MixedG2P-T5: G2P-free Speech Synthesis for Mixed-script texts using Speech Self-Supervised Learning and Language Model

Joonyong Park*, Daisuke Saito*, Nobuaki Minematsu*
* The University of Tokyo, Japan
E-mail: {jpark, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

***Abstract*—This study presents a novel approach to voice synthesis that can substitute the traditional grapheme-to-phoneme (G2P) conversion by using a deep learning-based model that generates discrete tokens directly from speech. Utilizing a pre-trained voice SSL model, we train a T5 encoder to produce pseudo-language labels from mixed-script texts (e.g., containing Kanji and Kana). This method eliminates the need for manual phonetic transcription, reducing costs and enhancing scalability, especially for large non-transcribed audio datasets. Our model matches the performance of conventional G2P-based text-to-speech systems and is capable of synthesizing speech that retains natural linguistic and paralinguistic features, such as accents and intonations.**

I. Introduction

Speech synthesis refers to the technology by which machines automatically generate speech audio signals and is commonly known as text-to-speech (TTS). With the advancement of deep learning, speech synthesis models have demonstrated performance that significantly surpasses traditional methods [1]. These models typically convert input text into acoustic feature vectors through an encoder, and subsequently generate Mel-spectrograms using techniques such as attention mechanisms or variational inference, which are then transformed into speech by a vocoder [2], [3]. The model learns the correspondence between audio samples and their respective "input representations."

Constructing such deep learning-based speech synthesis systems requires accurately labeled data corresponding to spoken utterances. In conventional approaches, phonemes are typically generated from sample text using grapheme-to-phoneme (G2P) conversion, which are then input into the speech synthesis model. In the case of Japanese, where texts often contain a mix of kanji and kana, phonemes are generated from the mixed-script input, which are subsequently used to synthesize speech. Specifically, some methods rely on rule-based systems to assign required TTS information—such as accent and prosody—based on morphological analysis, while others adopt neural G2P models using CTC or encoder-decoder structures to model the alignment between text and phoneme sequences of differing lengths.

Such transcription tasks are largely conducted manually. While it is possible to incorporate additional information—such as accents and syllable durations—by referring to pronunciation or accent dictionaries, two major challenges remain in building G2P systems. The first is the cost associated with data construction. Generating phonetic elements requires various resources, including pronunciation and accent dictionaries and linguistic rules. Since these supplementary inputs cannot be derived solely from raw text, they must be individually integrated into the system, thereby incurring high annotation costs.

The second challenge lies in the limited support for multilingual text. When dealing with texts that include multiple languages, it becomes difficult for a single G2P model to provide adequate coverage. This necessitates the development of separate models for each language, which further increases costs. Additionally, pronunciation errors can arise when the same character is pronounced differently depending on the language, presenting a significant difficulty in multilingual speech synthesis.

To address these challenges, this study aims to develop a G2P-free, multilingual-capable speech synthesis model by utilizing discrete representations derived from speech self-supervised learning (SSL) models. To achieve this goal, the following key aspects are investigated.

- Performance comparison between input representation using SSL model and conventional G2P representation
- Implementation of G2P for mixed Kanji and Kana situations using discrete representation

II. Related Works

To address the aforementioned challenges in speech synthesis—particularly the reliance on costly and language-dependent grapheme-to-phoneme (G2P) conversion—recent studies have explored the use of discrete tokens, which are directly extracted from raw audio using speech self-supervised learning (SSL) models trained on large-scale unlabeled speech corpora. Unlike conventional phoneme representations, these discrete tokens encode not only linguistic and semantic content, but also speaker-specific paralinguistic features such as accentuation, intonation, and prosody. This rich and compact representation is especially advantageous because it avoids the information loss typically incurred during the intermediate conversion of speech into text.

979-8-3315-7206-8/25/$31.00

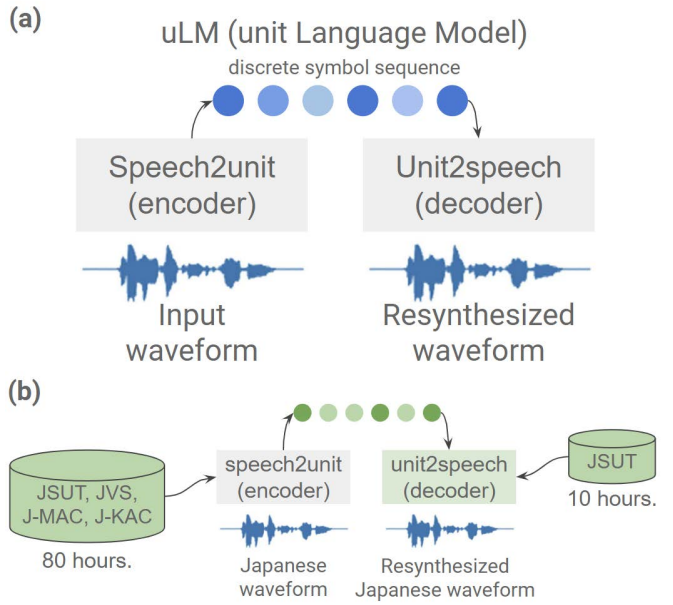

Fig. 1. (a) Architecture of GSLM and (b) Application to the Japanese language

A prominent framework that leverages this representation for speech synthesis is the Generative Spoken Language Model (GSLM) [4]. GSLM establishes a three-stage pipeline: (1) an encoder that converts the input speech waveform into a sequence of discrete symbols via quantized SSL embeddings, (2) an optional unsupervised language model (uLM) that captures long-range dependencies across the symbol sequence, and (3) a decoder that reconstructs the speech waveform from the symbolic input. Encoders such as wav2vec 2.0 [5] and HuBERT [6] are commonly employed, offering robust and generalizable speech representations. These discrete tokens act as a learned alternative to phonemes or graphemes, forming a flexible intermediate layer between raw audio and synthesis.

As depicted in Figure 1(a), the encoder module transforms the continuous speech signal into a compressed symbolic form using a k-means clustering layer trained on SSL features. The decoder module, typically a neural vocoder or sequence-to-sequence model such as Tacotron 2 [2], learns to reconstruct the speech waveform from these symbolic representations. Interestingly, it has been demonstrated (Figure 1(b)) that this symbolic representation is not strictly language-specific: once the encoder-decoder pipeline is trained on one language, it can be transferred to other languages through fine-tuning, enabling zero-shot or low-resource language synthesis. This cross-lingual transferability opens the door to G2P-free synthesis across many languages, even those lacking well-defined phonemic resources.

III. Research Approach

In this work, we extend this idea by using the encoder of GSLM not only for decoding speech but also for constructing a pseudo-language label predictor, which can infer symbolic representations directly from raw text. The goal is to replace

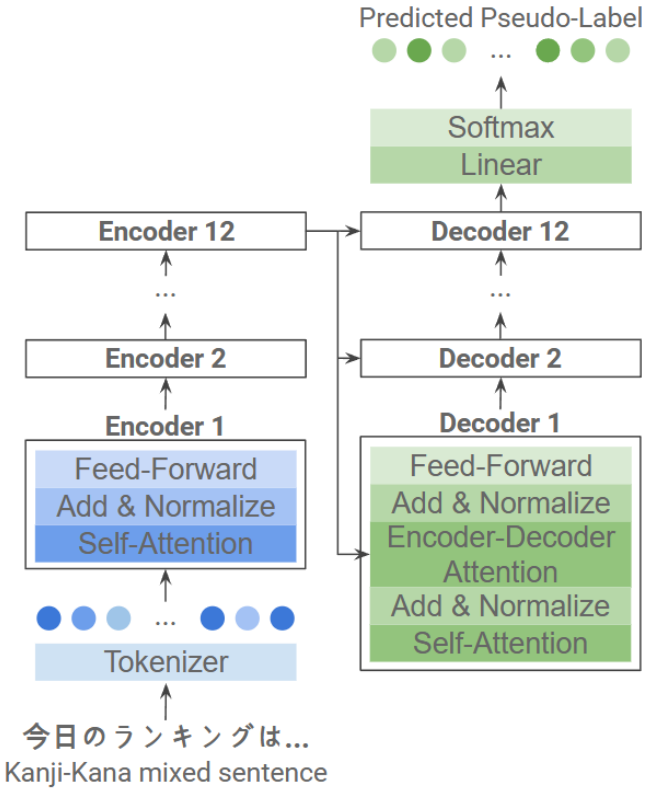

Fig. 2. Architecture of the Pseudo-Language Label Prediction Model

the G2P module with a learned mapping from text, which may contain mixed scripts, such as Kanji and Kana in Japanese, to symbolic tokens that function analogously to phonemes. Once these pseudo-language labels are predicted, they can be fed into the spectral predictor to produce natural-sounding synthesized speech.

The pseudo-language label predictor is built on top of the T5 architecture [9], a powerful Transformer-based sequence-to-sequence model known for its success in a wide range of natural language processing tasks. T5's encoder-decoder structure allows it to handle inputs and outputs of varying lengths, making it suitable for converting raw text into sequences of symbolic labels. Moreover, T5 is pretrained on a massive multilingual corpus, which equips it with broad linguistic understanding and generalization capabilities. In our setting, it enables the conversion of Japanese text—including diverse orthographic patterns—to meaningful discrete symbol sequences. The architecture of this label predictor is illustrated in Figure 2. By leveraging this T5-based predictor, we eliminate the need for hand-crafted phoneme dictionaries, morphological analyzers, or accent dictionaries typically required in G2P pipelines.

While prior studies have explored the use of T5 for grapheme-to-phoneme conversion tasks, they remain constrained to alphabetic scripts and conventional G2P paradigms [15], [16]. Specifically, their work assumes a clear one-to-one correspondence between characters and phonemes,

TABLE I
DATASETS USED FOR TRAINING, FINE-TUNING, AND APPLYING EACH MODEL

Language	G2P	Language-Specific Pseudo-Language Label Predictor	Spectral Predictor
Japanese	OpenJTalk[1] + Mecab[2] + Marine [7]	Reazonspeech [8] (T5 [9] Pre-training) JSUT [10], JVS [11], JKAC [12], JMAC [13], JSSS [14] (Fine-tuning for k-means/T5) tohoku-BERTv3-tokenizer[3] (T5 Tokenizer)	JSUT [10]

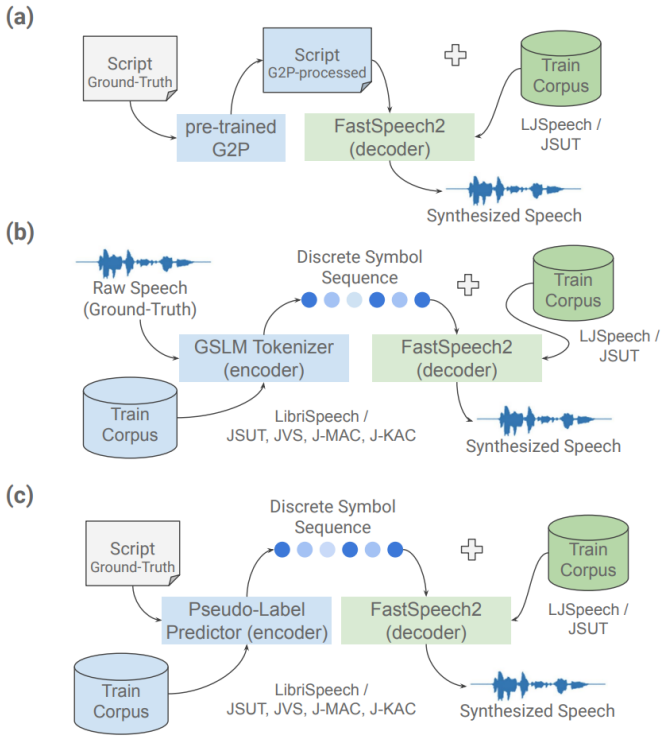

Fig. 3. Construction of speech synthesis systems using (a) phonemes obtained via G2P (Baseline), (b) pseudo-language labels obtained from original speech (Oracle), and (c) pseudo-language labels predicted from raw text (Proposed)

and does not address the combined demonstration of mixed-script languages. Also, multilingual G2P systems have also been proposed, but they still operate within a text-based inference thus does not have additional prosodic features essential for natural-sounding speech synthesis, such as accents and durations, that can be obtained from SSL models [17]–[19].

Therefore, this study demonstrate a fully G2P-free speech synthesis pipeline that leverages SSL-derived tokens and a T5-based pseudo-language label predictor to handle the structural and phonological intricacies of mixed-script languages, such as Japanese kanji and kana.

IV. EXPERIMENTAL SETUP

In order to verify performance of the pipeline, we designed three pipelines based on different encoding architectures to process Japanese sentences. The overall structure is illustrated in Figure 3.

[1]https://open-jtalk.sp.nitech.ac.jp/
[2]http://taku910.github.io/mecab/
[3]https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

The first structure serves as the conventional baseline, using G2P-derived representations obtained from reference scripts. This G2P system, trained using traditional methods such as HMMs, provides phoneme sequences along with information on duration and accent, which are used as inputs for speech synthesis.

The second structure serves as the oracle and utilizes pseudo-language labels derived directly from raw speech using the encoder of the GSLM system trained in a self-supervised manner. Since GSLM does not support synthesis from text scripts, this model generates labels from the original speech aligned with the reference scripts. These labels are then used to evaluate the performance of the pseudo-language label predictor, which is trained on the corresponding discrete representations.

The third structure is our proposed method, which employs a pseudo-language label predictor to infer labels from raw text. Discrete labels are first extracted from raw speech using the GSLM encoder, and a dataset of text–label pairs is constructed using the corresponding reference scripts. This dataset is then used to train the predictor model, enabling it to output predicted pseudo-language labels given new text input. The pseudo-language label predictor is trained on Japanese utterances using a T5 model with 12 encoder and decoder layers, pre-trained and fine-tuned on TTS data (see Table 1).

oracle and the proposed method both employ the GSLM architecture enhanced with speaker embedding capabilities using ContentVec [20], an SSL model known for effectively capturing speaker characteristics. The SSL and k-means models were fine-tuned on a Japanese speech corpus, resulting in a model capable of outputting 500 discrete tokens. In both cases, a preprocessing step was applied to remove repeated symbols from the pseudo-language label sequences.

The spectral predictor used to generate synthesized speech from the common input representations is FastSpeech 2 [21]. FastSpeech 2 has the advantage of being able to additionally control speech characteristics through duration, pitch, and accent information, in addition to textual inputs. The model is trained to predict Mel-spectrograms from input text, which are then converted to speech using a vocoder under the same conditions.

In the conventional approach, the model is trained on paired data consisting of speech and G2P-derived text. Here, phoneme sequences, duration, and accent information are learned using pre-constructed TextGrid files based on G2P processing.

In the oracle and proposed methods, training is conducted using pairs of speech and pseudo-language labels. Since the

TABLE II
EVALUATION METRICS OBTAINED FROM THE DISCRETE SYMBOLS GENERATED BY THE LANGUAGE LABEL PREDICTOR AND SYNTHESIZED SPEECH GENERATED BY THE SPECTRAL PREDICTOR

		GT	Baseline	Oracle	Proposed
Discrete Symbols	UER(%)↓			-	7.47
Synthesized Speech (Spectral Predictor)	CER(%)↓	16.21	18.24	20.63	21.28
	UTMOS↑	2.79	2.59	2.49	2.54
	WARP-Q↑	-	2.47	2.64	2.63
	SDR(dB)↑	-	-22.79	-23.12	-23.28

output pseudo-language labels may include repeated symbols, these repetitions can be used as cues to predict duration.

V. EXPERIMENTAL RESULTS

We first evaluate the pseudo-language label predictor by assessing the linguistic information captured in the predicted discrete symbols. Then, we evaluate the synthesized speech generated via the spectral predictor. Table 2 summarizes the evaluation metrics across the three model architectures, using 100 utterances selected from the JVS corpus that were not used during training.

A. Evaluation of Linguistic Intelligibility

To assess the impact of each input representation on linguistic intelligibility, we analyzed error rates with respect to the target language. For the discrete symbols, we compared the pseudo-language labels predicted by the label predictor with those obtained from the original audio via the SSL model, and calculated the Unit Error Rate (UER). For the synthesized speech, we used the Whisper-base [22] ASR model and compared its output with the reference transcription, calculating the Character Error Rate (CER) in Japanese.

Table 2 shows the UER and CER scores. While UER did not show large differences compared to prior experiments, prediction errors were still observed in the label predictor. For the synthesized speech, although the CER increased compared to the G2P-based baseline, the increase was smaller when compared to the oracle model. This suggests that synthesis errors are more heavily influenced by the spectral predictor than by the label predictor.

B. Evaluation of Naturalness

We employed the pretrained UTMOS model [23] for this evaluation. UTMOS is a model trained to predict Mean Opinion Scores (MOS) for multilingual synthetic speech in an automated fashion. Since it operates without comparison to reference audio, it can be used to assess absolute linguistic and paralinguistic qualities, though it cannot capture comparative variations between ground truth and synthesized audio.

Table 2 also includes UTMOS scores. While the baseline exhibited the highest naturalness, the proposed method achieved a higher UTMOS score than the oracle model. This further supports the finding that differences in naturalness between the baseline and proposed method are greater than those between oracle and proposed methods, indicating that the spectral predictor has a more substantial influence on naturalness than the label predictor.

C. Evaluation of Acoustic Quality

Next, we evaluated the overall acoustic quality of the synthesized speech.

Given that input representations are compressed speech features, they can be viewed as analogous to neural audio codecs. Thus, we treated the speech synthesis system as a single transmission pipeline and evaluated the output speech as a degraded version of a reference signal.

For this purpose, we employed WARP-Q [24], which is robust to common codec-related distortions and compensates internally for temporal misalignments between the ground truth and resynthesized signals—issues often encountered with metrics like PESQ [25], which can be sensitive to speaker identity and speech type.

We also assessed audio quality in terms of noise contamination using the Signal Distortion Rate (SDR), a metric typically used in source separation. SDR is suitable here as it quantifies the degree of noise in the synthesized audio relative to the original input.

As shown in Table 2, WARP-Q scores for the proposed model were higher than those of the baseline and comparable to the oracle model. Similarly, SDR results were consistent across the baseline, oracle, and proposed models. These results suggest that the overall pipeline involving the label predictor does not introduce significant acoustic degradation.

VI. CONCLUSION

This study presented a novel approach to speech synthesis that replaces conventional grapheme-to-phoneme (G2P) conversion by employing a model that directly generates discrete tokens from speech. Furthermore, we trained a pseudo-language label predictor and a spectral predictor using a speech SSL model on mixed Kanji-Kana Japanese texts and analyzed the outputs under various conditions. Experimental results demonstrated that it is possible to predict pseudo-language labels from text with comparable performance to traditional G2P-based models.

Several areas for improvement have been identified. For the language label predictor, we utilized a Japanese tokenizer pretrained on Japanese text. However, to support multilingual settings, it is necessary to explore alternative tokenizers such as Byte-Pair Encoding (BPE) tokenizers that do not rely on

language-specific training. For the spectral predictor, additional acoustic features—such as accentual information—should be integrated alongside the four currently used features.

From an evaluation perspective, it is also necessary to individually assess the contributions of each input factor (e.g., duration, pitch) fed into FastSpeech 2. Moreover, a broader and more diverse test dataset would allow for a more comprehensive and generalized analysis.

As future directions, we plan to conduct experiments on multilingual datasets to investigate the language dependency of the language label predictor. In order to minimize such dependency, we intend to explore strategies that avoid language-specific preprocessing. One possible solution is to convert raw text into language-aware discrete tokens using tokenizers such as mT5 [26] or ByT5 [27], which tokenize text based on Unicode character strings. These models are capable of performing tokenization while preserving minimal language-specific biases, thus enabling scalable and robust evaluation across a wider range of multilingual scenarios.

References

[1] L. Borgholt, J. D. Havtorn, J. Edin, L. Maaløe, and C. Igel, "A brief overview of unsupervised neural speech representation learning," *CoRR*, vol. abs/2203.01829, 2022. arXiv: 2203.01829.

[2] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS synthesis by conditioning Wavenet on mel-spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4779–4783.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 5530–5540.

[4] K. Lakhotia, E. Kharitonov, W.-N. Hsu, *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[6] 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.

[7] B. Park, R. Yamamoto, and K. Tachibana, "A Unified Accent Estimation Method Based on Multi-Task Learning for Japanese Text-to-Speech," in *Proc. Interspeech 2022*, 2022, pp. 1931–1935.

[8] Y. Yue, M. Daijiro, and F. Seiji, "Reazonspeech: A free and massive corpus for Japanese ASR," in *Proceedings of Annual Meeting of the Association for NLP*, 2023, 1134–1139.

[9] C. Raffel, N. M. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, 140:1–140:67, 2019.

[10] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," *CoRR*, vol. abs/1711.00354, 2017. arXiv: 1711.00354.

[11] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: Free Japanese multi-speaker voice corpus," *CoRR*, vol. abs/1908.06248, 2019. arXiv: 1908.06248.

[12] W. Nakata, T. Koriyama, S. Takamichi, *et al.*, "Audiobook speech synthesis conditioned by cross-sentence context-aware word embeddings," in *Proc. The 11th ISCA SSW*, 2021.

[13] S. Takamichi, N. Wataru, T. Naoko, and S. Hiroshi, "J-MAC: Japanese multi-speaker audiobook corpus for speech synthesis," in *Interspeech 2022*, ISCA, 2022, pp. 2358–2362.

[14] S. Takamichi, M. Komachi, N. Tanji, and H. Saruwatari, "JSSS: free japanese speech corpus for summarization and simplification," *CoRR*, vol. abs/2010.01793, 2020. arXiv: 2010.01793.

[15] M. Řezáčková, J. Švec, and D. Tihelka, "T5g2p: Using text-to-text transfer transformer for grapheme-to-phoneme conversion," in *Interspeech 2021*, 2021, pp. 6–10.

[16] J. Ao, R. Wang, L. Zhou, *et al.*, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol. 1, Dublin, Ireland, May 2022, pp. 5723–5738.

[17] A. Sokolov, T. Rohlin, and A. Rastrow, "Neural machine translation for multilingual grapheme-to-phoneme conversion," in *Interspeech 2019*, 2019, pp. 2065–2069.

[18] K. Vesik, M. Abdul-Mageed, and M. Silfverberg, "One model to pronounce them all: Multilingual grapheme-to-phoneme conversion with a transformer ensemble," in *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Jul. 2020, pp. 146–152.

[19] J. Zhu, C. Zhang, and D. Jurgens, "Byt5 model for massively multilingual grapheme-to-phoneme conversion," in *Interspeech 2022*, 2022, pp. 446–450.

[20] K. Qian, Y. Zhang, H. Gao, *et al.*, "ContentVec: An improved self-supervised speech representation by disentangling speakers," in *Proceedings of the 39th International Conference on Machine Learning*, 2022, pp. 18 003–18 017.

[21] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *ArXiv*, vol. abs/2006.04558, 2020.

[22] A. Radford, K. Jong Wook, X. Tao, B. Greg, M. Christine, and S. Ilya, *Robust speech recognition via large-scale weak supervision.* https://cdn.openai.com/papers/whisper.pdf, 2022.

[23] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022," in *Interspeech 2022*, ISCA, 2022, pp. 4521–4525.

[24] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Speech quality assessment with WARP-Q: From similarity to subsequence dynamic time warp cost," *IET Signal Processing*, vol. 16, no. 9, 1050—1070, 2022.

[25] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, 749–752 vol.2.

[26] X. Linting, C. Noah, R. Adam, *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021, pp. 5206–5210.

[27] X. Linting, B. Aditya, C. Noah, *et al.*, "ByT5: Towards a token-free future with pre-trained byte-to-byte models," in *TACL 2022*, 2021, pp. 5206–5210.