INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

High-Quality Automatic Voice Over with Accurate Alignment: Supervision through Self-Supervised Discrete Speech Units

Junchen Lu[1,2], *Berrak Sisman*[2], *Mingyang Zhang*[3], *Haizhou Li*[3,1]

[1]National University of Singapore, Singapore [2]The University of Texas at Dallas, USA
[3]Shenzhen Research Institute of Big Data, School of Data Science,
The Chinese University of Hong Kong, Shenzhen, China

junchen@u.nus.edu, Berrak.Sisman@UTDallas.edu, {zhangmingyang, haizhouli}@cuhk.edu.cn

Abstract

The goal of Automatic Voice Over (AVO) is to generate speech in sync with a silent video given its text script. Recent AVO frameworks built upon text-to-speech synthesis (TTS) have shown impressive results. However, the current AVO learning objective of acoustic feature reconstruction brings in indirect supervision for inter-modal alignment learning, thus limiting the synchronization performance and synthetic speech quality. To this end, we propose a novel AVO method leveraging the learning objective of self-supervised discrete speech unit prediction, which not only provides more direct supervision for the alignment learning, but also alleviates the mismatch between the text-video context and acoustic features. Experimental results show that our proposed method achieves remarkable lip-speech synchronization and high speech quality by outperforming baselines in both objective and subjective evaluations. Code and speech samples are publicly available.

Index Terms: Text-to-speech, lip-speech synchronization, automatic voice over, discrete speech units, speech synthesis

1. Introduction

Automatic Voice Over is a cutting-edge technology that utilizes artificial intelligence to generate speech that voice-synchronizes with a pre-recorded video [1]. AVO technology enables the automatic creation of a voice track that is perfectly aligned with the lip movement, facial expression, and conversational tone of the video, provided with text transcription. As a highly efficient AI-powered solution for voice over, AVO has the potential to revolutionize video-making in various industries, including movie dubbing, online education, and marketing.

The development of neural text-to-speech synthesis has played a crucial role in the advancement of AVO technology. In light of rapid emergence of deep learning, TTS systems built upon neural networks can generate high-quality speech [2]. End-to-end TTS systems, including Tacotron 1/2 [3, 4] and FastSpeech 1/2 [5, 6], work in a simplified pipeline of mapping sequences of character or phoneme input into mel-spectrogram acoustic features. With the help of neural vocoders [7, 8, 9], they generate speech with human-level naturalness.

With the ability to generate high-quality speech, neural TTS provides a foundation for AVO systems to produce accurate and natural-sounding voice over. To model speech with precise timing in sync with the video, AVO systems generally use lip motion or facial movement to guide the rendering of phonetic duration in TTS [1, 10, 11], as illustrated in Figure 1. Existing AVO approaches use attention-based alignment modules to align multi-modal features and produce text-video context for acoustic feature decoding. However, as a learning objective of an AVO system, acoustic feature reconstruction does not provide direct supervision for the model to learn accurate inter-modal alignment, thus affecting the acoustic decoding and leading to degradation of synthetic speech quality.

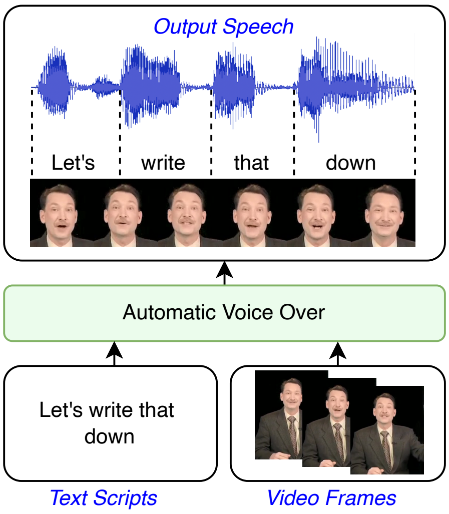

Figure 1: *An illustration of the AVO workflow: The input to the system consists of video frames and corresponding text scripts, and the output is voice over speech audio in synchronization with the video.*

Self-supervised learning (SSL) speech models trained on large amounts of unlabeled data are proved to be sufficient for capturing content information of speech [12, 13, 14, 15]. Recent studies show that discrete units derived from SSL speech models can be applied as speech content representation for speech synthesis tasks, including speech resynthesis [16] and voice conversion [17]. As speech content in AVO is modeled through the alignment of multi-modal information, content representation can provide more direct supervision for alignment learning.

Motivated by these, we propose a novel AVO approach leveraging discrete speech units: first, align multi-modal features and predict discrete speech units from the text-video context formed; then, synthesize speech conditioned on the predicted units. The main contributions of this paper include:

- We propose a new learning objective of discrete speech unit prediction for AVO, providing more direct supervision for

Code and voice over samples: https://ranacm.github.io/DSU-AVO/
This work is supported by 1) National Natural Science Foundation of China (Grant No. 62271432); 2) Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen (Grant No. B10120210117-KP02); and 3) Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A2b0046).

10.21437/Interspeech.2023-2179

text-video alignment learning at the context feature level, thus improving lip-speech synchronization;

- We propose to synthesize speech directly from discrete speech units with a pretrained unit vocoder, thus alleviating mismatch between text-video context and acoustic features in acoustic decoding and improving synthetic speech quality.

2. Related work

2.1. AVO Background

The industry of video content creation calls for high-quality and cost-efficient voice over solutions. As such, AVO research, drawing upon TTS, has become increasingly important. While general-purpose TTS systems generate speech from text, AVO systems need to produce speech that is not only natural-sounding but also synchronized with visual presentation of speakers in video. With this requirement, AVO systems have to model the timing of the speech, taking visual context of the video into account.

Generally, AVO frameworks consist of text and video encoders, which encode information from multi-modal inputs, an alignment module that models inter-modal feature relationships, an acoustic decoder that constructs acoustic features such as mel-spectrogram, and a vocoder to convert acoustic features into output speech waveform [1, 10, 11]. Existing approaches of alignment module are based on attention mechanism [18]. VisualTTS [1] and Neural Dubber [10] align textual and visual features explicitly by attention-based aligners and form text-video context which serves as the input to acoustic decoder. VDTTS [11] uses a multi-source attention mechanism for selecting which outputs of video encoder and text encoder to pass to acoustic decoder at each decoding timestep, forming the alignment between modalities in an implicit manner.

The existing AVO frameworks have limitations that hinder their ability to produce high-quality voice over. Firstly, the capability of the model to learn inter-modal alignment is limited to the acoustic decoding process with the learning objective of mel-spectrogram reconstruction. To establish the alignment, the entire AVO model must be trained as a whole, which is computationally expensive and time-consuming. Secondly, the supervision on the acoustic feature level is indirect for alignment learning due to the gap between phonetic information presented in input text and the acoustic features to model. Moreover, the mismatch between the context representation modeled by the alignment module and target acoustic features can negatively impact acoustic decoder training, resulting in degradation of synthetic speech quality. In this paper, we aim to address these challenges in AVO by imposing supervision on the context representation level with a learning objective of discrete speech unit prediction, which will be introduced in Sec. 3.

2.2. Self-supervised learning in speech synthesis

Self-supervised learning has emerged as a powerful approach to learning speech representations. SSL speech models leverage large amounts of unlabeled speech data by defining auxiliary tasks and generating pseudo-labeled data to be trained using supervised learning techniques [12, 13, 14]. The pretrained models can be used for various downstream tasks [19, 20], such as automatic speech recognition (ASR) [15, 21] and speech emotion recognition [22], and speech enhancement [23].

Recent studies have also explored the use of SSL speech representations in speech synthesis tasks. Du et al. [24] propose to reduce the complexity of the acoustic model in TTS with SSL vector-quantized acoustic features as its classification target. Huang et al. [25] propose to use discrete speech representations as a bottleneck to disentangle content information from speaker information and model acoustic features on top of them for voice conversion. Polyak et al. [16] demonstrate that discrete units derived from HuBERT models [14] can serve as the input of high-quality speech waveform generation.

Inspired by these, we propose to use discrete speech units as the content representation for speech generation and the supervision of alignment learning in AVO.

3. Proposed method

We formulate the AVO problem and propose *Discrete Speech Unit-AVO (DSU-AVO)*, with motivation of providing more direct supervision to alignment learning and establishing a strong connection between text-video context and speech in AVO.

3.1. Problem formulation

Given input phoneme sequence $X_p = (x_{p_1}, x_{p_2}, ..., x_{p_N})$ with length N and video represented by a sequence of image frames $X_v = (x_{v_1}, x_{v_2}, ..., x_{v_{T_v}})$ with length T_v, the goal of AVO is to generate speech audio that accurately reflects the phonetic content and is temporally aligned with the video.

In a voiced video clip, speech audio and video are both continuous signals of the same length. In practice, they are sampled at different frame rates. This allows us to encode ground-truth speech Y as a sequence of speech representation Z with length T_z that can be easily aligned with the video frame sequence X_v by upsampling the latter, given that the length of the speech representation sequence is n times that of the video frame sequence, where $n = \frac{T_z}{T_v} \in \mathbb{N}^+$.

Given the temporal correspondence of speech audio and video, the synchronization between speech and lip motion can be achieved by aligning phoneme information, which determines the speech content, with the lip motion information [1, 10]. An AVO framework aims to model accurate text-video alignment and produce context representation $C = f(X_p, X_v)$ with length T_v. Then, conditioned on the context and given the audio-video length ratio n, the framework generates speech representation $\hat{Z} = g(C, n)$. Finally, $\hat{Z}$ is converted to speech waveform $\hat{Y}$ through a pretrained vocoder. Training of an AVO framework can be seen as finding the optimal context modeling $f(\cdot)$ and speech representation generation $g(\cdot)$.

3.2. Supervision at context representation level with discrete speech units

As speech representation generation in AVO is conditioned on text-video context, context modeling $f(\cdot)$ is crucial for a framework to produce speech with accurate pronunciation and timing. Since the context is formed by aligning multi-modal representations, alignment learning is an integral part of context modeling.

Existing AVO frameworks [1, 10, 11] rely on acoustic features, typically mel-spectrogram, as the speech representation Z to model, utilizing acoustic decoders as $g(\cdot)$. With the learning objective of mel-spectrogram reconstruction, these frameworks guide the alignment learning through supervision at the acoustic feature level. However, mel-spectrogram does not directly capture linguistic information [4, 26], posing a gap between context and speech representation, thus providing indirect supervision for alignment learning in AVO.

We propose to guide the context modeling and alignment learning of AVO more directly by imposing discrete speech unit prediction as the supervision at the context representation level, given that discrete speech units are closely correlated with speech content. Additionally, compared with mel-spectrogram, discrete speech units are more disentangled from nuisance variation [16] and are easier to predict. With the proposed learning objective, $g(\cdot)$ essentially becomes a classification model instead of a regression model and gains better training efficiency. To be specific, we first encode inputs from different modalities, align inter-modal representations to form context, and predict discrete speech units; then, synthesize speech conditioned on the predicted units with a pretrained unit vocoder.

3.3. DSU-AVO system

As demonstrated in Figure 2, our proposed DSU-AVO consists of unit tokenizer, video encoder, text encoder, video-text aligner, unit predictor, and unit vocoder.

3.3.1. Unit tokenizer

A HuBERT model [14] followed by k-means clustering is used as the unit tokenizer to encode ground-truth speech into a sequence of discrete speech units as the prediction target $Z = (z_1, z_2, ..., z_{T_z})$, where $z_i \in \{0, 1, ..., K-1\}$ for $1 \leq i \leq T_z$ and K is number of k-means centroids. The unit tokenizer is pretrained and frozen during DSU-AVO training.

3.3.2. Encoders

As speech progression is inherently linked with lip motion in real life [1, 27], we use lip image sequence cropped from the video clip as the input X_v. The video encoder consists of a frozen visual feature extractor and feed-forward Transformer (FFT) [5, 10, 18] blocks. Visual features extracted by such an extractor pretrained on visual speech recognition tasks have strong correlation with phonetic information [27] and are efficient for aiding speech-related tasks [28, 29]. Input X_v is encoded into hidden visual representation $H_v \in \mathbb{R}^{T_v \times d}$ where d is the dimension of hidden representations.

We adopt the same text encoder that is used in FastSpeech 2 [6] for TTS and in Neural Dubber [10] for AVO. It consists of an embedding layer followed by FFT blocks, to encode input phoneme X_p into hidden textual representation $H_p \in \mathbb{R}^{T_p \times d}$.

3.3.3. Text-video aligner

We utilize a text-video aligner [1, 10] to temporally align textual and visual representations by scaled dot-product attention [18], and produce text-video context with length T_v:

$$C = \text{softmax}(\frac{H_v H_p^T}{\sqrt{d}})H_p + H_v \quad (1a)$$

$$= AH_p + H_v \in \mathbb{R}^{T_v \times d} \quad (1b)$$

where H_v serves as the query, H_p serves as the key and the value, and $A \in \mathbb{R}^{T_v \times T_p}$ is the attention weight matrix. H_v is added through residual connection to enhance alignment learning. C is then upsampled to match T_z by simply duplicating each frame of representation n times, where n is the audio-video length ratio.

We adopt the diagonal constraint loss $\mathcal{L}_{diag}$ following [10, 30] to shape diagonal attention.

3.3.4. Unit predictor

The accuracy of the prediction has a direct impact on the content correctness and intelligibility of the synthetic speech. A higher prediction accuracy results in a better perceptual quality of the synthetic speech. Hence, DSU-AVO focuses on predicting accurate discrete speech units.

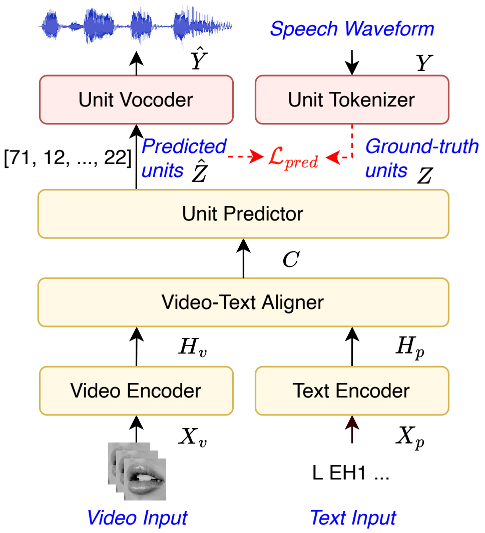

Figure 2: *The model architecture of the proposed DSU-AVO. Modules denoted with red color are pretrained and frozen during DSU-AVO training. Dotted arrows denote loss calculation.*

The unit predictor consists of an FFT block that further models the context into a more deterministic representation for accurate unit prediction, followed by a softmax layer that maps output of the FFT block onto a probability distribution over a set of output classes, i.e., the discrete speech units $\{0, 1, ..., K-1\}$. Taking the context C produced by the text-video aligner as input, this module predicts a sequence of discrete speech units $\hat{Z} = (\hat{z_1}, \hat{z_2}, ..., \hat{z}_{T_z})$ with the same length of ground-truth discrete speech units Z. The prediction is guided by minimizing the cross entropy loss $\mathcal{L}_{pred}$:

$$\mathcal{L}_{pred} = -\sum_{t=1}^{T_z} \sum_{k=0}^{K-1} \mathbf{z}_{t,k} \text{log} p_{t,k}(C) \quad (2)$$

where $\mathbf{z}_t$ is the one-hot vector representing speech unit truth label at the t-th frame, and $p_{t,k}(C)$ is the predicted softmax probability for unit k at the t-th frame, given context C. Finally, the overall training criterion is $\mathcal{L} = \mathcal{L}_{pred} + \mathcal{L}_{diag}$.

3.3.5. Unit vocoder

We utilize Unit HiFi-GAN [16], which is pretrained on ground-truth <units, waveform> pairs without the speaker encoder and the F0 encoder in a single-speaker setting [31], as the unit vocoder. Given predicted units $\hat{Z}$, it generates speech audio as the output $\hat{Y}$ of AVO. As Unit HiFi-GAN models the mapping from discrete speech units to speech waveform without any spectrogram estimation [16], DSU-AVO does not need to model acoustic features for speech generation, thus alleviating the mismatch between context and acoustic features.

4. Experiments

4.1. Experimental setup

4.1.1. Dataset and preprocessing

We utilize Chem dataset [10] from Lip2Wav [32] to evaluate the performance of AVO frameworks. Chem dataset is a single-speaker audio-visual English speech dataset with official transcripts collected from YouTube. We use 6088 samples for train-

Table 1: *Evaluation results of LSE-C, LSE-D, FD, WER, and MOS (with 95% confidence intervals). Arrows indicate whether higher or lower metric values are better.*

Method	LSE-C ↑	LSE-D ↓	FD ↓	WER(%) ↓	MOS ↑
Ground Truth	7.00	7.31	NA	11.4	4.69 ± 0.06
Mel Resynthesis	6.89	7.40	0.66	11.8	4.58 ± 0.07
Unit Resynthesis	6.99	7.39	0.82	20.6	4.06 ± 0.08
FastSpeech 2 [6]	2.69	11.78	40.38	25.4	3.10 ± 0.09
Neural Dubber [10]	6.11	8.47	9.39	75.8	2.43 ± 0.12
DSU-AVO	**6.81**	**7.56**	**3.23**	**24.7**	**3.98 ± 0.08**

ing, 200 samples for validation, and 200 samples for testing. As a preprocessing step, we follow the same process described in [33] to crop 88 × 88 lip region-of-interest for the video input. All video clips are sampled at 25Hz frame rate, and audio samples are sampled at 16kHz.

4.1.2. Model configurations

We evaluate several systems, including: 1) Mel Resynthesis, where we convert the ground-truth audio into mel-spectrogram and convert it back to waveform with HiFi-GAN[1] [9]; 2) Unit Resynthesis, where we synthesize speech audio conditioned on ground-truth discrete speech units using Unit HiFi-GAN[2] [16]; 3) FastSpeech 2 [6], a TTS baseline[3] that generates speech conditioned only on text, without taking visual information into consideration; 4) Neural Dubber [10], an AVO baseline with the learning objective of mel-spectrogram reconstruction; and 5) DSU-AVO, our proposed framework. As there is no publicly available implementation of Neural Dubber, we implement the framework based on an open-source FastSpeech 2 implementation[3] without the image-based speaker embedding module as we conduct AVO in a single-speaker setting.

System 1), 3), and 4) use the same HiFi-GAN for a fair comparison. System 2) and 5) use the same Unit HiFi-GAN. Both vocoders are trained on Chem dataset. We set the number of FFT blocks in text encoders of 3), 4), and 5) to 4, the one in video encoders of 4) and 5) to 2, the one in decoders of 3) and 4) to 6. d is set to 256 for 3), 4), and 5). For visual feature extractor in both 4) and 5), we use the same AV-HuBERT + Self-Training model[4] [33] pretrained on 1,758h of unlabeled Voxceleb2 data [34] and finetuned on 433h of labeled LRS3 data [35] for visual speech recognition. The extracted visual feature is projected to d dimensions as the input to FFT blocks in video encoder. We note that in our implementation, Neural Dubber and DSU-AVO use identical text encoder, video encoder, and text-video aligner for a fair comparison. In DSU-AVO, we use a HuBERT Base model [14] pretrained on 960h LibriSpeech corpus [36] and an accompanying k-means model trained on LibriSpeech clean-100h dataset [36] as the unit tokenizer, following [16]. Ground-truth speech is encoded to discrete speech units with $K = 100$ centroids at 50Hz.

4.2. Experimental results

4.2.1. Objective evaluation

We measure lip-speech synchronization between the synthetic speech and input video with Lip Sync Error - Confidence (LSE-C) and Lip Sync Error - Distance (LSE-D) [37], using a pretrained SyncNet model [38]. LSE-C denotes the confidence score of audio-video synchronization time offset, where higher values indicate more accurate synchronization. LSE-D measures the distance between audio and video features, where lower values indicate better synchronization. As shown in Table 1, DSU-AVO outperforms the baselines by achieving LSE-C of 6.81, and LSE-D of 7.56.

Table 2: *Evaluation results of the BWS listening test on lip-speech synchronization. N/P stands for no preference.*

Method	Best(%)	Worst(%)	N/P(%)
FastSpeech 2 [6]	4.0	72.0	24.0
Neural Dubber [10]	12.0	26.7	61.3
DSU-AVO	**84.0**	**1.3**	14.7

We utilize Frame Distrubance (FD) [39] to measure the duration deviation between generated speech and ground-truth speech from the test set. Since ground-truth speech is in sync with video, FD also indicates lip-speech synchronization performance for AVO [1]. We note that DSU-AVO exhibits remarkable results and outperforms baselines with an FD of 3.23. LSE-C, LSE-D, and FD results prove the effectiveness of DSU-AVO in alignment learning.

We report the Word Error Rate (WER) obtained by the Wav2Vec 2.0 Large ASR model[5] [13] pretrained and finetuned on 960h LibriSpeech [36] data as an assessment of synthetic speech intelligibility. Note that WER is for relative comparison only, since the ASR model is not finetuned on Chem dataset. DSU-AVO achieves a WER value of 24.7, demonstrating a strong ability to model correct speech content given input text.

4.2.2. Subjective evaluation

Human perception plays a crucial role in evaluating the performance of AVO, as the goal of AVO is to generate speech that appears natural to human observers. We conduct listening experiments, in which 15 listeners participate. In the mean opinion score (MOS) evaluation, each participant listens to 12 speech samples produced by each system and rate the speech audio quality on a five-point scale. As shown in Table 1, our proposed DSU-AVO produces speech with a higher level of naturalness than both baselines by achieving a MOS score of 3.98 ± 0.08.

We also conduct a Best-Worst Scaling (BWS) test [39] on lip-speech synchronization, where each subject watches in total 10 scaling sets of videos and selects the best and the worst lip-speech synchronization from each set. The original pre-recorded speech samples in the test set videos are replaced with synthetic speech samples produced by system 3), 4), and 5). Table 2 shows that DSU-AVO outperforms baselines in terms of lip-speech synchronization, with the highest best votes (84.0%) and the lowest worst votes (1.3%).

5. Conclusions

In this paper, we propose DSU-AVO, a novel AVO approach leveraging discrete speech units as the content representation. Our proposed method not only provides more direct supervision for alignment learning, but also alleviates the mismatch between context and acoustic features. Experimental results show that DSU-AVO outperforms baselines in terms of synthetic speech quality and lip-speech synchronization. In future work, we will investigate further modeling speech expressiveness with SSL speech representations for AVO.

[1] https://github.com/jik876/hifi-gan

[2] https://github.com/facebookresearch/speech-resynthesis

[3] https://github.com/ming024/FastSpeech2

[4] https://github.com/facebookresearch/av_hubert

[5] https://github.com/facebookresearch/fairseq/tree/main/examples

6. References

[1] J. Lu, B. Sisman, R. Liu, M. Zhang, and H. Li, “Visualtts: Tts with accurate lip-speech synchronization for automatic voice over,” in *Proc. ICASSP*, 2022.

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, “A survey on neural speech synthesis,” *arXiv preprint arXiv:2106.15561*, 2021.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards End-to-End Speech Synthesis,” in *Proc. Interspeech 2017*, pp. 4006–4010.

[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech: fast, robust and controllable text to speech,” in *Proc. NeurIPS*, vol. 32, 2019.

[6] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *Proc. ICLR*, 2021.

[7] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[8] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” in *Proc. NeurIPS*, vol. 32, 2019.

[9] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NeurIPS*, vol. 33, 2020.

[10] C. Hu, Q. Tian, T. Li, W. Yuping, Y. Wang, and H. Zhao, “Neural dubber: Dubbing for videos according to scripts,” in *Proc. NeurIPS*, vol. 34, 2021.

[11] M. Hassid, M. T. Ramanovich, B. Shillingford, M. Wang, Y. Jia, and T. Remez, “More than words: In-the-wild visually-driven prosody for text-to-speech,” in *Proc. CVPR*, 2022, pp. 10 587–10 597.

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *Proc. Interspeech 2019*, pp. 3465–3469.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[15] A. Pasad, J.-C. Chou, and K. Livescu, “Layer-wise analysis of a self-supervised speech representation model,” in *Proc. ASRU*, 2021, pp. 914–921.

[16] 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*, pp. 3615–3619.

[17] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, “A comparison of discrete and soft speech units for improved voice conversion,” in *Proc. ICASSP*, 2022.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. NeurIPS*, 2017.

[19] X. Yang, D. Zhou, S. Liu, J. Ye, and X. Wang, “Deep model reassembly,” in *Proc. NeurIPS*, vol. 35, 2022.

[20] X. Yang, J. Ye, and X. Wang, “Factorizing knowledge in neural networks,” in *Proc. ECCV*, 2022.

[21] S.-J. Chen, J. Xie, and J. H. Hansen, “FeaRLESS: Feature Refinement Loss for Ensembling Self-Supervised Learning Features in Robust End-to-end Speech Recognition,” in *Proc. Interspeech 2022*, pp. 3058–3062.

[22] L. Goncalves and C. Busso, “Improving Speech Emotion Recognition Using Self-Supervised Learning with Domain-Specific Audiovisual Tasks,” in *Proc. Interspeech 2022*, pp. 1168–1172.

[23] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, “Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration,” in *Proc. CVPR*, 2023, pp. 18 795–18 805.

[24] C. Du, Y. Guo, X. Chen, and K. Yu, “VQTTS: High-Fidelity Text-to-Speech Synthesis with Self-Supervised VQ Acoustic Feature,” in *Proc. Interspeech 2022*, pp. 1596–1600.

[25] W.-C. Huang, Y.-C. Wu, and T. Hayashi, “Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations,” in *Proc. ICASSP*, 2021, pp. 5944–5948.

[26] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, “Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis,” in *Proc. NeurIPS*, vol. 35, 2022.

[27] H. Chen, J. Du, Y. Hu, L.-R. Dai, B.-C. Yin, and C.-H. Lee, “Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement,” *Neural Networks*, 2021.

[28] Z. Pan, R. Tao, C. Xu, and H. Li, “Muse: Multi-modal target speaker extraction with visual cues,” in *Proc. ICASSP*, 2021, pp. 6678–6682.

[29] ——, “Selective listening by synchronizing speech with lips,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1650–1664, 2022.

[30] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, “Multispeech: Multi-speaker text to speech with transformer,” in *Proc. Interspeech 2020*, pp. 4024–4028.

[31] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, and W.-N. Hsu, “Direct speech-to-speech translation with discrete units,” in *Proc. ACL*, 2022, pp. 3327–3339.

[32] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, “Learning individual speaking styles for accurate lip to speech synthesis,” in *Proc. CVPR*, 2020, pp. 13 796–13 805.

[33] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, “Learning audio-visual speech representation by masked multimodal cluster prediction,” in *Proc. ICLR*, 2022.

[34] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep Speaker Recognition,” in *Proc. Interspeech 2018*, pp. 1086–1090.

[35] T. Afouras, J. S. Chung, and A. Zisserman, “Lrs3-ted: a large-scale dataset for visual speech recognition,” *arXiv preprint arXiv:1809.00496*, 2018.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Proc. ICASSP*, 2015.

[37] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, “A lip sync expert is all you need for speech to lip generation in the wild,” in *Proc. ACM Multimedia*, 2020, pp. 484–492.

[38] J. S. Chung and A. Zisserman, “Out of time: automated lip sync in the wild,” in *Proc. ACCV*, 2017, pp. 251–263.

[39] B. Sisman, J. Yamagishi, S. King, and H. Li, “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.