# Transfer Learning for Articulatory Synthesis

*Anonymous submission to Interspeech 2024*

## Abstract

Articulatory data is extremely limited, and particularly so when compared to acoustic speech data. We propose three transfer learning techniques to noticeably improve articulatory-to-acoustic synthesis performance in these low-resource settings: (1) pre-trained weight initialization, (2) pre-training part of the model, and (3) multimodal pre-training. On single-speaker MRI-, EMG-, and EMA-to-speech tasks, the intelligibility of synthesized outputs improves noticeably. For example, compared to prior work, utilizing our proposed transfer learning methods improves the MRI-to-speech performance by 57% word error rate (WER). We also propose a deep speech representation that outperforms self-supervised learning features and spectrums as an intermediate for articulatory synthesis.

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

## 1. Introduction

Articulatory synthesis aims to incorporate information about the vocal tract into speech synthesizers to improve interpretability, generalizability, and efficiency [1, 2, 3, 4, 5, 6]. Additionally, these models can be applied to decoding speech from biosignals for health technology applications [7, 8, 9, 10, 11, 12]. Articulatory data is limited compared to other types of language data like acoustics and text [13, 14, 15]. Thus, to improve generalizability, deep articulatory synthesizers could utilize transfer learning from other datasets and modalities, a methodology successful in a variety of deep learning tasks [16]. In this paper, we propose three transfer learning methods: (1) pre-trained weight initialization, (2) pre-training part of the model, and (3) multimodal pre-training. Through single-speaker magnetic resonance imaging (MRI), electromyography (EMG), and electromagnetic articulography (EMA) to speech tasks, we find these transfer techniques effective for improving synthesized speech quality. With less than 10 minutes of single-speaker training data, our MRI-to-speech model achieves a test-set automatic speech recognition (ASR) word error rate (WER) of 33%, compared to 90% from the previous model [6]. Our EMA- and EMG-to-speech models similarly noticeably outperform the baseline, and human listening tests results match our ASR trends.

## 2. Deep Articulatory Synthesis

Deep articulatory synthesis involves synthesizing acoustics from articulatory features using a deep learning model [8, 9, 10, 5, 6, 11]. Current approaches can generally be described as either direct or involving an intermediate representation. Direct synthesis maps articulatory inputs to acoustics with a single end-to-end model, whereas synthesis with intermediates maps inputs to intermediate features, which are then mapped to acoustics.

To map articulatory or intermediate features to waveforms, we use HiFi-CAR, an auto-regressive temporal convolutional network optimized with adversarial training [5, 17, 18]. To map articulatory inputs to intermediate features, we build on the EMG-to-spectrum model proposed by [10]. Specifically, we map articulatory representations to the intermediate representations using a six-layer Transformer [19] prepended with three residual convolution blocks.

## 3. Articulatory Datasets

### 3.1. 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 [20, 5]. We use MNGU0, a single-speaker dataset containing 67 minutes of 16 kHz speech and 200 Hz EMA [13]. 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 [14]. To maintain consistency with prior work [5, 21, 22], 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.

### 3.2. Magnetic Resonance Imaging (MRI)

Another articulatory modality that we experiments with is real-time magnetic resonance imaging (MRI), which provides a more comprehensive feature set of the human vocal tract than EMA [23, 24, 6, 25]. In addition to the six locations described by EMA, midsagittal MRI images contain locations of the hard palate, pharynx, epiglottis, velum, and larynx, all of which are useful for speech synthesis [6]. In this work, we used the same 11-minute, single-speaker real-time MRI dataset as [6]. This dataset is comprised of 20 kHz speech and $83.\overline{3}$ Hz midsagittal MRI data, with 170 x-y points annotated for each MRI frame. Following [6], 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.

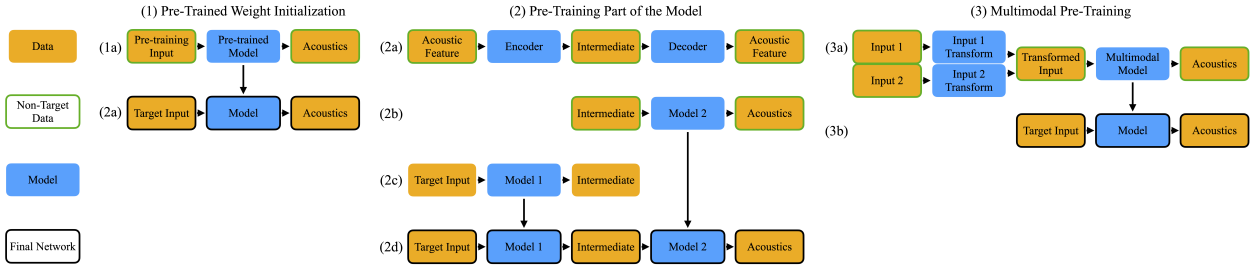


Figure 1: *Three transfer learning approaches for articulatory synthesis.*

### 3.3. Electromyography (EMG)

Surface electromyography (EMG) measures electrical potentials caused by nearby muscle activity using electrodes placed on top of the skin [26]. When placed near articulators, EMG provides another low-dimensional manifold of articulatory movements [26, 27, 10, 11]. 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 waveforms have a sampling rate of 16 kHz, and EMG 1000 Hz.

## 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 [6], visualized in Figure 1 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 5.1.

### 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 [28], and many classifiers pre-train their encoder [29]. Popular vocoder input representations include spectrums, high-dimensional self-supervised features, learnt representations, and units [28, 30, 31, 32]. This pre-training method has also shown success with ultrasound-speech tasks [33]. 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 [34] self-supervised representation in order to reduce the complexity of mapping to this intermediate feature, visualized as (2a) in Figure 1. We choose HuBERT given its success with other synthesis tasks [35, 36], 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) [37]. 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 tables 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 1). 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 2), visualized as step (2b) in Figure 1. Thirdly, in step (2c), we train an articulatory-to-intermediate Transformer (Section 2), 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 5.2.

### 4.3. Multimodal 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 [38, 39]. 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 1 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 al., 2023) [21]. 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 result-

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

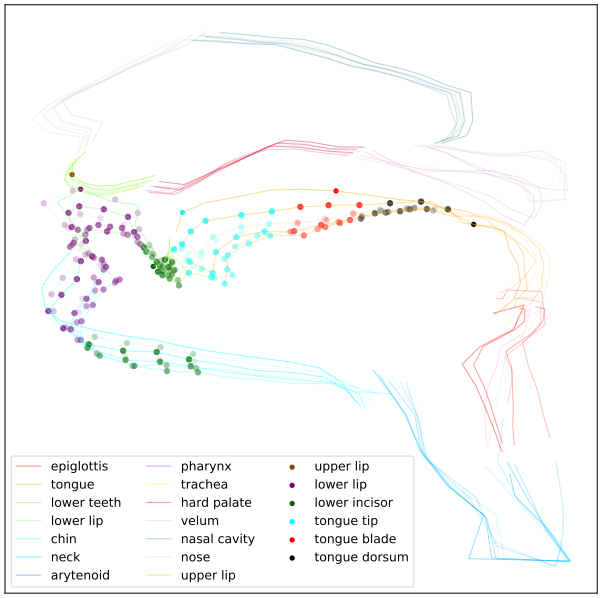


Figure 2: *Extracted MRI-features for the utterance "apa." Lighter is earlier in time. Each point is colored with the highest-correlation EMA feature. Points with maximum correlation magnitude below 0.3 are omitted for readability.*

ing 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 5.3.

To provide more intuition on multimodal pre-training, Figure 2 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. 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 [40]. 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 [40] 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 [6, 10], 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.

### 5.1. 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 3.1. Our EMA-to-speech model here is HiFi-CAR, described in Section 2, 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 [15] HiFi-GAN mel-spectrogram to speech vocoder weights in [17, 6]. 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 [41], We compute pitch using CREPE [42, 43] 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 [5]. 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 [44], with WER results in Table 1. 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.

Table 1: *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) |

Table 2: *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) |

### 5.2. 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 3.2 and 3.3, 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 an 85%-5%-10% train-validation-test split [45].

Our baseline for MRI-to-speech is [6], labeled Direct in Tables 2 and 3. Specifically, this is the HiFi-CAR model described in Section 2 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

Table 3: *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 |

Table 4: *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) |

have a baseline to our knowledge [10], 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 4 and 3. For partially pre-trained models, we map inputs to intermediates (Section 4.2) using the Transformer in Section 2, 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 5.1 and human evaluation. As shown in Tables 2 and 4, 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 3 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.

### 5.3. 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 5.2, with the model being modified during the multi-modal pre-training step as described in Section 4.3. We pre-train our model with: (1) all of the EMA data in the HPRC dataset described in Section 3.1, and (2) the training set of our MRI dataset described in Section 3.2. The pre-training and fine-tuning steps both use the Adam optimizer with a learning rate of $10^{-4}$ [40]. 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.

Table 5: *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 6: *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 | **0.67 ± 0.12** | 0.34 ± 0.12 |
| HuBERT | **0.84 ± 0.24** | 0.17 ± 0.24 |

We evaluate these models with the same ASR metric as Section 5.2. Table 5 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 [6]. 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 5.2, we average scores for each model to give a number between 0 and 1, with 1 being the best possible score. Table 6 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. Conclusion

In this work, we devise three transfer learning techniques for improving the performance of articulatory synthesizers: (1) pre-trained weight initialization, (2) pre-training part of the model, and (3) multimodal pre-training. Through EMA-, MRI-, and EMG-to-speech experiments, we validate the effectiveness of these approaches. Additionally, we propose a deep speech representation that outperforms self-supervised speech representations and spectrums as an intermediate for articulatory synthesis. Training on less than 10 minutes of single-speaker data, our MRI-to-speech model achieves a test-set automatic speech recognition (ASR) word error rate (WER) of 33%, which is much better than the 90% WER from the previous model [6]. 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.

# 7. References


[1] C. P. Browman, L. Goldstein, J. S. Kelso, P. Rubin, and E. Saltzman, "Articulatory synthesis from underlying dynamics," *The Journal of the Acoustical Society of America*, vol. 75, no. S1, pp. S22–S23, 1984.

[2] C. Scully, "Articulatory synthesis," in *Speech production and speech modelling*. Springer, 1990, pp. 151–186.

[3] B. J. Kröger and P. Birkholz, "Articulatory synthesis of speech and singing: State of the art and suggestions for future research," *Multimodal Signals: Cognitive and Algorithmic Issues: COST Action 2102 and euCognition International School Vietri sul Mare, Italy, April 21-26, 2008 Revised Selected and Invited Papers*, pp. 306–319, 2009.

[4] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016.

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

[6] ——, "Deep speech synthesis from mri-based articulatory representations," *Interspeech*, 2023.

[7] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–685.

[8] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "Dnn-based ultrasound-to-speech conversion for a silent speech interface," *Interspeech*, 2017.

[9] N. Kimura, M. Kono, and J. Rekimoto, "Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11.

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

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

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

[13] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Interspeech*, 08 2011, pp. 1505–1508.

[14] M. K. Tiede *et al.*, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *JASA*, 2017.

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

[16] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, pp. 1–40, 2016.

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

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

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

[20] P. 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.

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

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

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

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

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

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

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

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

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

[30] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 644–12 652.

[31] M. Kim, M. Jeong, B. J. Choi, D. Lee, and N. Kim, "Transduce and speak: Neural transducer for text-to-speech with semantic token prediction," in *ASRU*, 12 2023, pp. 1–7.

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

[33] J.-X. Zhang, K. Richmond, Z.-H. Ling, and L. 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*, vol. 35, no. 16, 2021, pp. 14 402–14 410.

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

[35] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Proc. Interspeech 2021*, 2021.

[36] R. Huang, J. Liu, H. Liu, Y. Ren, L. Zhang, J. He, and Z. Zhao, "Transpeech: Speech-to-speech translation with bilateral perturbation," *ICLR*, 2023.

[37] R. T. Chen *et al.*, "Neural ordinary differential equations," *NeurIPS*, 2018.

[38] P. P. Liang *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," *NeurIPS*, 2021.

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

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

[41] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "Ddsp: Differentiable digital signal processing," *ICLR*, 2020.

[42] J. W. Kim, J. Salamon, P. Li, and J. P. 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.

[43] M. Morrison, C. Hsieh, N. Pruyne, and B. Pardo, "Cross-domain neural pitch and periodicity estimation," *arXiv preprint arXiv:2301.12258*, 2023.

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

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