# A Real-Time Speech-Driven Vocal Tract Avatar

*Anonymous submission to Interspeech 2024*


## Abstract

Recent advances in deep articulatory inversion have enabled the development of high quality offline speech-driven avatars. However, real-time speech-driven avatars remain an unsolved problem, with applications in language acquisition, automated facial animation, and virtual embodiment for multimedia and virtual reality. Thus, we build a speech-driven avatar system for real-time facial and inner-mouth animation. Using a novel pseudolabeling technique, we approximate self-supervised learning based inversion models in streaming scenarios and integrate a mapping between predicted articulatory features and a web-based 3D avatar. We evaluate our models with varying context lengths and achieve an average latency of 34ms per 100ms of streamed audio and 0.725 correlation with 20ms of forward context, improving the state-of-the-art by 64%. Finally, we conduct a subjective user study to validate the realism of the speech-driven avatar and obtain a 77% improvement over the baseline.



**Index Terms**: speech driven, real-time, talking avatar, articulatory


## 1. Introduction

Speech-driven avatar animation is useful for many applications in speech and linguistics. It can facilitate second language pronunciation learning via visual feedback and aid hearing-impaired individuals to lip-read when only an audio signal is available during communication [1, 2, 3]. In addition, accurate facial and tongue animation has been shown to help with virtual embodiment for paralyzed patients [4].

Key tasks in speech-driven avatar research include both offline and real-time animation of the face and inner mouth. Such systems have use in interactive multimedia such as video games and virtual reality [5]. Solutions to the offline speech-driven avatar task date back to [6, 7], which propose predicting phonemes with optical motion tracking and electromagnetic articulography (EMA) data. More recent works focusing on offline synthesis have achieved success with linguistic phoneme-viseme mappings and mesh deformations using audio-mesh datasets [8, 9, 10, 11, 12]. However, these systems require manually defined complex rules and cannot realistically model inner mouth movements during articulation.

Recently, advances in deep articulatory inversion have demonstrated the ability to approximate physiology-grounded vocal tract models from solely speech inputs [13, 14, 15, 16, 17, 18, 19]. Notably, self-supervised learning (SSL) speech features encode vocal tract information for generalizable prediction of articulatory kinematics [20, 21]. These inversion models have inspired offline speech-driven avatars powered by articulatory representations of the vocal tract, such as magnetic resonance imaging (MRI) and EMA [22, 23, 24]. In particular, [22] achieves high fidelity, realistic, offline avatar animations using speech-EMA inversion followed by a rig optimizer model to map EMA to a 3D avatar.

However, these offline successes have not been extended to streaming solutions yet. Previously, ultrasound imaging has been streamed to a live tongue animation system, but it lacks full facial animations and requires a real-time ultrasound machine during inference [25]. Deep neural networks have shown a real-time acoustic-visual mapping is possible without needing an imaging machine [26], but do not yet offer avatar visualizations via a 3D facial mesh or latency analysis for streaming.

In this work, we aim to apply deep articulatory inversion to an integrated real-time speech-driven facial and tongue animation system. Due to the scarcity of EMA data, we first deploy an an SSL-based model to pseudolabel the VCTK speech dataset. We then propose a streaming inversion model to predict real-time EMA data from buffered speech, animate a joint-based 3D avatar on the web, and evaluate predicted animations by comparing the generated motion capture to ground truth labels. We achieve average latencies of 34ms per 100ms of streamed audio and demonstrate average correlations of 0.725 when comparing streamed predictions to ground truth EMA. Additionally, we evaluate the speech-driven avatar in a subjective evaluation, improving over the baseline by 77%.

## 2. Related Work

State-of-the-art articulatory inversion models in [13, 20, 21] utilize self-supervised learning (SSL) speech features as model inputs to take advantage of SSL generalizability to unseen speakers and languages. This source agnosticity is crucial for a general speech-driven avatar for unseen speakers and makes SSL speech information valuable for this task. However, these models are only suitable for offline scenarios due to their high inference times. Thus, we motivate our work in pursuit of bridging the gap between offline inversion model quality and low latency inference.

## 3. Streaming Articulatory Inversion

### 3.1. Datasets

MNGU0 is a single-speaker dataset with 67 minutes of 16 kHz British English speech and 200 Hz EMA [27]. The EMA features capture the $x, y$ midsagittal movements of 6 articulator positions: lower incisor, upper lip, lower lip, tongue tip, tongue body, and tongue dorsum. We refer to MNGU0's vocal tract as represented by this EMA data as the common vocal tract for the remainder of this paper. Having a common target vocal tract for any input speaker is desirable for the real-time avatar task, as we

can build a general speech-avatar mapping rather than creating speaker-specific avatar meshes.

The CSTR VCTK dataset contains 44 hours of 48 kHz English speech from 109 speakers [28]. We pseudolabel VCTK with EMA features for our streaming inversion models as described in Section 3.2.

### 3.2. Speech Pseudolabeling

To achieve a streaming speech-driven avatar, we aim to capture the generalizability of SSL-based inversion models to unseen speakers and languages without the overhead of generating SSL features on the fly. To do this, we propose a waveform-based inversion model trained on pseudolabeled speech-EMA data. In particular, we follow [21] and train a linear probing model on WavLM-large SSL features for the MNGU0 dataset (80/5/5 training split) [29]. Following [13, 20, 21], we use the articulatory score (average Pearson correlation coefficient) as a standard metric for measuring plausibility of predicted EMA trajectories. This linear "teacher" model achieves an articulatory score of $0.871 \pm 0.045$ when comparing the predicted EMA with the MNGU0 ground truth.

A benefit of the SSL-based linear model is its learned bijection between any speaker's WavLM speech features and the common vocal tract. We represent this relationship as:

$$L : W \rightarrow Y \tag{1}$$

$$L' : Y \rightarrow W \tag{2}$$

where $W$ represents WavLM SSL features for any speaker, $Y$ represents the common vocal tract, and $L$ and $L'$ represent invertible mappings between the two.

We wish to learn $S : X \rightarrow Y$, where $X$ represents any speaker's raw waveform, and $S$ represents a new mapping to the common vocal tract without using explicit SSL features. We propose that our streaming model learns $S$ by modeling the joint probability of a streamed EMA prediction $\boldsymbol{y}$ as the product of conditional probabilities of a sliding window of waveform inputs:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{T} p(y_t|x_t, x_{t-1}, \ldots, x_{t-\ell}) \tag{3}$$

where $x_t$ is each waveform batch and $\ell$ is the length of the sliding window. However, without the rich SSL speech information already encoded within $W$, we require a much larger amount of data than the 67 minutes of MNGU0 to model $p(\boldsymbol{y})$ effectively. Further, since $S$ maps any speaker's speech to the common vocal tract, a multi-speaker dataset is necessary for generalizable inversion.

Thus, we introduce a pseudolabeling technique for EMA data augmentation. Using the linear model $L$ from Section 3.2, we label the 44 hours of speech data in VCTK with predicted EMA. This synthetic data is used to train streaming inversion models learning $S$.

### 3.3. Streaming Inversion

We train two streaming inversion models using a chunked autoregressive convolutional neural network (CAR-CNN) and the efficient memory Transformer (Emformer) [30, 31, 32]. We motivate these choices in twofold: the CAR-CNN represents a simple waveform-based model for fast inference, and the Emformer attempts to replace the full-context Transformer used in predicting WavLM SSL features. Incoming waveform inputs are first fed through the pretrained WavLM convolutional feature encoder. Additionally, we use the mean absolute error loss (L1) as the objective function for training both models.

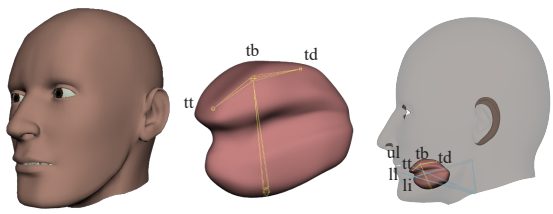


Figure 1: *Three-quarter view of full 3D avatar (left), three-quarter view of tongue with joint-based rig labeled (middle), and midsagittal view of face and tongue with joint-based rigs labeled (right). Key: ul - upper lip, ll - lower lip, li - lower incisor, tt - tongue tip, tb - tongue body, td - tongue dorsum.*

The CAR-CNN autoregressively models the distribution of EMA $\boldsymbol{y}$ as:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{T} p(y_t|x_t, \ldots, x_{t-\ell}, y_{t-1}, \ldots, y_{t-\ell}) \tag{4}$$

The chunked autoregressive component $y_{t-1}, \ldots, y_{t-\ell}$ is encoded using a three-layer CNN and concatenated with the waveform inputs following [33]. The overall network architecture is a fully convolutional neural network following the Multi-Receptive Field Fusion module from [34]. We vary kernel sizes and dilation rates for multiple residual blocks and sum their outputs.

The Emformer models the original joint probability in Equation 3. It is optimized for the real-time scenario by saving repeated left context attention computations in an updated summary state used in subsequent $y_t$ predictions. We experiment with varying context lengths and monitor the effect on both EMA prediction quality and latency during inference.

For our streaming baseline, we use the BiGRU inversion model in [13]. To our knowledge, it is state-of-the-art for offline speaker-independent inversion. In a streaming scenario, we chunk incoming audio in batches of 100ms, infer SSL features within a sliding window using HuBERT [35], then use the baseline BiGRU model to predict streamed EMA.

## 4. Streaming 3D Avatar

### 4.1. Avatar Mesh and Rig

The facial model in Figure 1 was constructed in Autodesk Maya using teeth from [36], the face from [37], and a custom tongue. A joint-based rig was built to control each of the 6 EMA articulators. The deformable avatar mesh can thus be directly controlled by the position of each articulator as predicted by the streaming inversion models. We use the common vocal tract's minimum and maximum values for each EMA feature for renormalization.

### 4.2. Streaming Speech Processing

While streaming avatar animations, we receive 100ms of audio per batch. To mitigate noise, we use a voice activity detector (VAD) to make sure we only run inference on waveform batches containing speech. Otherwise, the previous avatar frame is maintained for the batch duration.

Table 1: *Mean ± standard deviation PCCs and latency reports for all inversion models as future context is varied. Success rate is defined as the ratio of complete-system inference below the buffer size of 100ms. Past context is fixed to 500ms for each model.*

| Model | Future Context (ms) | MNGU0 PCC [↑] | VCTK PCC [↑] | Inference Time (ms) [↓] | Success Rate [↑] |
|---|---|---|---|---|---|
| SSL-BiGRU | 0 | $0.031 \pm 0.138$ | – | $135.76 \pm 8.46$ | $0.0 \pm 0.0$ |
| | 20 | $0.441 \pm 0.129$ | – | $137.56 \pm 23.16$ | $0.0 \pm 0.0$ |
| | 80 | $0.468 \pm 0.122$ | – | $123.266 \pm 23.71$ | $0.0 \pm 0.0$ |
| | 1000 | $0.616 \pm 0.100$ | – | $169.84 \pm 22.67$ | $0.0 \pm 0.0$ |
| CAR-CNN | 0 | $0.574 \pm 0.088$ | $0.672 \pm 0.066$ | $31.27 \pm 19.60$ | $0.998 \pm 0.008$ |
| | 20 | $0.608 \pm 0.079$ | $0.717 \pm 0.061$ | $34.92 \pm 31.26$ | $0.995 \pm 0.020$ |
| | 80 | $0.648 \pm 0.079$ | $0.752 \pm 0.059$ | $30.97 \pm 6.531$ | $1.0 \pm 0.0$ |
| | 1000 | $0.707 \pm 0.083$ | $0.822 \pm 0.053$ | $47.22 \pm 26.38$ | $0.996 \pm 0.014$ |
| Emformer | 0 | $0.653 \pm 0.079$ | $0.818 \pm 0.051$ | $34.13 \pm 5.97$ | $0.998 \pm 0.004$ |
| | **20** | $\mathbf{0.725 \pm 0.077}$ | $\mathbf{0.875 \pm 0.043}$ | $\mathbf{34.27 \pm 7.65}$ | $\mathbf{0.997 \pm 0.006}$ |
| | 80 | $0.720 \pm 0.072$ | $0.855 \pm 0.046$ | $64.16 \pm 10.71$ | $0.990 \pm 0.022$ |
| | 1000 | $0.646 \pm 0.088$ | $0.700 \pm 0.078$ | $346.91 \pm 45.77$ | $0.0 \pm 0.0$ |

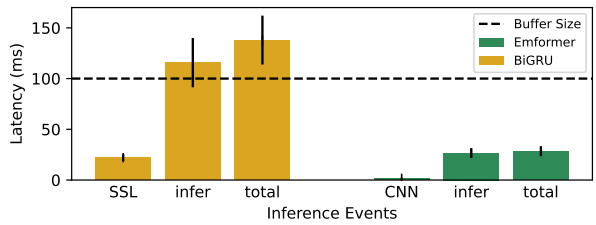


Figure 2: *Latency contributions during inference events for the Emformer and SSL-BiGRU with 20ms forward context. SSL refers to WavLM SSL feature extraction and CNN refers to WavLM convolutional feature extraction.*

Since the streaming model initially has no prolonged audio context to condition predictions on, we artificially inject context in the form of a random speech utterance prepended to the audio stream. We find this reduces initial noise in the predicted EMA. Each batch of EMA is then sent to the 3D avatar for animation.

### 4.3. Web-Based Avatar Animation

We host the avatar on the web, allowing any user with an OpenGL device to test the proposed system. A WebRTC backend hosts the streaming model and engages in two-way streamed communication with the client for real-time avatar animation. A queue of EMA is maintained in the client as batches are received from the streaming model. The avatar is animated by absolute joint movements correspondingly. The demo website link and code will be publicly available post-anonymity.

## 5. Results

### 5.1. Articulatory Scores

To achieve an optimal configuration for streaming inversion, we assess articulatory scores of the baseline SSL-BiGRU, the CAR-CNN, and Emformer with varying future context lengths, reported in Table 1. We fix the backward context to 500ms for each model. For the streaming models, we first evaluate articulatory scores on the VCTK pseudolabeled test data. The CAR-CNN improves with increasing future context, while the Emformer achieves a higher peak of 0.875 with 20ms of future context then declines with additional context.

To verify this performance, we compare each model to the baseline on the ground truth MNGU0 test data. Both streaming models far outperform the SSL-BiGRU, which requires at least one full second of future context for scores greater than 0.6. We observe a similar trend of increasing scores with future context for the SSL-BiGRU and the CAR-CNN. The best model, the Emformer with 20ms future context, achieves 0.725 average correlations with the ground truth. Notably, the Emformer with 0ms of future context achieves a 0.653 score, outperforming all SSL-BiGRU configurations and $<$ 1000ms CAR-CNN configurations. In contrast, the SSL-BiGRU with 0ms future context does not accurately represent the ground truth (average correlation of 0.031), which suggests that our method excels at inference from smaller, moving contexts.

This trend is further illustrated in Figure 3, where we compare baseline and Emformer EMA trajectories with the ground truth for the tongue tip (tt) feature across increasing future context lengths. We observe decreases in high frequency jitter as future context increases for both models. Additionally, the supporting figure on the right highlights higher scores for the Emformer model across all future context lengths, suggesting that it serves as more accurate system in streaming scenarios.

### 5.2. System Latency

In a streaming scenario, real-time inference is required for realistic avatar animations. Average syllable lengths can vary from 50ms to 400ms in length [38]. To appear real-time, we target latencies between the range of 50ms and 100ms, as avatar animations will be less than one syllable length behind streamed speech. To evaluate this, we measure average inference times using an RTX 3060 Max-Q GPU for each inversion model while varying future context lengths, as presented in Table 1.

For each model, inference times generally increase with increasing forward contexts. The CAR-CNN and Emformer have

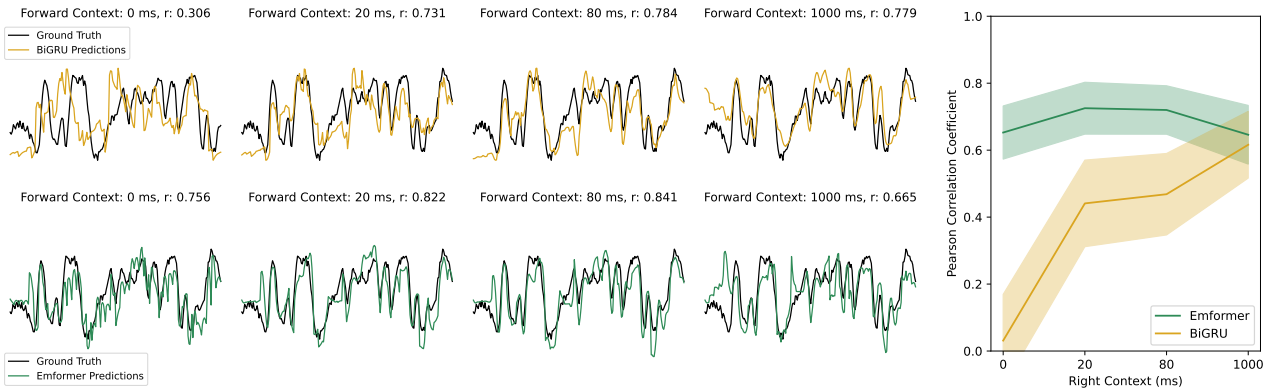


Figure 3: *Left: Streamed tongue tip (tt) EMA prediction trajectories of a test MNGU0 utterance compared to the ground truth with PCCs when varying future context lengths. Right: PCCs comparing streamed EMA prediction trajectories and the ground truth for the MNGU0 test set when varying future context lengths. Both Emformer and BiGRU have a constant 500ms left context window.*

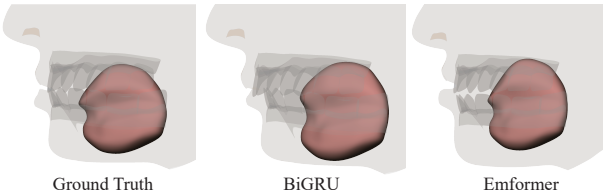


Figure 4: *Avatar visualizations of the /j/ phoneme during utterance of "alleged". Both models use 20ms forward context length.*

Table 2: *Subjective mean opinion scores (MOS) of streamed BiGRU and Emformer-based avatars for MNGU0 test set utterances, with 95% confidence intervals in parentheses.*

| | **Pref. Score** (%) [↑] | **MOS** [↑] |
|---|---|---|
| Offline BiGRU | – | 4.31 (4.03-4.58) |
| Real-Time BiGRU | 2.8 | 2.22 (1.85-2.59) |
| Emformer | **88.9** | **3.94 (3.64-4.25)** |

comparable latencies when forward context is limited, while Emformer inference times scale more with larger forward context lengths. However, both streaming models outperform the SSL-BiGRU by a wide margin, achieving 34ms inference times on average with 20ms forward context length. In contrast, the BiGRU's lowest average inference time is 123ms with 80ms forward context. Thus, the BiGRU never completes inference within the buffer size of 100ms.

To further quantify this result, we report the success rate of each inversion model configuration in Table 1, calculated as the ratio of complete-system inference below the buffer size of 100ms. Every streaming model except the Emformer with 1000ms of forward context achieves a success rate of at least 99%, while all BiGRU configurations have a 0% success rate.

We observe that these gains in inference speed can be attributed to significant improvements in both feature extraction and subsequent EMA inference. In Figure 2, we report the latency contributions from each stage for the BiGRU and the Emformer (20ms forward context). The replacement of SSL features with the WavLM convolutional features saves over 20ms of inference time. The Emformer architecture saves an additional 90ms, allowing the system to remain firmly under both the 100ms buffer size and the 50ms syllable length lower bound.

### 5.3. Avatar

To investigate the avatar visualization during streaming, we conducted a paired comparison test and a subjective quality evaluation test ($n = 9$). In the paired comparison tests, participants were presented with video pairs containing midsagittal views of the avatar using streamed MNGU0 test set predictions (BiGRU vs Emformer, 20ms forward context). Examples of the midsagittal avatar view are shown in Figure 4. They were then asked to choose which avatar video they preferred (with an option to indicate no preference). In the subjective quality evaluation test, participants watched MNGU0 test set avatar videos from a real-time BiGRU, an offline BiGRU, and the Emformer. Each participant was asked to rate how well the visualization represented the speech in a five-point Likert scale score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).

The results from this study are detailed in Table 2. In the paired comparison test, the Emformer-based avatar is largely preferred over the real-time BiGRU with a preference score of 88.9%. In the subjective evaluation test, the Emformer-based avatar achieves a 77% increase from the baseline BiGRU mean opinion scores (MOS) and approaches comparable scores to the offline BiGRU.

## 6. Conclusion

We develop high fidelity, low latency streaming inversion models for use in a real-time, web-based 3D speech-driven avatar system. Using a pseudolabeling technique, we approximate SSL-based articulatory inversion models with waveform-based models optimized for streaming. We validate these models with Pearson correlations across varying forward context lengths and observe that with just 20ms of forward context, streaming models noticeably outperform SSL-based models with average correlations of 0.725 against ground truth EMA. The avatar system is subjectively evaluated and surpasses the baseline by 77%. Future work may extend these results with video to enable expressions conditioned on vocal articulations.

# 7. 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 *et al.*, "11. ultrasound imaging applications in second language acquisition," 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:63438867

[4] S. L. Metzger *et al.*, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. 7976, pp. 1037–1046, Aug. 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06443-4

[5] M. Radovan and L. Pretorius, "Facial animation in a nutshell," in *Proceedings of the 2006 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing couuntries - SAICSIT '06*. ACM Press, 2006. [Online]. Available: https://doi.org/10.1145/1216262.1216270

[6] J. Beskow *et al.*, "Synface - a talking head telephone for the hearing-impaired," *Lecture Notes in Computer Science*, 2003.

[7] G. Salvi *et al.*, "SynFace—speech-driven facial animation for virtual speech-reading support," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, pp. 1–10, 2009. [Online]. Available: https://doi.org/10.1155/2009/191940

[8] P. Edwards *et al.*, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on Graphics*, vol. 35, no. 4, p. 111, Jul. 2016. [Online]. Available: http://dx.doi.org/10.1145/2897824.2925984

[9] E. Setyati *et al.*, "Phoneme-viseme mapping for indonesian language based on blend shape animation," in *IAENG International Journal of Computer Science*, 2015.

[10] D. Cudeiro *et al.*, "Capture, learning, and synthesis of 3D speaking styles," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 101–10 111. [Online]. Available: http://voca.is.tue.mpg.de/

[11] J. Xing *et al.*, "Codetalker: Speech-driven 3d facial animation with discrete motion prior," *arXiv preprint arXiv:2301.02379*, 2023.

[12] M. V. Aylagas *et al.*, vol. 41, no. 8, pp. 255–265, Dec. 2022. [Online]. Available: https://doi.org/10.1111/cgf.14640

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

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

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

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

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

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

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

[20] C. J. Cho *et al.*, "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)*. IEEE, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1109/ICASSP49357.2023.10094711

[21] ——, "Self-supervised models of speech infer universal articulatory kinematics," 2024.

[22] S. Medina *et al.*, "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

[23] I. Steiner and S. Ouni, "Progress in animation of an ema-controlled tongue model for acoustic-visual speech synthesis," 2012.

[24] R. Li and J. Yu, "An audio-visual 3d virtual articulation system for visual speech synthesis," in *2017 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE)*. IEEE, Oct. 2017. [Online]. Available: https://doi.org/10.1109/have.2017.8240348

[25] S. Chen *et al.*, "Direct, near real time animation of a 3d tongue model using non-invasive ultrasound images," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4994–4998.

[26] K. Zhao *et al.*, "A real-time speech driven talking avatar based on deep neural network," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–4.

[27] K. Richmond *et al.*, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Interspeech*, 08 2011, pp. 1505–1508.

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

[29] S. Chen *et al.*, "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. [Online]. Available: https://doi.org/10.1109/jstsp.2022.3188113

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

[31] Y. Shi *et al.*, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6783–6787.

[32] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[33] M. Morrison *et al.*, "Chunked autoregressive gan for conditional waveform synthesis," in *Submitted to ICLR 2022*, April 2022.

[34] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle *et al.*, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf

[35] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 34513460, oct 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

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

[37] M. M. Design, "Male head," https://www.turbosquid.com/3d-models/male-head-obj/346686, 2007.

[38] S. Greenberg *et al.*, "Temporal properties of spontaneous speech—a syllable-centric perspective," *Journal of Phonetics*, vol. 31, no. 3, pp. 465–485, 2003, temporal Integration in the Perception of Speech. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447003000603