Multimodal Segmentation for Vocal Tract Modeling

Rishi Jain[*]*, Bohan Yu*[*]*, Peter Wu, Tejas Prabhune, Gopala Anumanchipalli*

University of California, Berkeley, United States

rishiraij@berkeley.edu, ybhtim@berkeley.edu

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. Together, we set a new benchmark for vocal tract modeling in MRI video segmentation and use this to 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 rishiraij.github.io/multimodal-mri-avatar/.

Index Terms: articulatory speech, 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 inherently learn features correlated with articulators [7]. Modeling is also necessary in healthcare applications such as brain-computer interfaces for communication [4, 8] 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]. As a result, current work using real-time articulatory MRI falls into two broad categories: (1) methods 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, 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.

2. Datasets

2.1. USC-TIMIT Dataset

We use the labeled 8-speaker RT-MRI USC-TIMIT dataset of the vocal tract described in [17] for training. 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 spoken audio during the RT-MRI scan. Following previous articulatory MRI work, we further enhance this audio using Adobe Podcast to reduce reverbation [15]. For training, we use 7 of the 8 speakers (roughly 66 minutes of RT-MRI video) and leave out the remaining speaker as "unseen".

[*]These authors contributed equally to this work

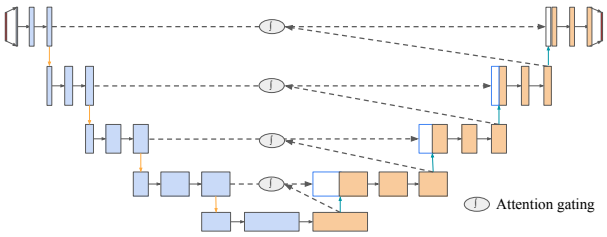

Figure 1: *The attention U-Net model. Dotted lines represent the paths of attention gating in contracting/expanding layers.*

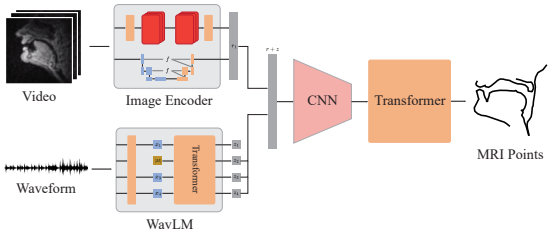

Figure 2: *Architecture of the multimodal segmentation model.*

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.

3. Models and Training

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. This is followed by 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 was a natural fit. 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 apply attention gating following the Attention U-Net [22] with the modification of using additive attention as opposed to multiplicative, visualized in Figure 1. While minimally increasing complexity, the model learns to suppress the components of the signal which are not important for the labeling task.

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 guides the model into producing an output that also appears Gaussian in nature and is intuitively well suited 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. In this context, articulator "importance" is determined by the effect that dropping the articulator has on downstream speech synthesis. Both the importance and standard deviation were calculated using the 7 training speakers by previous works [15]. We multiply 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.

3.3. Multimodal 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 [23] (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 [24, 25]. 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 ex-

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 4.1.*

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

perimented with other image representation models including ResNet [26] and ConvNeXt [27]. To represent audio, we used the 10th layer of WavLM [28] 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.

4. 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.

4.1. Vision-only U-Net

The first experiment compared L2 (mean squared error) loss against 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. As RMSE and MSE have the same convergence point, articulatory weighting predictably appears worse using this metric. However, manual inspection reveals that most of this error can be attributed to shifts in less phonologically important articulators such as the hard palate, with significant improvement on the more important articulators.

4.2. Labeling the Speech MRI Open Dataset

The vision-based U-Net above was used to provide labels to RT-MRI video for the 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 3, 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

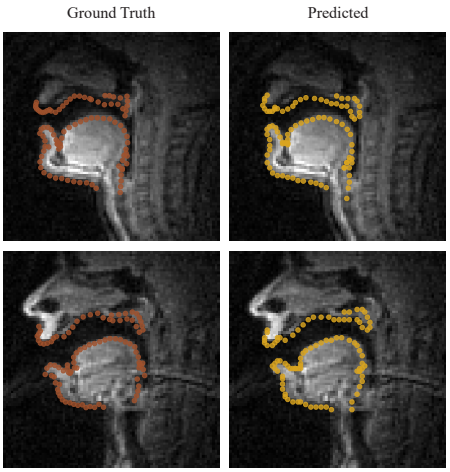

Figure 3: *Two representative examples of predicted MRI points (right) compared to expert hand labels (left). The examples are spoken by unseen Female (bottom) and Male (top) speakers in the Speech MRI Open Dataset.*

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 `rishiraij.github.io/multimodal-mri-avatar/`.

4.3. 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 4 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.

Similarly, we evaluate our segmentation methods on downstream speech tasks using speech synthesis within seen and unseen speaker contexts. Using the state-of-the-art MRI synthesis model [29] pretrained on the newly-labeled 75-speaker dataset, we finetune on the projected MRI trajectories of a USC-TIMIT speaker provided by the different feature extraction models (i.e. baseline, U-Net, and multimodal). To evaluate the intelligibility of synthesized speech, we compute the word error rate (WER) on test unseen examples using Whisper [30], a state-of-the-art automatic speech recognition (ASR) model. For seen speakers, speech synthesized using the multimodal U-Net + WavLM based segmentations is more intelligible than speech synthe-

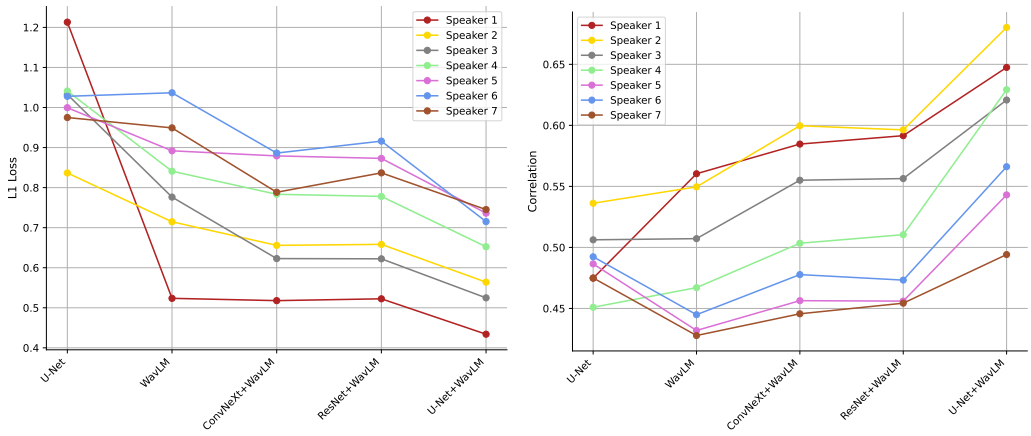

Figure 4: *L1 losses [↓] (left) and Pearson Correlation Coefficients (PCCs) [↑] (right) comparing MRI trajectories of unseen examples from seen speakers of a given model with the USC-TIMIT ground truth. Varying through a subset of six representative models.*

Table 2: *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	**31.3% (16.4%-49.3%)**
U-Net	36.4% (20.9%-55.1%)
Ground Truth	34.7% (18.6%-53.2%)
U-Net + WavLM (S)	34.9% (20.3%-52.8%)

Table 3: *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	**33.3% (20.2%-49.8%)**
U-Net	35.2% (17.2%-56.8%)
Ground Truth	49.7% (34.8%-66.6%)
U-Net + WavLM (S)	50.1% (28.0%-72.8%)

sized from either the ground truth baseline or the U-Net outputs, 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 2 summarizes these results. The results in Table 3 highlight that the U-Net + WavLM based model has the lowest WER when testing on an unseen USC-TIMIT speaker, documenting that the segmentations from the multimodal model on unseen speakers still capture representative articulatory kinematics for naturalistic speech. Pretraining the synthesis model on the 75 speakers also results in much better unseen speaker generalization, demonstrating that the new labels for the Speech MRI Open Dataset are beneficial for future work in articulatory speech.

4.4. 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 hypothesize 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 deep learning approaches do not have the same level of frame-dependent noise, 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 the baseline 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. Each participant repeated this process for five test examples. Our results reveal the participants (n=21) prefer the outputs of our algorithm over the baseline segmentations ($p < 0.001$). For visualization of these results, we invite you to watch our demo video at `rishiraij.github.io/multimodal-mri-avatar`.

5. Conclusion

In this work, we looked at developing a generalizable articulatory segmentation algorithm from RT-MRI videos of the vocal tract. We used the limited existing articulatory labeling to train vision-based and multimodal models which efficiently and accurately extract physiological features from MRI videos of unseen speakers. Through speech synthesis, we demonstrate that our approach results in higher quality segmentations for downstream speech tasks than existing baselines, while also being more accurate representations of speech audio. While MRI-based articulatory modeling is less studied than other approaches such as EMA, we hope that our released labeling of 75 speakers will allow future work in speech modeling and linguistics to take advantage of the more-complete physiological representation that RT-MRI provides.

6. References

[1] A. Suemitsu and J. 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] J. S. Levitt and W. 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] B. Gick, B. M. Bernhardt, P. Bacsfalvi, and I. Wilson, "11. ultrasound imaging applications in second language acquisition," in *Phonology and Second Language Acquisition*, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:63438867

[4] S. L. Metzger, K. T. Littlejohn, A. B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger, I. Zhuravleva, A. Tu-Chan, K. Ganguly, G. K. Anumanchipalli, and E. F. Chang, "A high-performance neuroprosthesis for speech decoding and avatar control." *Nature*, vol. 620, pp. 1037–1046, 2023.

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

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

[7] C. J. Cho, P. Wu, A. Mohamed, and G. 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] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, p. 493–498, Apr. 2019. [Online]. Available: http://dx.doi.org/10.1038/s41586-019-1119-1

[9] Y. Lu, C. E. Wiltshire, K. E. Watkins, M. Chiew, and L. 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] A. Richard, C. Lea, S. Ma, J. Gall, F. de la Torre, and Y. 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] S. Medina, D. Tome, C. Stoll, M. Tiede, K. Munhall, A. Hauptmann, and I. Matthews, "Speech driven tongue animation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2022. [Online]. Available: https://doi.org/10.1109/cvpr52688.2022.01976

[12] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. 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] D. Gaddy and D. 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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 175–181. [Online]. Available: https://aclanthology.org/2021.acl-short.23

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

[15] P. Wu, T. Li, Y. Lu, Y. Zhang, J. Lian, A. W. Black, L. Goldstein, S. Watanabe, and G. K. Anumanchipalli, "Deep speech synthesis from mri-based articulatory representations," 2023.

[16] E. Bresch and S. 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] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. 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, p. 1307–1311, Sep. 2014. [Online]. Available: http://dx.doi.org/10.1121/1.4890284

[18] Y. Yu, A. H. Shandiz, and L. Tóth, "Reconstructing speech from real-time articulatory mri using neural vocoders," 2021.

[19] Y. Otani, S. Sawada, H. Ohmura, and K. Katsurada, "Speech Synthesis from Articulatory Movements Recorded by Real-time MRI," in *Proc. INTERSPEECH 2023*, 2023, pp. 127–131.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[21] Y. Lim, A. Toutios, Y. Bliesener, Y. Tian, S. G. Lingala, C. Vaz, T. Sorensen, M. Oh, S. Harper, W. Chen, Y. Lee, J. Töger, M. L. Monteserin, C. Smith, B. Godinez, L. Goldstein, D. Byrd, K. S. Nayak, and S. 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. [Online]. Available: http://dx.doi.org/10.1038/s41597-021-00976-x

[22] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[23] S. A. Hebbar, R. Sharma, K. Somandepalli, A. Toutios, and S. 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.

[24] F. Loewenich and F. 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: IOS Press, 2006, p. 25–30.

[25] Y. Chen, D. Zhao, and H. 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.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[27] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022.

[28] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. 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, p. 1505–1518, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2022.3188113

[29] Anonymous, "Transfer learning for articulatory synthesis," 2024, preprint. [Online]. Available: https://openreview.net/pdf?id=pEW9cXrY4O

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.