ACOUSTIC BPE FOR SPEECH GENERATION WITH DISCRETE TOKENS

Feiyu Shen, Yiwei Guo, Chenpeng Du, Xie Chen, Kai Yu[*]

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

{francis_sfy,yiwei.guo,duchenpeng,chenxie95,kai.yu}@sjtu.edu.cn

ABSTRACT

Discrete audio tokens derived from self-supervised learning models have gained widespread usage in speech generation. However, current practice of directly utilizing audio tokens poses challenges for sequence modeling due to the length of the token sequence. Additionally, this approach places the burden on the model to establish correlations between tokens, further complicating the modeling process. To address this issue, we propose *acoustic BPE* which encodes frequent audio token patterns by utilizing byte-pair encoding. Acoustic BPE effectively reduces the sequence length and leverages the prior morphological information present in token sequence, which alleviates the modeling challenges of token correlation. Through comprehensive investigations on a speech language model trained with acoustic BPE, we confirm the notable advantages it offers, including faster inference and improved syntax capturing capabilities. In addition, we propose a novel rescore method to select the optimal synthetic speech among multiple candidates generated by rich-diversity TTS system. Experiments prove that rescore selection aligns closely with human preference, which highlights acoustic BPE's potential to other speech generation tasks.

Index Terms— discrete audio token, byte-pair encoding, language modeling, rescore

1. INTRODUCTION

The emergence of self-supervised learning (SSL) models in the audio domain has introduced a new option for audio feature selection. For instance, wav2vec [1] utilizes a contrastive loss objective to exploit high-level representation from audio signals. HuBERT [2] utilizes a masked prediction objective where discrete targets derive from k-means [3] clustering. W2v-BERT [4] combines both contrastive learning and masked prediction in an end-to-end fashion. Aside from modeling speech content in the audio signal, WavLM [5] encodes speaker-related information through a denoising objective, where the model is asked to predict the pseudo labels of original audio that is overlapped by a noisy or secondary utterance clip. vq-wav2vec [6] and wav2vec 2.0 [7] use vector quantization to learn a discrete representation of audio signals.

Discretization further compresses redundant information and enables the direct application of algorithms from natural language processing (NLP) communities. With k-means clustering or vector quantization, these aforementioned SSL features can be utilized as pseudo text for speech generation purposes. GSLM [8] extracts SSL features from pretrained HuBERT which is then discretized using k-means clustering. These discrete tokens are uti-

lized as inputs to Tacotron2 [9] for synthesizing mel-spectrograms. VQTTS [10, 11, 12] replaces the mel-spectrogram with vq-wav2vec tokens and a 3-dimensional prosody feature to bridge the acoustic model and the vocoder, achieving competitive synthesis quality compared to its mel-spectrogram counterpart. AudioLM [13] trains a language model on tokens discretized by w2v-BERT and k-means clustering, which demonstrates the effectiveness of discrete audio tokens as textual units in language modeling. This observation is further substantiated by SPEAR-TTS [14], which leverages discrete audio tokens as pseudo-textual units in low-resource scenarios.

However, the existing approaches directly utilize audio tokens, which poses challenges in sequence modeling. This is primarily due to the typically lengthy token sequences and the reliance on the model to capture correlations between tokens. While some solutions have been proposed to mitigate this issue, such as removing sequential repetitions of units used in [8, 13], these approaches suffer from corruptive encoding, making them unsuitable for speech generation tasks. In text language modeling, this problem is relieved by combining a segment of consecutive tokens according to a certain rule. For example, Byte-pair encoding (BPE) [15] dynamically creates subword units based on frequency, encoding morphological information.

Taking inspiration from text-based approaches, we propose acoustic BPE for speech generation tasks, which extends BPE to discrete audio token sequences to reduce sequence length and leverage the morphological information present in token sequence. Previous studies have investigated the effectiveness of acoustic BPE in SSL model pretraining and automatic speech recognition (ASR). For instance, HuBERT-AP[16] employs BPE to encode the pseudo target label used in HuBERT pretraining to bridge the gap between audio signal and natural language. However, it does not extend BPE to the inference procedure. [17] explores the improvements brought by acoustic BPE in the ASR task. [18] formulates text-to-speech (TTS) as a machine translation task, where the discrete VQVAE[19] sequence is encoded by BPE. However, less attention has been paid to exploring the detailed benefits brought by acoustic BPE in speech generation tasks.

In this study, we employ our proposed acoustic BPE to train a language model, which we refer to as the speech language model (SLM). The SLM serves as a generative model with various applications, including speech continuation and speech evaluation. Through comprehensive investigations, we uncover several notable benefits resulting from the use of acoustic BPE, including faster inference, better syntax capturing abilities, and improved diversity and richness in generation. These benefits can enhance various speech generation tasks, including text-to-speech, voice cloning, and speech enhancement. As an example, we introduce a novel rescore method designed

---

[*]Corresponding author.

for speech generation utilizing acoustic BPE. This method leverages the speech language model to evaluate the quality of different candidates generated by rich-diversity TTS systems. By selecting the optimal synthetic speech, our rescore method achieves a balance between diversity and naturalness. Experimental results demonstrate that the rescore selection closely aligns with human preference, further highlighting the effectiveness of acoustic BPE for other speech generation tasks.

In Section 2, we introduce the acoustic BPE and the speech language model used in experiments. We detailed our investigations on the SLM in Section 3 and present the novel rescore method in Section 3.5 as an application of acoustic BPE in speech generation tasks.

2. SPEECH LANGUAGE MODEL WITH ACOUSTIC BPE

In this section, we first give a comprehensive explanation of acoustic BPE (aBPE). Then, we introduce the unconditional speech language model (SLM) used in our experiments. Finally, we present the novel rescore method that effectively selects the optimal synthetic speech from various candidates generated by rich-diversity TTS.

2.1. Acoustic BPE

On text corpus, the byte-pair encoding (BPE) works by initiating a vocabulary that contains all unique characters in the training text. Then it iteratively merges the most frequent character pair into one unit which is added to the vocabulary. It finishes till a desired vocabulary size is achieved. To adopt the BPE algorithm on discrete audio tokens, we first convert the token sequence into Unicode text, then apply BPE training and encoding on the Unicode text. The detailed process is depicted in Figure 1 and involves the following steps:

1. We discretize audio into tokens utilizing pretrained HuBERT models and k-means clustering.

2. Then we convert the token sequence to Unicode text by mapping integer to common Chinese characters located within the Unicode region 4E00 $\sim$ 9FFF. This region contains 20992 Chinese characters, which is sufficient for our purposes.

3. Training BPE model on the obtained Unicode text with desired vocabulary size.

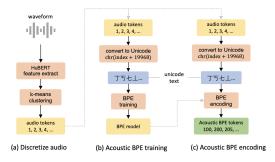4. Encode Unicode text to acoustic BPE tokens with the trained BPE model.

Fig. 1. Acoustic BPE training and encoding. (a) discretize audio into tokens leveraging the pretrained HuBERT model and k-means clustering. (b) convert discrete tokens into Unicode text for BPE training. (c) encode with the trained BPE model to obtain acoustic BPE tokens.

2.2. Speech language model

The language model (parameterized by θ) is designed to estimate the probability of a given sequence $\boldsymbol{x} = \{x_1, ..., x_t\}$. It can be expressed as:

$$p(\boldsymbol{x}|\theta) = \prod_{i=1}^{t} p(x_i|x_{<i}, \theta) \tag{1}$$

Given a sequence, the language model can calculate the probability of this sequence or predict the probability distribution of the next token to perform tasks such as continuation. The speech language model (SLM) extends language modeling to discrete audio tokens. However, unlike previous approaches [8, 13, 20] that directly utilize discrete audio tokens, our speech language model leverages acoustic BPE. This approach inherits the advantages of BPE encoding, such as enhanced sequence modeling and improved syntax capturing abilities. We use a decoder-only Transformer similar to [?] and briefly explain the process of training and generating continuations.

Training: we first discretize audio into tokens utilizing HuBERT and k-means clustering, then, we encode these tokens into acoustic BPE tokens as in Section 2.1. The SLM (parameterized by θ) is optimized to maximize the joint probability of aBPE sequence $\boldsymbol{x}$:

$$\theta = \arg\max_{\theta} p(\boldsymbol{x}|\theta) = \arg\max_{\theta} \prod_{i=1}^{t} p(x_i|x_{<i}, \theta) \tag{2}$$

Generating continuation: The speech language model can generate continuations after a given prompt $\boldsymbol{p} = \{p_1, .., p_l\}$ by autoregressively sampling the next token x_i:

$$x_i \sim p(x_i|x_{<i}, \boldsymbol{p}, \theta) \tag{3}$$

2.3. Acoustic rescoring with SLM

Recent TTS systems can generate diverse synthetic speeches. However, the naturalness of these synthetic speeches can vary considerably. To strike a balance between diversity and naturalness, one approach is to generate multiple candidates from the TTS system and manually select the most natural one. Nonetheless, this manual evaluation process is laborious and time-consuming. To address this issue, we propose a novel rescore method leveraging the SLM to select the optimal synthetic speech that ensures both diversity and naturalness.

With multiple synthetic speeches $\boldsymbol{u}^1, ..., \boldsymbol{u}^n$, we perform rescoring with SLM by first calculating the probability y^i of each speech $\boldsymbol{u}^i$ and then select the synthetic speech with the highest probability as the optimal one as follows:

$$\boldsymbol{x}^i = \{x_1^i, ..., x_t^i\} = \text{Tokenizer}(\boldsymbol{u}^i) \tag{4}$$

$$y^i = p(\boldsymbol{x}^i|\theta) = \prod_{k=1}^{t} p(x_k^i|x_{<k}^i, \theta) \tag{5}$$

$$\boldsymbol{u}^{best} = \boldsymbol{u}^{\arg\max_i y^i} \tag{6}$$

where the Tokenizer first discretizes audio into tokens by HuBERT and k-means, and then encodes tokens by acoustic BPE.

3. EXPERIMENT

3.1. Experimental setup

Dataset: We conduct experiments using two datasets: LibriTTS [21], which consists of 580 hours of speech data, and the 6k-hour clean

subset of LibriLight [22]. For LibriLight, we utilize the official scripts to segment the original long recordings into utterances of 60 seconds. All the speech waveforms are downsampled to 16kHz before training and inference.

Acoustic BPE: We extract speech features from the final layer of pretrained HuBERT Large model[1]. Then we train k-means clustering with 2000 centroids on a randomly selected 100-hour subset of speech features from the LibriTTS train-960 subset. All the speech data from LibriTTS and LibriLight is encoded to discrete tokens with the trained k-means model as illustrated in Figure 1 (a).

The acoustic BPE model is trained on the LibriTTS train-960 subset. We first convert discrete tokens to Unicode text and train the acoustic BPE model using SentencePiece[2] with the desired vocabulary size as outlined in Section 2.1. Then we use the trained BPE model to encode the discrete tokens of LibriLight into acoustic BPE tokens. In the following experiments, we compare four acoustic BPE variants: without acoustic BPE and acoustic BPE of vocabulary size 5k, 10k, and 20k.

Speech language model: We use a decoder-only Transformer with 12 layers, 16 attention heads, an embedding dimension of 1024, and a T5-style [23] relative positional embedding mechanism. During training, we use random cropping to equivalent input lengths of 15 seconds. Models with four acoustic BPE variants are trained on LibriLight 6k subset for 10 epochs with learning rate linearly increases from 0 to 1×10^{-5} for the first epoch and cosine decay to 1×10^{-6} for subsequent epochs.

Decoding discrete tokens to waveform: We train a CTX-vec2wav vocoder proposed in [24] to decode waveform from discrete tokens. The CTX-vec2wav is configured to have a frameshift of 20ms and kernel sizes (16,10,8,4) in its HifiGAN upsampling layers. It is trained on LibriTTS train-960 subset with discrete speech tokens discretized by HuBERT and k-means clustering. When decoding, acoustic BPE tokens are first decoded and then synthesized to waveform. In all experiments, we use the same speaker (speaker ID 121) as a prompt during vocoding.

In the subsequent subsections, we compare of the inference speed of speech continuation using various acoustic BPE variants in subsection 3.2. Following that, we examine the syntax capturing abilities of the SLM in subsection 3.3 and generation diversity and richness in subsection 3.3 and 3.4, respectively. Lastly, in subsection 3.5 we demonstrate the effectiveness of our novel rescore method, emphasizing the advantages conferred by the utilization of acoustic BPE.

3.2. Inference speedup

Acoustic BPE combines frequent tokens into one single unit, significantly reducing the sequence length. In Table 1, we present a comparison of the average sequence length for LibriLight-6k across four acoustic BPE variants: without acoustic BPE and acoustic BPE with vocabulary size 5k, 10k, and 20k. Also, We assess their inference speed of speech continuation with the SLM. We use the first 3 seconds of a randomly selected utterance from the LibriTTS test-clean subset as a prompt and generate 10 continuations each of 20 seconds long on an NVIDIA V100 GPU with 32GB memory. We compare the speedup brought by acoustic BPE in Table 1.

From Table 1, the employment of acoustic BPE compresses the sequence by 1.6 to 2.4 times, which is believed to ease the sequence

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/hubert
[2] https://github.com/google/sentencepiece

Encoding	Num. aBPE	Avg. sequence length	Inference speedup
w/o aBPE	-	2513.8	1.0x
aBPE	5k	1547.0	2.8x
	10k	1241.0	3.8x
	20k	**1053.0**	**5.0x**

Table 1. Comparison of sequence lengths and inference speedup across four acoustic BPE variants.

modeling. A shorter sequence also accelerates the autoregressive inference procedure by 2.8 to 5.0 times.

3.3. Syntax capturing with acoustic BPE

Following GLSM [8] and AudioLM [13], we assess the syntax modeling capability of SLM by distinguishing between a pair of syntactically correct and non-correct utterances. Similar to the aforementioned rescore method, we consider the utterance with higher probability as syntactically correct. To construct such test pairs, we filter out too short utterances from the LibriTTS test-all subset and use the remaining 5497 utterances as syntactically correct ones. To create non-meaningful utterances, we randomly shuffle the words in the syntactically correct text. Both syntactically correct and non-correct utterances are synthesized to waveforms using VQTTS [10] trained on LibriTTS train-960. Subsequently, the SLM is used to classify the two utterances in each test case into syntactically correct and non-correct. This classification accuracy is referred to as syntax accuracy. The syntax accuracy among four acoustic BPE variants is presented in Table 2.

Moreover, we examine the diversity of generated speech content similar to [8]. We randomly select 3 utterances from the LibriTTS test-clean subset and use the first 3 seconds of each utterance as a prompt to generate a 20-second continuation using the SLM. For each prompt, we repeat the continuation process 50 times, resulting in a total of 150 utterances. We use the open-sourced whisper[3] to transcribe speech into text. To evaluate the diversity of these texts, we employ the n-gram VERT metric proposed in [8]. This metric is a geometric mean of the n-gram self-BLEU and n-gram auto-BLEU scores, measuring the n-gram diversity both across sentences and within sentences. Higher VERT values indicate lower diversity. The results of 3-gram VERT are presented in Table 2.

Encoding	Num. aBPE	Syntax acc.($\uparrow$)	3-gram VERT($\downarrow$)
w/o aBPE	-	75.35%	6.99
aBPE	5k	83.45%	7.55
	10k	**85.37%**	5.97
	20k	85.34%	**4.71**

Table 2. Comparison of syntax probing accuracy and generation diversity among four acoustic BPE(aBPE) variants.

As expected, the incorporation of acoustic BPE enhances the speech language model's ability to accurately capture syntax structure and effectively model a wider range of diverse syntax patterns.

3.4. Generation richness with acoustic BPE

With acoustic BPE, the SLM can generate diverse outputs that are different and vary in content. In this section, we explore other en-

[3] https://github.com/openai/whisper/blob/main

hancements brought by acoustic BPE, specifically in generating outputs that can convey more information within a limited time.

To quantify the informativeness of the SLM, we employ cross-entropy H of the text content from the SLM generated speech with respect to a well trained text language model. The text LM can be viewed as an approximation of the true distribution of all meaningful text-content. Hence, regarding each SLM with different aBPE as an information source and assuming speech recognition does not introduce much errors, we can view cross-entropy as the measurement of the information contained in the synthetic speech. The cross-entropy is calculated as follows: Firstly, we generate a set of prompted continuations $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_n\}$ from the SLM (parameterized by θ), which are transcribed into text with ASR. Next, for each continuation, we calculate the log-probability of its transcript with a pretrained text language model (parameterized by γ). The cross-entropy $H(\text{SLM}_\theta|\text{TextLM}_\gamma)$ is obtained by averaging the negative log-probability over all generated continuations.

$$\{\boldsymbol{u}_1, ..., \boldsymbol{u}_n\} \sim p(\boldsymbol{u}|\theta) \tag{7}$$

$$H(\text{SLM}_\theta|\text{TextLM}_\gamma) = -\frac{1}{n}\sum_{i=1}^{n}\log p(\text{ASR}(\boldsymbol{u}_i)|\gamma) \tag{8}$$

In our experiments, we crop the first 3 seconds from a randomly select utterance in the LibriTTS test-clean subset as prompt and generate 150 continuations each lasting 20 seconds. we use Whisper to transcribe speech to text and use a pretrained text LM[4] to calculate cross-entropy. We present the results in Table 3.

Encoding	Num. aBPE	cross-entropy
w/o aBPE	-	352.4
aBPE	5k	409.1
	10k	457.6
	20k	**469.6**

Table 3. Comparison of cross-entropy among four acoustic BPE(aBPE) variants.

The usage of acoustic BPE increases the amount of information conveyed by the SLM within a limited time, indicating that acoustic BPE contributes to the richness of generation.

3.5. Text-to-speech rescoring with SLM

The benefits discussed above regarding acoustic BPE highlight its potential application in speech generation tasks. Here, we introduce a novel rescore method that selects the optimal synthetic speech from many candidates generated by diverse TTS systems. This method aims to balance between maintaining diversity and ensuring naturalness in the generated speech.

We use a rich-diversity TTS system[5] trained on LibriTTS train-960 subset. We randomly select 106 utterances from the LibriTTS test-clean subset and synthesize each utterance 5 times. Before proceeding to rescore with SLM, we conduct a subjective listening test where 10 listeners are asked to rank the 5 synthetic speeches based on naturalness. Subsequently, We use SLM to rescore the 5 synthetic speeches as described in Section 2.3. To quantitatively evaluate how close rescore selection aligns with human preference, we calculate

---

the top-x accuracy which denotes the success rate of rescore selection appearing within the top-x of human rankings. We compare the top-1 to top-3 accuracy with four acoustic BPE variants in Table 4.

Encoding	Num. aBPE	Top1 acc.	Top2 acc.	Top3 acc.
random	-	20.0%	40.0%	60.0%
w/o aBPE	-	29.3%	52.9%	74.3%
aBPE	5k	26.4%	50.7%	73.6%
	10k	**31.4%**	**57.9%**	**77.9%**
	20k	29.3%	55.0%	76.4%

Table 4. Rescore accuracy among various acoustic BPE(aBPE).

As shown above, the rescore selection aligns with human preference. The use of acoustic BPE further improves rescore performance. Moreover, we conducted two preference tests to verify the effectiveness of rescore and acoustic BPE. The first preference test involves a random selection and the rescore selection from SLM with acoustic BPE 10k. The second test comprises selections from two acoustic BPE variants: without acoustic BPE and acoustic BPE 10k. Results are shown in Figure 2.
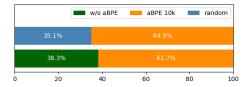

Fig. 2. Preference test. Above: random vs. aBPE 10k; Below: w/o aBPE vs. aBPE 10k.

The preference test further confirms the effectiveness of the rescore method and highlights the potential of acoustic BPE for speech generation tasks.

4. CONCLUSION

In this work, we introduce the acoustic BPE to speech generation tasks. We trained a generative speech language model using acoustic BPE and conducted thorough investigations on its property. These investigations uncover significant advantages associated with acoustic BPE. Firstly, the use of acoustic BPE results in shorter sequences, facilitating sequence modeling and accelerating autoregressive inference. Secondly, by leveraging the morphological information present in the token sequence, it alleviates the burden on the speech language model to construct token correlations, thereby enhancing its syntax capturing abilities and generation diversity and richness. Moreover, we demonstrate the application of acoustic BPE on a novel TTS rescore method, which selects the optimal one from multiple diverse TTS syntheses to strike a balance between diversity and naturalness. Experimental results provide empirical evidence of the effectiveness of acoustic BPE. These findings open up possibilities for applying acoustic BPE to other speech generation tasks such as text-to-speech, which can be explored in future research.

5. ACKNOWLEDGEMENTS

---

[4]https://github.com/pytorch/fairseq/tree/master/examples/language_model

[5]https://github.com/lifeiteng/vall-e

6. REFERENCES

[1] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019*, 2019, pp. 3465–3469.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[3] Stuart P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[4] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu, "w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *ASRU 2021*. 2021, pp. 244–250, IEEE.

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[6] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR 2020*. 2020, OpenReview.net.

[7] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS 2020*, 2020.

[8] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al., "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[9] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *ICASSP 2018*. 2018, pp. 4779–4783, IEEE.

[10] Chenpeng Du, Yiwei Guo, Xie Chen, and Kai Yu, "VQTTS: high-fidelity text-to-speech synthesis with self-supervised VQ acoustic feature," in *Interspeech 2022*. 2022, pp. 1596–1600, ISCA.

[11] Chenpeng Du, Yiwei Guo, Feiyu Shen, and Kai Yu, "Multi-speaker multi-lingual VQTTS system for LIMMITS 2023 challenge," *CoRR*, 2023.

[12] Chenpeng Du, Yiwei Guo, Xie Chen, and Kai Yu, "Speaker adaptive text-to-speech with timbre-normalized vector-quantized feature," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[13] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour, "Audiolm: A language modeling approach to audio generation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2523–2533, 2023.

[14] Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matthew Sharifi, Marco Tagliasacchi, and Neil Zeghidour, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *CoRR*, vol. abs/2302.03540, 2023.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*, 2019, pp. 4171–4186.

[16] Shuo Ren, Shujie Liu, Yu Wu, Long Zhou, and Furu Wei, "Speech pre-training with acoustic piece," in *Interspeech 2022*, 2022, pp. 2648–2652.

[17] Xuankai Chang, Brian Yan, Yuya Fujita, Takashi Maekaku, and Shinji Watanabe, "Exploration of efficient end-to-end ASR using discretized input from self-supervised learning," *CoRR*, vol. abs/2305.18108, 2023.

[18] Tomoki Hayashi and Shinji Watanabe, "Discretalk: Text-to-speech as a machine translation problem," *CoRR*, vol. abs/2005.05525, 2020.

[19] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura, "VQVAE unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019," in *Interspeech 2019*. 2019, pp. 1118–1122, ISCA.

[20] Soumi Maiti, Yifan Peng, Takaaki Saeki, and Shinji Watanabe, "Speechlmscore: Evaluating speech generation using speech language model," in *ICASSP 2023*, 2023, pp. 1–5.

[21] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech 2019*. 2019, pp. 1526–1530, ISCA.

[22] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux, "Libri-light: A benchmark for ASR with limited or no supervision," in *ICASSP 2020*, 2020, pp. 7669–7673.

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[24] Chenpeng Du, Yiwei Guo, Feiyu Shen, Zhijun Liu, Zheng Liang, Xie Chen, Shuai Wang, Hui Zhang, and Kai Yu, "Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding," *CoRR*, vol. abs/2306.07547, 2023.