hsieh, Nathan Pruyne, and Bryan Pardo, "Cross-domain neural pitch and periodicity estimation," *arXiv preprint arXiv:2301.12258*, 2023.

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

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

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