Bluemarble: Bridging Latent Uncertainty in Articulatory-to-Speech Synthesis with a Learned Codebook

Seyun Um*, Miseul Kim*, Doyeon Kim*, Hong-Goo Kang*
* Department of Electrical and Electronic Engineering , Yonsei University, Seoul, Korea
E-mail: syum, miseul4345, ehyeon24@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

***Abstract*—In this paper, we present Bluemarble, a novel neural articulation-to-speech (ATS) system designed to generate high-quality speech from articulatory recordings. Traditional ATS methods often demand a substantial dataset of corresponding articulatory and speech signal pairs for effective model training. This requirement arises because these approaches treat the estimation of acoustic features as a regression task, which can pose difficulties for accurately mapping latent articulatory features to the desired speech-related features. To tackle this challenge, we propose a method that involves employing a finite learned codebook to constrain the uncertainty latent space. Operating within a framework consisting of a speech mel-spectrogram encoder, quantizer, and decoder, our model establishes a mapping rule between the latent embeddings derived from electromagnetic articulography (EMA) signals and quantized discrete speech tokens. During the inference stage, EMA embeddings are first transformed into speech-related discrete tokens, which are then input into a neural vocoder to synthesize the speech waveform. Experimental results show that our approach outperforms existing state-of-the-art methods in both qualitative and quantitative assessments. Audio samples are available online.**[1]

I. Introduction

Articulation-to-speech (ATS), the task of utilizing articulatory movements in speech synthesis, is a field of research that holds significant importance with numerous practical applications [1]. It provides essential communication tools for individuals facing challenges in producing audible speech freely, such as those who have undergone laryngectomy. In addition, ATS systems can play a crucial role in establishing a connection between neural activity in the brain and speech signals, making it a valuable intermediate step in the development of brain-to-speech systems [2]–[4].

Previous works have used a wide variety of approaches for modeling ATS systems. Gaussian mixture model (GMM)-based ATS synthesizers have been designed to estimate the vocal tract spectrum from articulatory movements [5]. Hidden Markov model (HMM)-based methods have also been adopted for this task, combined with a codebook-based network [6]. More recently, the emergence of deep neural models has led to significant improvements in the speech synthesis quality of ATS systems compared to traditional statistical-based approaches [7]–[11]. For example, ATS models often incorporate fully connected layers along with uni- and bi-directional LSTM-based networks [7], [12]. The inclusion of attention-based transformer modules [13] have further contributed to advances in the field; in [8], transformer blocks are employed to estimate mel-frequency cepstral coefficients (MFCCs) from articulatory features. Based on the CARGAN vocoder [14], [9] directly predicts waveforms from EMA signals without intermediate acoustic features using an adversarial training criterion. Text-to-speech (TTS)-based methods have been adopted [10] to extract pitch and energy information from articulatory recordings and generate high-quality mel-spectrograms. Voice conversion-based schemes have also been utilized to synthesize speech from articulated content using the voices of different speakers for vocalization [11].

Although the aforementioned frameworks have demonstrated impressive speech synthesis capabilities, they have typically required a substantial number of parallel articulation and speech recording pairs for model training. Furthermore, they often treat the ATS task as a regression problem, minimizing the mean-squared error (MSE) between estimated and target acoustic features. Mean-squared error primarily aims to minimize the energy difference between two features, resulting in a lack of fine-grained feature estimation (coarse granularity) in complex latent domains [15]–[17]. This can make it challenging to precisely reconstruct intricate target acoustic features from the complex latent embeddings.

In this paper, we propose a neural network-based ATS system designed to decode high-quality covert speech from articulatory kinematics by adopting acoustic tokens extracted from pretrained codebooks. Our overall system utilizes two network training flows: one for mel-spectrogram codebook training and another for articulation-to-acoustic token modeling. Initially, acoustic tokens are estimated from a neural speech compression model through adversarial training [18], utilizing mel-spectrograms. Upon the completion of mel-spectrogram codebook training, an electromagnetic articulograph (EMA) encoder is trained to predict the corresponding acoustic tokens from articulatory recordings, estimating the appropriate codebook indices using cross-entropy loss. The mel-spectrograms are then reconstructed using the estimated acoustic tokens. In the final step, raw speech waveforms are generated by passing the synthesized mel-spectrograms through a pretrained vocoder

[1] https://sam-0927.github.io/Bluemarble/

(e.g. HiFi-GAN [19]). The key advantages of our framework can be summarized as follows:

- **Discretized prediction task:** Thanks to the simpler prediction task involving discrete, quantized acoustic tokens rather than regression in a continuous latent space, our model can more effectively generate speech from articulatory movements.
- **Less paired data required:** Our proposed articulation-to-acoustic token modeling framework can be trained using a significantly smaller amount of paired recordings between articulatory and speech signals compared to previous methods.

II. Related work

A. Style modeling for multi-speaker ATS systems

The most closely related work to ours is [10], which addresses the task of estimating ground-truth mel-spectrograms from EMA signals [20]. Our baseline approach is centered around the reconstruction of mel-spectrograms in a multi-speaker setting, simultaneously estimating both the target speaker's speaking style and contextual information from the provided EMA signals. The entire framework is built using modules that are integrated using convolutional neural network (CNN) and self-attention-based layers. It then utilizes the HiFi-GAN neural vocoder [19] to reconstruct raw speech waveforms from mel-spectrograms.

B. Acoustic feature tokenization in generative models

There has recently been an increasing focus on utilizing discrete acoustic tokens as speech representations for generative speech modeling [21], [22]. The key idea is to predict a sequence of tokens from a codebook, which can then be decoded into more concrete speech representations (e.g. raw audio). High-quality speech can be generated by splitting the task into multiple stages that model semantic and acoustic details successively. This approach simplifies the generation task due to the inherent constraints on the codebook's size, and allows models to be trained using abundant unlabeled audio-only data. AudioLM [23], WavLM [24], CLaM-TTS [25], and LM-VC [26] are key examples of this modeling approach. SPEAR-TTS [27] is an extension to this approach that learns a mapping between text and acoustic tokens to perform TTS. This work leverages the benefits of the aforementioned token-based acoustic modeling in the development of an articulation-to-speech framework.

III. Proposed model

In this section, we provide a comprehensive overview of our proposed network, Bluemarble. Our model employs a two-step learning process to predict mel-spectrograms from the EMA signals, shown in Figure 1. First, we train an autoencoder-based neural speech compression model to learn a codebook of acoustic tokens. These tokens are used to encode a mel-spectrogram into a set of discrete representations. The autoencoder's decoder is trained to reconstruct the mel-spectrogram given a set of codebook indices. Second, we train an EMA encoder to predict discrete acoustic tokens in the codebook from articulatory signals while freezing the codebook and mel-spectrogram decoder. During inference, we input the acoustic tokens predicted by the EMA encoder into the pretrained decoder to generate the target mel-spectrogram. Then, we synthesize a speech waveform from the mel-spectrogram using a pretrained HiFi-GAN vocoder [19].

A. Mel-spectrogram codebook training

Since speech signals contain complex and intertwined information, including content and speaker characteristics, it can be challenging to directly predict these intricate continuous spaces from articulatory signals. We alleviate this problem by changing the prediction domain from a continuous latent space to a discrete latent space, bridging the two distinct domains with reduced uncertainty.

In our proposed model, we encode the mel-spectrograms $S_{1:T} = s_1, ..., s_T$, where $s \in \mathbb{R}^{1\times F_s}$, of a target speech signal to create a discrete codebook. The model has an autoencoder structure consisting of encoder $\mathbf{E_{Mel}}$, codebook $\mathbf{Q}$, and decoder $\mathbf{D_{Mel}}$ [28]. Both the encoder and decoder that down-samples and up-samples the acoustic features [29] are equipped with an attention module and residual blocks. The encoder in our model is comprised of five residual blocks, each consisting of a 2D convolutional layer with dropout and group normalization, an attention layer, and five additional 2D convolutional layers. The decoder is composed of seven residual blocks, an attention layer, and five 2D convolutional layers. The encoder estimates hidden features, denoted as $\hat{Z} \in \mathbb{R}^{F'\times T'\times d}$, from the mel-spectrogram and subsequently converts them into discrete tokens Z_q through a codebook, where d represents the dimension of the codebook entries, and $F' \times T'$ corresponds to a reduced frequency and time dimension. Meanwhile, the decoder predicts mel-spectrograms based on the quantized representations.

$$\begin{aligned} \hat{Z} = \mathbf{E_{Mel}}(S_{1:T}), \quad Z_q = \mathbf{Q}(\hat{Z}), \\ \hat{S}_{1:T} = \mathbf{D_{Mel}}(Z_q). \end{aligned} \tag{1}$$

We train our model using the LSGAN [30] method to enhance the intelligibility of the reconstructed output. The discriminator is built with a 2D convolutional layer featuring skip connections and the Exponential Linear Unit (ELU) activation function. The overall loss for training the autoencoder is defined as follows:

$$\mathcal{L}^{AE}_{total} = \mathcal{L}_G + \mathcal{L}_D + \mathcal{L}_{VQ}, \tag{2}$$

where $\mathcal{L}_G$ and $\mathcal{L}_D$ are generator and discriminator losses and the VQ loss is defined as:

$$\mathcal{L}_{VQ} = \lambda||\hat{Z} - sg(Z_q)||^2_2 + ||sg(\hat{Z}) - Z_q||^2_2, \tag{3}$$

where λ is 0.25. The sg means stop gradient.

A crucial consideration in the process of codebook design is how to determine the number of codebook entries. As the codebook size expands, the task of accurately predicting acoustic tokens from EMA signals becomes progressively

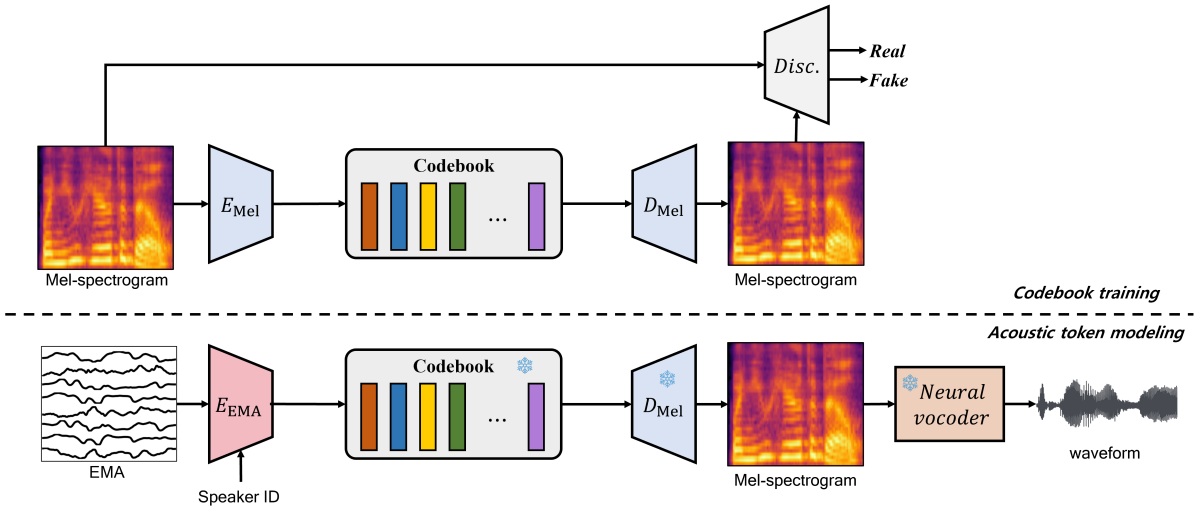

Fig. 1: Architecture of Bluemarble. During codebook training, the target mel-spectrogram is fed into the mel encoder, and it is converted to discrete acoustic tokens by utilizing vector quantization (VQ) codebooks. These tokens are subsequently decoded back into the original mel-spectrogram. The codebook training is carried out concurrently with the reconstruction of the input mel-spectrogram. During acoustic token modeling, the electromagnetic articulography (EMA) encoder utilizes speaker embeddings to predict the target indices of acoustic tokens in the pretrained codebook. The snow symbol means that the module's weights are frozen.

more challenging. Conversely, when the codebook size is too small, the decoder may encounter difficulties in learning how to reconstruct high-quality acoustic features. Taking these factors into account, we chose a codebook size of 32. We conducted an ablation experiment to assess how variations in codebook size affect the model's performance in Section IV.

B. Articulatory to acoustic token modeling

After the codebook has been trained, we proceed to freeze it, as well as the decoder. We then train an encoder that takes in EMA signals and predicts the acoustic tokens for the corresponding target acoustic feature. Constructing a codebook as we do above enables us to map EMA signals to discrete acoustic tokens instead of directly predicting acoustic features.

To predict acoustic tokens, we employ a 2D convolutional module with residual connections and group normalization, maintaining a structure identical to that of the mel encoder. We also apply a 1D convolution to the EMA signal input $A_{1:T} = \{a_1, a_2, ..., a_T\}$, where $a \in \mathbb{R}^{1\times F_a}$, with the aim of increasing the channel size to align it with the target mel-spectrogram. Additionally, we incorporate speaker information into the model because the characteristics of articulatory signals differ from one person to another. The EMA encoder processes the input to predict the acoustic token index within the pretrained codebook, then the decoder estimates the acoustic features.

$$\hat{Z}' = \mathbf{E_{EMA}}(EMA_{1:T} + e_{spk}), \quad e_{spk} = \mathbf{H}(i), \tag{4}$$

where e_{spk} denotes speaker embedding created by the fully connected layer $\mathbf{H}$. We train the EMA encoder using cross-entropy loss between predicted and target acoustic token indices.

$$\mathcal{L}_{CE} = -\frac{1}{L}\sum_{i=1}^{L} y_i \log(\mathbf{F}(\hat{Z}')_i), \tag{5}$$

where $\mathbf{F}$ denotes the classifier consisting of a fully connected layer and softmax. y and L denote the target and codebook size, respectively.

IV. Experimental Setup

A. Dataset

We utilize the publicly available Haskins dataset **haskins**, which consists of paired speech and EMA signals collected from 4 male and 4 female speakers. Each speaker's data covers approximately 55 minutes of recordings, with the EMA signals recorded at a sampling rate of 100 Hz and the speech signals at 44.1 kHz. We excluded silence regions and applied a pretrained denoising model [31] among speech enhancement models [32]–[35] to eliminate background noise from all speech samples. For computational convenience, we downsampled the speech samples from 44.1 kHz to 16 kHz. For further processing, the speech data was transformed into 40-dimensional mel-spectrograms using Short-Time Fourier Transforms (STFTs) at intervals of 10 ms, using a 25 ms window. The EMA data is represented as an 18-dimensional vector, where each dimension corresponds to one of the 6 sensors (TR, TB, TT, UL, LL, jaw), and each sensor comprises 3 trajectory orientations (X: posterior to anterior, Y: right to left, Z: inferior to superior).

In Table I, both our model and baseline models undergo training with a paired dataset spanning 5 hours. In contrast, we train the codebook and decoder using a speech-only dataset lasting 5 hours, while the EMA encoder is trained using paired datasets of 5 and 1 hour as detailed in Section V-C.

TABLE I: Objective test results.

Model	Objective	
	NISQA-T↑	NISQA-E↑
Reference	2.63±0.02	3.86±0.03
[8] (ACL 21')	2.08±0.02	2.9±0.03
[9] (INTERSPEECH 22')	1.84±0.02	2.22±0.02
[10] (ICASSP 23')	2.29±0.02	3.02±0.03
Bluemarble (Ours)	**2.66±0.02**	**4.22±0.02**

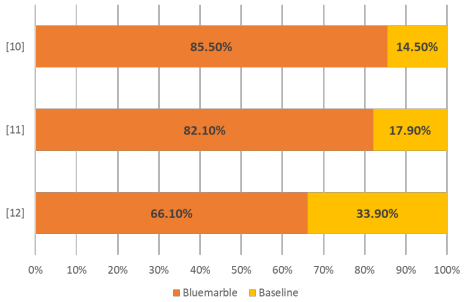

Fig. 2: A/B preference test results.

B. Training

We performed training on a single NVIDIA RTX 3090 GPU for a total of 150 epochs. We employed the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.9$, and $\epsilon = 1e-8$) for training both the EMA encoder and EMA discriminator, setting their respective learning rates to $1e-4$ and $1e-5$. We utilized a HiFi-GAN neural vocoder pre-trained on the VCTK dataset [36].

V. Experimental Results

A. Objective measurement

To evaluate the clarity and overall quality of synthesized speech, we use speech quality assessment metrics, NISQA-TTS (NISQA-T) [37] and NISQA-Enhance (NISQA-E) [38]. NISQA-TTS is relevant to evaluating the naturalness of the synthesized speech, while NISQA-Enhance focuses on measuring fundamental audio clarity and quality. As shown in Table I, our proposed model shows superior performance compared to all baseline models. Notably, Bluemarble achieves NISQA-TTS scores of 2.66 vs. 2.29 and NISQA-Enhance scores of 4.22 vs. 3.02 compared to the best baseline [10], demonstrating its capabilities for generating high quality speech that is both natural and intelligible. Furthermore, these results show that our framework of predicting discrete acoustic tokens from a codebook, rather than directly estimating acoustic features, significantly enhances ATS performance.

B. Subjective measurement

To assess generated sound quality as perceived by human listeners, we conducted an A/B preference listening test between each baseline and our proposed model. 14 listeners participated in the test. After listening to pairs of speech samples, consisting of one from Bluemarble and one from a baseline, participants were asked to choose the preferred speech sample based on the quality of sound. Results are shown in Figure 2. For all comparisons, participants showed much higher preference for samples synthesized by Bluemarble, demonstrating its superiority in terms of subjective evaluation in addition to objective metrics.

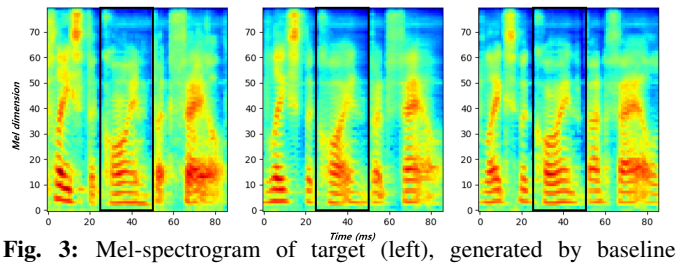

Fig. 3: Mel-spectrogram of target (left), generated by baseline (middle) and Bluemarble (right). The Bluemarble's sample has better quality than that of the baseline. The distinct pitch is especially evident in the region delineated by a black box. To facilitate a detailed comparison, we generate a mel-spectrogram of synthesized speech using an 80-dimensional resolution from a pre-trained vocoder, as opposed to the 40-dimensional output from the decoder.

TABLE II: Performance evaluation in terms of a codebook size and training dataset size.

Codebook	Data	NISQA-T↑	NISQA-E↑
32	5h	**2.66±0.02**	**4.22±0.02**
64	5h	2.59±0.02	4.10±0.02
16	5h	2.63±0.02	4.15±0.02
32	1h	**2.55±0.02**	**4.17±0.03**
64	1h	2.52±0.02	3.77±0.03
16	1h	**2.55±0.02**	4.08±0.03

C. Ablation study

To assess the influence of codebook and training dataset size on Bluemarble's performance, we conducted model training by varying codebook and training dataset size.

Codebook size. As shown in Table II, compared to the standard model with codebook size 32, model performance decreases when adopting smaller (16) or larger (64) codebook sizes. A larger codebook size allows for a more comprehensive representation of the signal, facilitating the reconstruction of acoustic features during modeling. However, it also means that the EMA encoder must learn to predict a larger number of tokens, which is more difficult. We conjecture that this caused a reduction in the quality of synthesized speech due to the more challenging token prediction task. Conversely, when the codebook size is too small, the acoustic tokens fail to adequately encode all of the information in the acoustic features, leading to degradation in reconstructing them. These findings highlight the importance of selecting an appropriate codebook size that strikes a balance between encoding sufficient information to reconstruct the mel spectrogram features and allowing the model to effectively predict acoustic tokens for high-quality speech synthesis.

Dataset size. Table II shows the NISQA-T and NISQA-E scores for Bluemarble trained on 1 hour vs. 5 hours of paired EMA and speech data. We can see that our model still demonstrates reasonable performance even when trained with only 1 hour of data, achieving 2.55 and 4.17 NISQA scores. This demonstrates that decoupling the training of the codebook and mel-spectrogram autoencoder with that of the

EMA encoder allows for high-quality speech synthesis even with only a small amount of paired data.

VI. Conclusion

In this paper, we proposed Bluemarble, an articulation-to-speech (ATS) system that uses electromagnetic articulography (EMA) signals. Utilizing concepts from recent trends in generative audio modeling, our model predicts a sequence of discrete acoustic tokens in a codebook rather than predicting continuous acoustic features. Specifically, it learns a mapping from the EMA signals to discrete latent acoustic embeddings represented by vector quantization codebooks, which are then decoded to produce mel-spectrograms. Experiments demonstrate that our model surpasses previous state-of-the-art models in terms of the clarity and naturalness of synthesized speech, as confirmed by both objective and subjective evaluation metrics.

VII. Acknowledge

This work was supported by the 'Alchemist Project' (Fully implantable closed loop Brain to X for voice communication) funded By the Ministry of Trade, industry & Energy (MOTIE, Korea), under Grant 20012355 and NTIS 1415181023.

References

[1] B. Cao, M. J. Kim, J. R. Wang, J. P. van Santen, T. Mau, and J. Wang, "Articulation-to-speech synthesis using articulatory flesh point sensors' orientation information.," in *INTERSPEECH*, 2018, pp. 3152–3156.

[2] P. Li, S. Cai, E. Su, and L. Xie, "A biologically inspired attention network for eeg-based auditory attention detection," in *IEEE Signal Processing Letters*, 2022, pp. 284–288.

[3] M. Angrick, C. Herff, E. Mugler, *et al.*, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, p. 036 019, 2019.

[4] M. Kim, Z. Piao, J. Lee, and H.-G. Kang, "Braintalker: Low-resource brain-to-speech synthesis with transfer learning using wav2vec 2.0," in *BHI*, 2023, pp. 1–5.

[5] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *ISCA Workshop on Speech Synthesis*, 2004, pp. 31–36.

[6] S. Hiroya and M. Honda, "Determination of articulatory movements from speech acoustics using an hmm-based speech production model," in *ICASSP*, 2002, pp. I-437-I–440.

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

[8] D. Gaddy and D. Klein, "An improved model for voicing silent speech," in *ACL/IJCNLP*, 2021, pp. 175–181.

[9] P. Wu, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Deep speech synthesis from articulatory representations," in *INTERSPEECH*, 2022, pp. 779–783.

[10] M. Kim, Z. Piao, J. Lee, and H.-G. Kang, "Style modeling for multi-speaker articulation-to-speech," in *ICASSP*, 2023, pp. 1–5.

[11] K. Scheck and T. Schultz, "Multi-speaker speech synthesis from electromyographic signals by soft speech unit prediction," in *ICASSP*, 2023, pp. 1–5.

[12] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, vol. 385, pp. 37–45, 2012.

[13] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Proc. NIPS*, vol. 30, 2017.

[14] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive gan for conditional waveform synthesis," in *ICLR*, 2021.

[15] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *ICLR*, 2015.

[16] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *ICLR*, 2016.

[17] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," *NIPS*, vol. 28, 2015.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139–144, 2020.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *NIPS*, vol. 33, pp. 17 022–17 033, 2020.

[20] T. Rebernik, J. Jacobi, R. Jonkers, A. Noiray, and M. Wieling, "A review of data collection practices using electromagnetic articulography," *Laboratory Phonology*, vol. 12, p. 6, 2021.

[21] K. Takagi, T. Akiba, and H. Tsukada, "Semi-supervised asr based on iterative joint training with discrete speech synthesis," in *APSIPA ASC*, 2022, pp. 923–928.

[22] C. Yuan and Y.-C. Huang, "Personalized end-to-end mandarin speech synthesis using small-sized corpus," in *APSIPA ASC*, 2020, pp. 837–840.

[23] Z. Borsos, R. Marinier, D. Vincent, *et al.*, "Audiolm: A language modeling approach to audio generation," *ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023.

[24] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2022.

[25] Anonymous, "CLam-TTS: Improving neural codec language model for zero-shot text-to-speech," in *ICLR*, 2023.

[26] Z. Wang, Y. Chen, L. Xie, Q. Tian, and Y. Wang, "Lm-vc: Zero-shot voice conversion via speech generation based on language models," in *IEEE Signal Processing Letters*, 2023, pp. 1157–1161.

[27] E. Kharitonov, D. Vincent, Z. Borsos, *et al.*, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1703–1718, 2023.

[28] V. Iashin and E. Rahtu, "Taming visually guided sound generation," in *BMVC*, 2021.

[29] J. Kim and H.-G. Kang, "Contrastive learning based deep latent masking for music source separation," in *INTERSPEECH 2023*, 2023, pp. 3709–3713.

[30] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[31] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *INTERSPEECH*, 2020, pp. 3291–3295.

[32] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *APSIPA ASC*, 2017, pp. 006–012.

[33] S.-W. Fu, C. Yu, T.-A. Hsieh, *et al.*, "Metricgan+: An improved version of metricgan for speech enhancement," in *Interspeech 2021*, 2021, pp. 201–205.

[34] D. Kim, S.-W. Chung, H. Han, Y. Ji, and H.-G. Kang, "Hd-demucs: General speech restoration with heterogeneous decoders," in *INTERSPEECH 2023*, 2023, pp. 3829–3833.

[35] D. Kim, H. Han, H.-K. Shin, S.-W. Chung, and H.-G. Kang, "Phase continuity: Learning derivatives of phase spectrum for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6942–6946.

[36] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *The Centre for Speech Technology Research*, vol. 6, p. 15, 2017.

[37] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *INTERSPEECH*, 2020, pp. 1748–1752.

[38] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *INTERSPEECH*, 2021, pp. 2127–2131.