Analytic Study of Text-Free Speech Synthesis for Raw Audio using a Self-Supervised Learning 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*—We examine the text-free speech representations of raw audio obtained from a self-supervised learning (SSL) model by analyzing the synthesized speech using the SSL representations instead of conventional text representations. Since raw audio does not have paired speech representations as transcribed texts do, obtaining speech representations from unpaired speech is crucial for augmenting available datasets for speech synthesis. Specifically, the proposed speech synthesis is conducted using discrete symbol representations from the SSL model in comparison with text representations, and analytical examinations of the synthesized speech have been carried out. The results empirically show that using text representations is advantageous for preserving semantic information, while using discrete symbol representations is superior for preserving acoustic content, including prosodic and intonational information.**

I. Introduction

Current speech synthesis has significantly advanced through deep learning models, greatly surpassing the performance of traditional speech synthesis models [1]. These models generally obtain speech feature vectors from the input text through an encoder, then output a Mel-spectrogram using methods such as Attention or Variational Inference, and finally convert it into speech through a Vocoder [2]–[4]. At this stage, speech synthesis models essentially learn the correspondence between the training speech source and the 'input representation,' which describes the speech source and is traditionally represented as a transcribed text script [5].

However, a constraint of such models is that enhancing performance necessitates the additional task of pairing input representations as labels to the training speech source, which involves human effort and thus incurs significant costs. Furthermore, such text-based input representations vary by language, requiring even more resources when creating multilingual speech synthesis. To overcome these constraints, the use of Self-Supervised Learning (SSL) models, which can extract usable information from the raw speech source, is being considered.

This study evaluates the performance of speech synthesis through several different input expressions, which analyzes the factors to consider when achieving such as "zero-resource synthetic speech", and evaluates the intelligibility, naturalness, and quality of the synthesized speech to see how each factor affects those speech.

II. Prior Studies

First, traditional speech synthesis models have focused on converting transcribed natural language text into a format more suitable for speech synthesis to obtain potential input representations. This includes text normalization, tokenization, the assignment of morphemes and parts of speech, phonemizing, and assigning accents to natural language texts through pre-prepared dictionaries and libraries. However, such methods have limitations in generalizing beyond single-language or high-resource language domains.

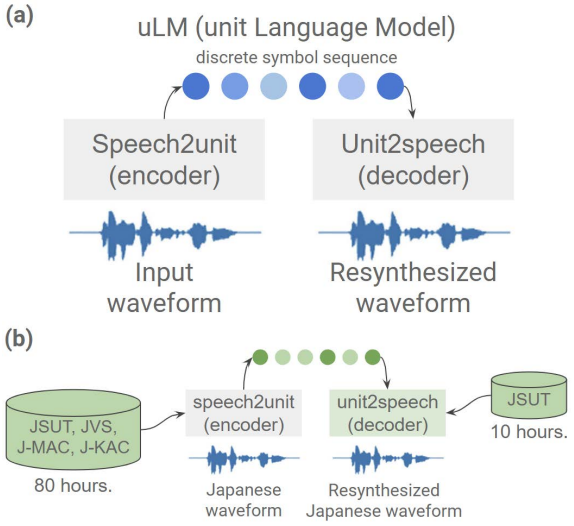

Fig. 1. (a) Architecture of GSLM and (b) Application to Japanese Language

Therefore, methods that can obtain effective intermediate representations for speech synthesis tasks without text transcription, using audio data as a modality, have recently been researched. Self-supervised learning models are trained on a large amount of unlabeled audio data to predict hidden units or symbols not directly observed in the input audio data. Consequently, obtaining corresponding input representations for unlabeled speech has become more accurate, and it has also become possible to incorporate paralinguistic and non-verbal information into these input representations.

In a previous study on this approach, the Generative Spoken Language Modeling (GSLM) [6] provided a methodology for speech synthesis from raw audio. Figure 1(a) shows a schematic illustration of the GSLM model architecture. GSLM processes and resynthesizes speech not through commonly used natural language texts but through discrete symbols, which are feature representations extracted from speech using SSL models. The model consists of an encoder called

speech2unit, a decoder called unit2speech, and a unit language model (uLM). The speech2unit module converts audio waveforms into discrete symbols in text form through feature representations via SSL models like CPC [7], wav2vec 2.0 [8], and HuBERT [9]. It then quantizes the features with a predetermined codebook to obtain the discrete symbol sequence. The codebook is obtained by applying k-means clustering to the framewise features of the training data.

Conversely, the unit2speech module generates audio waveforms from sequences of discrete symbols using traditional text-to-speech models like Tacotron 2 [5] and neural vocoder models. In this process, the speech synthesis model is trained with pairs of training speech and discrete symbols outputted by the encoder. In the case of speech resynthesis, there is a pipeline where the speech encoded through the speech2unit module is transformed back into synthesized speech through the unit2speech module. However, since unit2speech is basically trained in a single language, language dependency needs to be resolved to use it for multiple languages. Figure 1(b) shows a Japanese application of the GSLM model architecture, where language dependence was resolved to some extent by training the speech2unit's k-means clustering model and unit2speech module with a Japanese dataset. The uLM module, positioned between the two modules mentioned above, treats discrete symbols like character symbols and operates as a language model using the Transformer [10] network.

Additionally, as a prior study on the evaluation of synthesized speech without labels, the Zero Resource Speech Challenge (ZRC) can be mentioned. The research sets several tasks with the goal of constructing language processing models using only audio data by removing text labels from a ground truth speech corpus. The objective is to surpass the topline models, which are trained from transcribed text, in metrics for each task. Among the tasks, in the 'Discrete Resynthesis' task, the input representations obtained through SSL models are resynthesized, and clarity and naturalness are evaluated through Character Error Rate (CER) and Mean Opinion Score (MOS). The experimental results have shown that models adopting certain SSL methods achieved better MOS results than the topline, and it was also confirmed that the more bit-information the input representations contained, the higher the metrics of synthesized speech were. [11]–[13]

In addition, through structures such as GSLM, research on solving audio tasks as end-to-end using untranscribed audio data has continued to show superior results compared to previous baselines [14]–[16]. From such prior studies, it is suggested that by obtaining input representations through self-supervised learning methods for audio data without pre-transcribed labels, it is possible that SSL representations may show superior performance compared to text transcriptions for speech synthesis in general.

III. Comparative Study on Methods Based on Input Representations

In the case of the Zero Resource Speech Challenge, only the evaluation of intelligibility and naturalness after synthesis was present, while the quality of the acoustics was not evaluated. Additionally, prior studies have not conducted complete analysis of multiple languages or multiple encoder-decoder pairs, and this has not been analyzed with several conditions such as symbol size and output layer of Transformer.

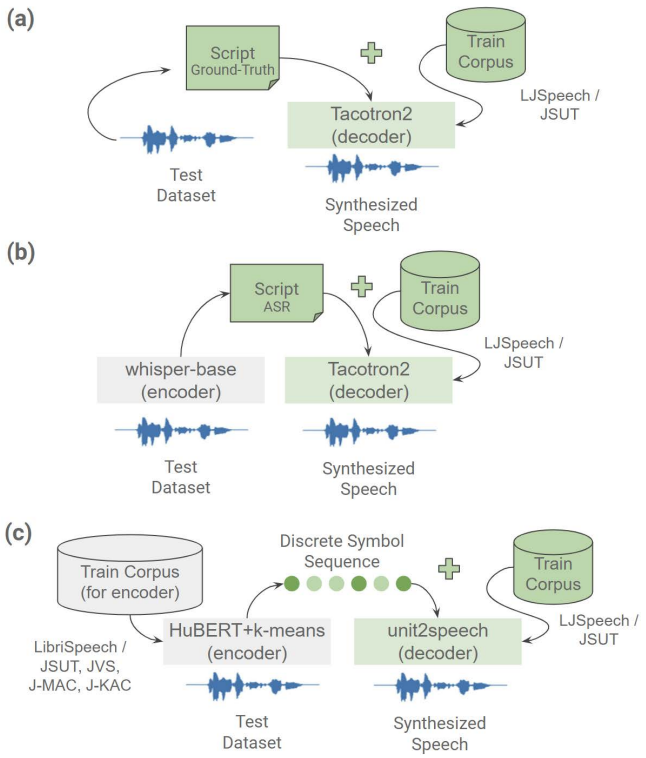

Fig. 2. Building a speech synthesis system using (a) ground-truth script labels, (b) speech recognition (ASR) model, and (c) self-supervised learning (SSL) model

Therefore, this study evaluates by adding metrics that can measure acoustic quality, in addition to the experimental results on clarity and naturalness, along with changes in the models used. Furthermore, SSL models are often trained in a specific language, which could show degraded performance in speech synthesis for multiple languages due to language dependency. Hence, this research proceeds with experiments not only in English but also in Japanese, investigates the presence or absence of language dependency in the representations outputted through SSL models, and ultimately analyzes the impact of such factors on the nature of synthesized speech.

A. Experiment Setting

In this experiment, models for three different input representations were created for two languages: English and Japanese. The structure is shown in Figure 2.

The first model (2-(a)) is used as a reference, synthesizing speech using text scripts as input representations, which serve as the correct (gold) labels.

The second model (2-(b)) serves as a baseline, synthesizing speech after recognizing the training speech dataset through an ASR model. The ASR model used is Whisper [24], with

TABLE I
CORPORA USED FOR TRAINING HYPOTHESIS MODEL

Language	S2u(multiple speakers)	u2S(single speaker)
English	LibriSpeech[17]	LJSpeech[18]
Japanese	Reazonspeech [19] (HuBERT) JSUT[20], JVS[21], JKAC[22], JMAC[23] (k-means)	JSUT

the base model used for both English and Japanese. For these above two models, preprocessing includes the use of the English cleaner provided by Tacotron2 for English, and morphological analysis through Mecab for Japanese.

The third model (2-(c)) is used as a hypothesis, synthesizing speech by obtaining input representations in discrete symbol form via the speech2unit module from the self-supervised learning model presented by GSLM. [1] For this, pipelines for English and Japanese are presented. The datasets for each language used in speech2unit and unit2speech are shown in Table 1; the sampling rate of the audio was unified to 16k for all pre-trained, training, and test datasets. We use HuBERT-base for outputting the input representation in the form of a discrete symbol through a k-means model.

In addition, several differential elements were introduced in the SSL model and the model was analyzed using the ablation method. Firstly, in order to analyze the language dependency of speech2unit, the difference between the two synthetic speech is analyzed as shown in Figure 3, assuming that speech2unit is trained in the same language as unit2speech and in a different language. This creates four combinations of S2u-u2S pair in total. Secondly, in order to evaluate the effect of changing the code length of the discrete symbol, synthetic speech that is synthesized from three different discrete symbols are compared, by training the k-means model's clustering number to 50, 200, and 1000, respectively. For all cases, processing was performed to remove repeated symbols from the discrete symbol sequences. Thirdly, to identify differences in output between Transformer layers within speech2unit, synthetic speech from the representation by speech2unit's different layers is been compared: in this experiment, 6th and 12th layer. Attempts have been conducted to find effective features within SSL models, and it has also been found that the differences between these layers has made difference in 'linguistic' and 'acoustic' factors [25], [26]. These conditions also verify that the prior research is consistent with the variables of language and code length.

In common with all three models, the Tacotron 2 model was used for generating synthesized speech through input representations. The model is designed to train Mel-Spectrograms from the text of input representations and to output speech using a Vocoder under the same conditions for each language, except for the differences in input representations.

For the test datasets, 100 utterances selected from LibriSpeech-dev were used for English, and for Japanese, 100

[1]As this experiment focuses only on speech resynthesis, which does not require the uLM, we only analyze the speech2unit and unit2speech equivalent modules.

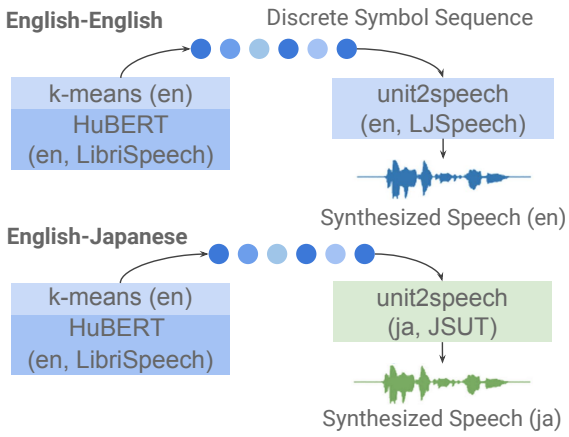

Fig. 3. An example of a SSL system that speech2unit-unit2speech pairs are language-matched and unmatched

utterances selected from utterances not used as training data in the JSSS corpus [27] were used. The test dataset is also converted into the form of same representations that are used in the training phase, such as transcribed text (GT), ASR script (ASR), and discrete symbol sequence (SSL) respectively. Those input representations are used for inference.

IV. EXPERIMENT RESULTS

The experiment is broadly divided into evaluations of speech language and acoustic quality. The speech language evaluation aims to assess changes in the linguistic semantic elements of synthesized speech based on input representations, while the acoustic quality evaluation focuses on elements other than the language contained in the synthesized speech audio.

The data are shown in Table 2 and are the result of SSL speech synthesis comparing with the reference GT and baseline ASR. The characters in the item represents 'S2u-u2S language matching', 'symbol size', and 'Transformer output layer', respectively. For example, in the case of *match-200-L12*, it refers a synthesized speech that used same S2u with u2S, 200 symbol size and discrete symbols obtained from the 12th layer.

A. Speech Intelligibility

For the language evaluation, the impact of each input representation on the intelligibility of speech language was examined by investigating the error rates for the target languages. The output from Whisper-base and the correct scripts of each test dataset were compared, investigating Word Error Rate (WER) for English and Character Error Rate (CER) for Japanese. As an initial step in evaluating the performance of the speech recognition model used in the experiment, the error rate was investigated against the correct scripts for the GT test dataset processed by speech recognition. The results showed a word error rate of 2.41 % for English and a character error rate of 19.35 % for Japanese.

Also, to perform a more detailed analysis using a language-independent metric, phonemes were obtained from each piece of speech information, and their errors were also investigated.

TABLE II

METRICS COMPARING SYNTHESIZED SPEECH FROM THE INPUT REPRESENTATION OF GROUND TRUTH, ASR MODEL, AND VARIOUS SSL MODELS FOR ENGLISH AND JAPANESE RESPECTIVELY. ITEM IS EMBOLDENED WHEN IT RECORDED THE MOST SUPERIOR VALUE AMONG SSL MODELS, AND BASELINE IS UNDERLINED IF THE BASELINE VALUE IS SUPERIOR THAN THE ITEM.

	English					Japanese				
	WER(%)↓	PER(%)↓	UTMOS↑	WARP-Q↑	SDR(dB)↑	CER(%)↓	PER(%)↓	UTMOS↑	WARP-Q↑	SDR(dB)↑
GT	4.22	8.14	3.93 ± 0.067	-	-	20.68	41.59	2.68 ± 0.045	-	-
ASR	5.41	9.92	3.41 ± 0.084	2.71	-18.55	24.62	44.91	2.57 ± 0.049	2.23	-23.53
match-50-L6	8.37	19.09	3.31 ± 0.096	2.62	-19.91	42.88	82.19	2.10 ± 0.064	2.20	-23.95
match-50-L12	7.82	16.06	3.49 ± 0.076	2.78	-19.32	35.73	69.02	2.21 ± 0.055	2.29	-23.81
match-200-L6	6.95	15.24	3.45 ± 0.068	2.86	-17.14	32.95	59.08	2.44 ± 0.041	2.37	-23.56
match-200-L12	6.63	14.53	3.60 ± 0.057	2.89	-17.05	27.04	50.07	2.62 ± 0.053	2.35	-23.41
match-1000-L6	6.25	13.09	3.63 ± 0.060	**2.93**	-17.74	29.54	53.35	2.45 ± 0.051	2.36	-23.45
match-1000-L12	**5.45**	**10.37**	**3.72 ± 0.063**	2.92	-17.37	**26.58**	**49.40**	**2.64 ± 0.043**	2.35	-23.39
unmatch-50-L6	10.25	25.42	3.29 ± 0.093	2.74	-19.44	43.67	83.75	2.12 ± 0.067	2.18	-23.92
unmatch-50-L12	8.90	20.51	3.41 ± 0.079	2.69	-18.94	38.06	77.29	2.18 ± 0.059	2.21	-23.61
unmatch-200-L6	8.84	20.13	3.37 ± 0.069	2.78	-17.46	33.58	62.76	2.43 ± 0.053	2.33	-23.80
unmatch-200-L12	7.37	15.42	3.47 ± 0.058	2.84	**-16.73**	28.22	52.32	2.59 ± 0.047	2.37	-23.44
unmatch-1000-L6	6.31	13.96	3.61 ± 0.057	2.80	-17.09	31.83	59.28	2.45 ± 0.054	2.36	**-23.07**
unmatch-1000-L12	6.29	13.59	3.64 ± 0.065	2.87	-16.69	27.45	50.81	2.57 ± 0.048	**2.40**	-23.47

The Phoneme Error Rate (PER) was investigated using the speech of the GT test dataset as a reference, with each speech is analyzed as hypothesis. We used the phoneme recognizer, *allosaurus* [28], to obtain phoneme directly from the speech.

The average error rates obtained for speech synthesized from ground-truth scripts, ASR model, and SSL model input representations are shown in Table 2. For both languages, speech synthesis using the GT script showed the lowest error rate. On the other hand, for label-less cases, synthesis through the ASR model showed lower error rates than synthesis through the SSL model. Also, it was confirmed that the relationship between PER and WER or CER showed a proportional relationship with all data. This shows that the intelligibility consistently affects the errors from low-level elements like phonemes to high-level elements like words. These results suggest that using input representations obtained through ASR, which is natural text, leads to better conveyance of linguistic elements like intelligibility. This aligns with the tendency in the ZRC's prior study, where the topline using text transcription for both English and other languages showed superior ABX values in the ABX task used for phoneme discrimination [13].

Moreover, regarding the intelligibility of synthesized speech through language matching, the language-matched model's results showed slightly lower error rates than the unmatched model's, showing that language dependence of SSL model has non-negligible consequences. For the codec length, performance improved as the codec length of k-means clustering increased, reaching a level comparable to the ASR result in the case of English WER. In addition, in all cases, the synthesized speech through the 12th layer was able to obtain a lower error rate than the synthesized speech through the 6th layer. This is consistent with the results of previous studies that the layer close to the final output has more 'semantic' information than the intermediate layer. [25]

B. Speech Naturalness

In the case of ZRC, the subjective metric Mean Opinion Score (MOS) was used to evaluate naturalness, which can be considered an indicator that comprehensively evaluates both linguistic and extra-linguistic information. On the other hand, this study employed the pre-trained UTMOS model. UTMOS, trained on multiple languages including English, predicts automated MOS. UTMOS is evaluated independently without a comparative subject, allowing for the analysis of absolute linguistic and paralinguistic information, but making it impossible to compare voice changes between GT and synthesized speech.

The average UTMOS values obtained for speech synthesized from correct labels, ASR model, and SSL model input representations are shown in Table 2. [2] Common across all languages, speech synthesis using correct scripts showed the highest values. Meanwhile, under label-less synthesis conditions, the MOS for synthesized speech from SSL models was slightly higher across all language conditions than for those from ASR models, with this difference being greater in English synthesized speech than in Japanese. Additionally, speech2unit-unit2speech pairs of the same language showed predominant MOS scores for both languages.[3]

Moreover, the performance improved as codec length increased; however, while there was an big increase without exception when the token increasing from 50 to 200, there was a mild increase or even decrease when the token increasing from 200 to 1000. Thus, it can be said that naturalness is correlated to intelligibility, but this correlation decreases as the number of tokens increases. Also, the output obtained in the 12th layer obtained better results in terms of naturalness compared to those obtained in the 6th layer, following the trend has shown in the intelligibility: while some errors are within confidence levels.

These results propose a hypothesis that SSL discrete rep-

[2]Although 95 % confidence level was indicated in the table, as UTMOS is not a subjective value but an objective value from the model, the value was calculated from the standard deviation of the sample.

[3]The difference in absolute UTMOS values between Japanese and English can be attributed to the UTMOS evaluation model being primarily trained on English speech data.

resentations containing paralinguistic information like accents and intonations can enhance the naturalness of synthesized speech more than input representations from ASR models, suggesting that this can change depending on the language dependency of the SSL model and the amount of target language data it was trained on. It also suggests that increasing the language dependency of SSL models could incorporate more language-appropriate paralinguistic information within discrete representations. This hypothesis aligns with the results in ZRC, where the synthesized speech results for Indonesian, a language not trained in ZRC, more frequently failed to surpass the MOS of top-line models utilizing text transcription for English, a trained language [13].

C. Audio Quality and Noisiness

Following the evaluation of linguistic and naturalness aspects, which are related to speech language, the overall acoustic elements of the synthesized speech were assessed through quality evaluation.

Initially, the quality of digital audio was evaluated based on codecs, considering input representations as compressed audio representations, which can be viewed as neural speech codecs. Therefore, assuming the speech synthesis system as a single system, the output audio was analyzed as degraded audio compared to a reference. For such evaluations, PESQ is a representative metric, but in this case, speaker information and speech type can influence the results. Thus, to compensate for temporal mismatches between GT and resynthesized signals internally and to be resilient against errors commonly occurring in audio codecs, the WARP-Q metric was added to the evaluation. To address this, synthesized speech using correct scripts was used as a reference, and synthesized speech using ASR and SSL was compared as hypotheses.

The average values of WARP-Q [29] obtained for speech synthesized from ASR model and SSL model input representations are shown in Table 2. Generally, synthesized speech generated through the SSL model showed higher metrics compared to that synthesized through the ASR model. However, although there was a slight tendency due to Transformer layer and token length, differences due to the SSL model's encoder did not show a consistent trend across languages or tasks, with only minor increases or decreases observed.

The quality of audio in terms of noise inclusion was also evaluated. Signal Distortion Rate (SDR), although an indicator used in source separation, can be considered for evaluating the degree of noise inclusion in the output audio relative to the input audio, hence indicative of acoustic quality. The average SDR values obtained for speech synthesized from ASR model and SSL model input representations are shown in the table. Overall, although the values were low, synthesized speech through the SSL model showed slightly better metrics. Moreover, regarding differences due to the SSL model's encoder, systems crossing languages for English synthesized speech showed better results, while for Japanese synthesized speech, systems not crossing languages did better, indicating no clear trend. There is little trend for Transformer layer and token length as well.

Thus, speech synthesis models generated using speech recognition showed slightly more codec distortion and noise compared to those using SSL models, and synthesized speech's dependency on token language did not bring significant differences in speech quality.

V. Conclusion

This study investigated input representations in speech synthesis systems and created synthesized speech through systems constructed using each type of input representation. These were then compared and analyzed across languages and representations through newly considered metrics and results from prior research.

The findings revealed that no input representation demonstrated higher metrics than correct labels. Natural language input representations through speech recognition models showed dominance in linguistic vocal aspects, while discrete symbol input representations through self-supervised learning models were superior in aspects of naturalness and acoustic quality. Furthermore, the study highlighted differences in self-supervised learning model performance based on language dependency initiated by the language of the data. Also, the metrics shown overall improvement as the codec length of k-means clustering increased, more dramatically in the speech language metrics than the acoustic metrics. Furthermore, regardless of the language using, it is able to be found that Transformer layers close to the final output in linguistic elements had more semantic information than those that did not, and it was possible to demonstrate the tendency of previous studies.

Future efforts will focus on devising methods to minimize such language dependency. The goal is to advance the evaluation of tasks in more multilingual contexts than currently possible by transforming text into language-informed discrete tokens through tokenizers without preprocessing the text according to language-specific rules.

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*, 2018. arXiv: 1712.05884.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. arXiv: 1807.03748.

[8] 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.

[9] 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.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[11] E. Dunbar, J. Karadayi, M. Bernard, *et al.*, "The Zero Resource Speech Challenge 2020: Discovering Discrete Subword and Word Units," in *Proc. Interspeech 2020*, 2020, pp. 4831–4835.

[12] E. Dunbar, M. Bernard, N. Hamilakis, *et al.*, "The zero resource speech challenge 2021: Spoken language modelling," in *Proc. Interspeech 2021*, 2021, pp. 1574–1578.

[13] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1211–1226, 2022.

[14] 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," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6562–6566, 2021.

[15] X. Li, Y. Jia, and C.-C. Chiu, "Textless direct speech-to-speech translation with discrete speech representation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096797.

[16] H. Kim, S. Kim, J. Yeom, and S. Yoon, "UnitSpeech: Speaker-adaptive Speech Synthesis with Untranscribed Data," in *Proc. INTERSPEECH 2023*, 2023, pp. 3038–3042. DOI: 10.21437/Interspeech.2023-2326.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[18] K. Ito and L. Johnson, *The lj speech dataset*, https://keithito.com/LJ-Speech-Dataset/, 2017.

[19] Y. Yue, M. Daijiro, and F. Seiji, *Reazonspeech: A free and massive corpus for Japanese ASR*, https://research.reazon.jp/_static/reazonspeech_nlp2023.pdf, 2023.

[20] 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.

[21] 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.

[22] 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.

[23] S. Takamichi, N. Wataru, T. Naoko, and S. Hiroshi, "J-MAC: Japanese multi-speaker audiobook corpus for speech synthesis," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association*, ISCA, 2022, pp. 2358–2362.

[24] 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.

[25] P. Kumar, V. N. Sukhadia, and S. Umesh, "Investigation of robustness of hubert features from different layers to domain, accent and language variations," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6887–6891.

[26] 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.

[27] 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.

[28] X. Li, S. Dalmia, J. Li, *et al.*, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8249–8253.

[29] 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, pp. 1050–1070, 2022.