Interspeech 2024
1-5 September 2024, Kos, Greece

Audio-conditioned phonemic and prosodic annotation for building text-to-speech models from unlabeled speech data

Yuma Shirahata[1*], *Byeongseon Park*[1*], *Ryuichi Yamamoto*[1], *Kentaro Tachibana*[1]

[1]LY Corporation., Japan

yuma.shirahata@lycorp.co.jp

Abstract

This paper proposes an audio-conditioned phonemic and prosodic annotation model for building text-to-speech (TTS) datasets from unlabeled speech samples. For creating a TTS dataset that consists of label-speech paired data, the proposed annotation model leverages an automatic speech recognition (ASR) model to obtain phonemic and prosodic labels from unlabeled speech samples. By fine-tuning a large-scale pre-trained ASR model, we can construct the annotation model using a limited amount of label-speech paired data within an existing TTS dataset. To alleviate the shortage of label-speech paired data for training the annotation model, we generate pseudo label-speech paired data using text-only corpora and an auxiliary TTS model. This TTS model is also trained with the existing TTS dataset. Experimental results show that the TTS model trained with the dataset created by the proposed annotation method can synthesize speech as naturally as the one trained with a fully-labeled dataset.

Index Terms: prosodic annotation, unlabeled data, text-to-speech, data augmentation

1. Introduction

The field of Text-to-speech (TTS) has experienced significant progress owing to the rapid advancements of deep neural network-based approaches [1].

For training TTS models, a sufficient amount of speech-text paired data is essential. While collecting a large amount of unlabeled speech data is comparatively straightforward as demonstrated by the datasets used for training audio self-supervised learning (SSL) models [2, 3, 4], the latter often necessitates accurate phonemic and prosodic labels for the development of high-quality TTS systems [5, 6, 7], which are challenging to obtain in large quantities. Thus, the acquisition of reliable labels from speech is crucial to leverage the vast amounts of unlabeled speech data in the TTS field.

To obtain phonemic and prosodic labels from unlabeled speech, a typical approach is the sequential application of automatic speech recognition (ASR) models followed by text processing [5, 8, 9]: 1) employing ASR models that output grapheme sequences given unlabeled speech samples; 2) performing text-based processing such as grapheme-to-phoneme (G2P) conversion [10, 11] and prosody prediction [12, 13] on the output of the ASR model. A key advantage of this approach is the use of extensive dictionary data and ASR models trained on large text corpora. Nonetheless, the task of predicting phonemic and prosodic labels from grapheme sequences inherently presents a one-to-many mapping challenge, making accurate annotation difficult without audio information. This is because a text can be interpreted and vocalized in multiple ways, influenced by factors such as the speaker's dialect, age, and speech disfluencies, among others.

On the other hand, there are some studies that utilize audio information to annotate prosodic labels on speech samples for creating TTS datasets [14, 15]. These studies successfully improved the accuracy of prosody prediction owing to the information derived from input speech. However, they are limited to scenarios where the correct text and phonemic information are provided. Research has not yet advanced to address performance on entirely unlabeled speech data, which represents a more realistic scenario.

To address the limitations of the previous works, this paper proposes an annotation model that predicts phonemic and prosodic labels (hereinafter *TTS labels*) simultaneously from unlabeled speech data, conditioned on input speech information. For creating a TTS dataset from unlabeled speech samples, the proposed annotation model leverages an ASR model to obtain TTS labels corresponding to the input speech samples. Specifically, we can construct the annotation model by fine-tuning a large-scale pre-trained ASR model with a limited amount of labeled speech data within an existing TTS dataset. Furthermore, to address the challenge of amassing a sufficient amount of label-speech paired data for training the annotation model, we propose a data augmentation method utilizing TTS. In this method, an auxiliary TTS model is first trained on a limited amount of label-speech paired data within the existing TTS dataset, and the model is then used to generate pseudo label-speech paired data from text-only corpora. The combination of the pre-trained ASR model and data augmentation enables the construction of a model capable of generating highly accurate TTS labels, even with a limited amount of label-speech paired dataset. For the architecture of the annotation model, we adopted the Transformer for its superior ability in sequence-to-sequence problems [16]. The model receives raw speech sequences as input and predicts the corresponding TTS labels in an auto-regressive manner. Once the annotation model is trained, it is applied to unlabeled speech samples to get the label-speech paired data for TTS model training.

Through experiments, we find that the proposed method is able to annotate unlabeled speech more accurately than the baseline method that cascades an ASR model and text processing even when the number of the ground truth labels is less than 5,000 samples of a single speaker (character error rate (CER) on phonemic label prediction: 6.45% vs. 2.44%, F_1 score on prosodic label prediction: 68.51% vs. 95.96%). Moreover, TTS models trained with the TTS datasets generated by the proposed method achieved comparable performance to those trained with the fully-labeled $f$dataset in terms of naturalness. Audio samples are available on our demo page[1].

2. Method

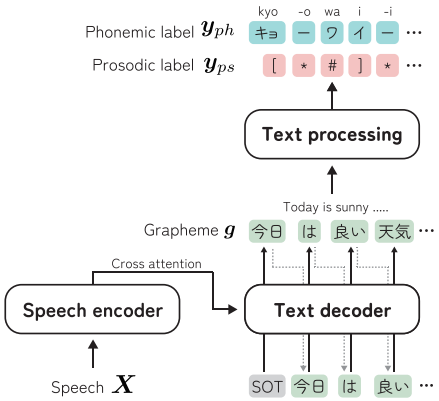
2.1. Problem formulation

To train a TTS model from unlabeled speech data, this study aims to construct an annotation model that can estimate a TTS

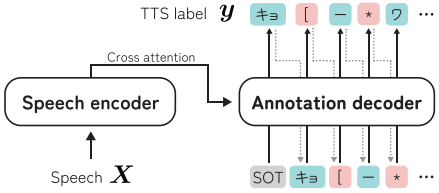
*Equal contribution.

[1]https://yshira116.github.io/pp_annotation/

10.21437/Interspeech.2024-342



(a) baseline framework



(b) proposed framework

Figure 1: *Overview of the baseline and the proposed annotation model. In the baseline framework, phonemic and prosodic labels are predicted from the grapheme sequence. In contrast, they are predicted directly from speech in the proposed method.*

label sequence $\boldsymbol{y} = \{y_m \in \mathcal{Y}\}_{m=1}^{M}$ from an unlabeled speech sample $\boldsymbol{X} = \{x_n \in \mathbb{R}^{D_{in}}\}_{n=1}^{N}$. Here, $\mathcal{Y}$ and M are the vocabulary of TTS input tokens (i.e., a mixed vocabulary of phonemic and prosodic labels) and the length of output TTS labels, D_{in} and N are the dimensions of acoustic features of input speech and its length, respectively. In mathematical terms, we optimize the following conditional likelihood objective:

$$L = p(\boldsymbol{y}|\boldsymbol{X}). \tag{1}$$

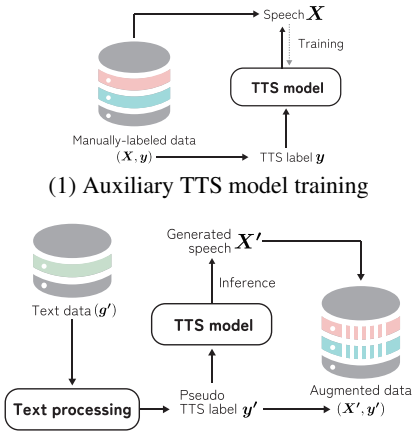
However, since $\boldsymbol{y}$ is a mixed representation of multiple sequences and difficult to predict at once, the following conditional dependency assumption is typically introduced in previous works:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X}) &= p(\boldsymbol{y}_{\mathrm{ph}}, \boldsymbol{y}_{\mathrm{ps}}|\boldsymbol{X}) \\ &= p(\boldsymbol{y}_{\mathrm{ph}}, \boldsymbol{y}_{\mathrm{ps}}|\boldsymbol{g})p(\boldsymbol{g}|\boldsymbol{X}), \end{aligned} \tag{2}$$

where $\boldsymbol{g}$, $\boldsymbol{y}_{\mathrm{ph}}$, and $\boldsymbol{y}_{\mathrm{ps}}$ are the corresponding grapheme sequence, phonemic label sequence, and prosodic label sequence, respectively. In (2), since the first term is independent of speech $\boldsymbol{X}$, it can be optimized using only text-based methods. In addition, since many high-quality grapheme-based ASR models are readily available online [17, 18], the optimization of the second term is also straightforward. However, since the first term cannot consider the speech information to estimate the label sequence, this method is inherently accompanied by errors in G2P and prosodic label estimation, which results in a suboptimal prediction. The overview of this method is depicted in Fig. 1 (a). To overcome this problem, we propose a model that directly optimizes (1) in 2.2.

2.2. Annotation model

The overview of the proposed annotation model is shown in Fig. 1 (b). Following successful prior works that predict a mixture of multiple sequences as a single sequence [19, 20, 21],



(1) Auxiliary TTS model training

(2) Data augmentation by the auxiliary TTS model

Figure 2: *Overview of the data augmentation method with the auxiliary TTS model. (1) First, the auxiliary TTS model is trained using manually labeled data. (2) Then, the auxiliary TTS model is used to generate augmented paired data from text-only corpora.*

we adopted the encoder-decoder Transformer architecture as the base structure of the annotation model. The model is composed of the speech encoder and the annotation decoder. The speech encoder encodes the input acoustic feature sequence $\boldsymbol{X}$ into a hidden speech embedding sequence. The annotation decoder then generates the corresponding TTS label sequence $\boldsymbol{y}$ conditioned on the embedding sequence in an auto-regressive manner:

$$\log p(\boldsymbol{y}|\boldsymbol{X}) = \sum_{m=1}^{M} \log p(y_m|y_1, \ldots, y_{m-1}, \boldsymbol{X}). \tag{3}$$

The annotation model is trained on a paired dataset of $(\boldsymbol{X}, \boldsymbol{y})$, to minimize the cross entropy loss of the model outputs and ground truth labels. During inference, given an unlabeled speech sample $\boldsymbol{X}$, the model infers the corresponding TTS label sequence $\hat{\boldsymbol{y}}$ as follows:

$$\hat{\boldsymbol{y}} = \underset{y \in \mathcal{Y}^*}{\arg\max}\, p(\boldsymbol{y}|\boldsymbol{X}), \tag{4}$$

where $\mathcal{Y}^*$ denotes a set of all possible hypotheses.

2.3. Text-to-speech data augmentation

Although we can train the annotation model with paired data consisting of $(\boldsymbol{X}, \boldsymbol{y})$, amassing substantial annotated data often proves challenging. This is because accurately labeling speech samples requires specialized expertise and is notably time-consuming. To deal with this issue, we propose a data augmentation method using an auxiliary TTS model. The overview of the proposed TTS data augmentation method is described in Fig. 2. As shown in Fig. 2, we first train an auxiliary TTS model $\boldsymbol{M}$ with a limited size of label-speech paired dataset $\boldsymbol{D} = \{\boldsymbol{X_i}, \boldsymbol{y_i}\}_{i=1}^{K}$, where K denotes the number of training samples with manually-annotated labels. Second, we prepare a large-scale text-only dataset $\boldsymbol{D'_g} = \{\boldsymbol{g'_i}\}_{i=1}^{K'}$ that has only grapheme sequences. Here, $\boldsymbol{K'}$ is the number of samples in the text dataset. Third, a text processing module is used to generate pseudo TTS labels $\boldsymbol{D'_y} = \{\boldsymbol{y'_i}\}_{i=1}^{K'}$ from $\boldsymbol{D'_g}$. Note that the text processing module here is not required to be correct, since the auxiliary TTS model $\boldsymbol{M}$ is expected to generate speech samples that are faithful to input TTS labels. In other words, if

the text processing module generates an incorrect phoneme sequence, the generated speech sample from it reflects the incorrect sequence, which is consistent as paired data for the training of the annotation model. Finally, $\boldsymbol{M}$ generates augmented speech samples $\{\boldsymbol{X'_i}\}_{i=1}^{K'}$ from $\boldsymbol{D'_y}$, and augmented training data $\boldsymbol{D'} = \{\boldsymbol{X'_i}, \boldsymbol{y'_i}\}_{i=1}^{K'}$ is obtained.

3. Experiments

To assess the performance of the proposed methods, we conducted two types of experiments. Section 3.1 objectively evaluates the accuracy of TTS labels generated from unlabeled speech datasets. Section 3.2 investigates the performance of the proposed method when applied to TTS tasks.

3.1. Annotation of unlabeled speech data

3.1.1. Experimental conditions

Dataset and pre-processing: For the training of the proposed annotation models, two datasets were prepared to investigate the performance of the models when trained on 1) a limited amount of labeled data, and 2) a large scale data with a variety of speakers. For the former, we adopted JSUT, which is a public Japanese speech corpus uttered by a single female speaker [22]. We used the *basic5000* subset and its manual TTS labels[2]. The dataset consists of 5,000 text samples and 6.78 hours of speech. We split the data into 4,500 and 250 samples for training and validation, respectively. The remaining 250 samples were not used in this experiment. For the latter, we used proprietary Japanese speech corpora recorded by six male and eleven female Japanese professional speakers with manually annotated labels. The corpora consist of 173,987 samples and 207.96 hours of speech. We held out the samples of two males and two females for evaluation, and the other speakers' data was used for training and validation. The number of data for training, validation, and evaluation were 153,551, 4,449, and 15,987, respectively. Hereinafter, this dataset will be referred to as LARGE.

TTS data augmentation: In our experiment, TTS data augmentation was applied to the JSUT dataset. The model architecture of the TTS model for data augmentation was based on Period VITS [23]. We used the same configuration that will be described in 3.2.1. To exclude the bias of the text domain, the augmented text data $\boldsymbol{D'_g}$ was taken from the training set of the LARGE dataset (153,551 samples). For the text processing module, Open JTalk[3] was used. The total amount of augmented speech data was 115.5 hours. Note that the augmented samples by the TTS model trained on the JSUT dataset generally had a faster speed than the LARGE dataset, which resulted in a smaller data size for the same text set.

phonemic/prosodic labels: For phonemic and prosodic labels, we used Kurihara et al. [24]'s design, as depicted in Fig. 3. In the method, the prosodic status of each mora is represented by five labels considering the rules of the Japanese pitch accent in the Tokyo dialect. The details of the labels are as follows: (1) Pause ("_" in Fig. 3); (2) Low to high accent change ("[" in Fig. 3); (3) High to low accent change ("]" in Fig. 3); (4) Accentual phrase boundary ("#" in Fig. 3); (5) Raise-type boundary pitch movement (for question sentence, "?" in Fig. 3). In this experiment, we additionally introduced a padding token for a mora that does not apply to the five categories above ("*" in Fig. 3). For the phonemic labels, we used Japanese katakana characters to represent the Japanese phonemic status of each mora.

Model details: All the proposed annotation models were fine-tuned from the encoder-decoder-based public speech recognition model Whisper [17]. We used the *small*[4] model for all the experiments. We fine-tuned each model for 100k steps, with a batch size of 36. The learning rate was increased to 0.0002 with warm-up steps of 500, and then linearly decreased to reach zero at the 100k step. The parameters in the encoder part were frozen during fine-tuning to stabilize the training. Model checkpoints with the best validation loss were used for the evaluation. In addition to the proposed models, two text-based baseline models were also prepared. The systems used in our experiments are summarized below:

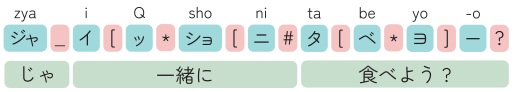


Figure 3: *Example of TTS labels for a Japanese text "じゃ一緒に食べよう?" (Well, let's eat together). The blue, red, and green squares denote phonemic labels, prosodic labels, and grapheme, respectively.*

ANNT-JSUT: Proposed annotation model trained on the manually annotated JSUT training data.

ANNT-JSUT-TTSAUG: Proposed annotation model trained on the manually annotated JSUT training data and TTS augmentation data.

ANNT-LARGE: Proposed annotation model trained on the manually annotated LARGE training data.

ASR-NLP: A baseline model that obtains grapheme transcription by Whisper small model and performs text-based post-processing to get TTS labels.

GT-NLP: A baseline model that obtains grapheme transcription from ground truth text data and performs text-based post-processing to get TTS labels.

For **ASR-NLP** and **GT-NLP**, Open JTalk was used to obtain the TTS labels from grapheme sequences.

3.1.2. Evaluation on annotation accuracy

To evaluate the performance of our proposed method on annotation tasks, we tested the models with 15,987 speech samples in our dataset. We used CER and F_1 scores as metrics to evaluate the phonemic and prosodic label annotation tasks, respectively. To independently evaluate phonemic and prosodic label annotation tasks, we separated phonemic and prosodic information from manually annotated ground truth and predicted labels. Hence, we only used phonemic labels for the calculation of CER. Since it is impossible to compare ground truth prosodic labels and predicted labels when predicted phonemic labels are corrupted, we only used 4,379 samples, in which all models correctly predicted phonemic labels of the test set, to evaluate the prosodic label annotation task. Additionally, for a fair comparison of the proposed and baseline methods, we excluded two prosodic labels on the evaluation: (1) Pause; (2) Raise-type boundary pitch movement. This is because ground truth texts of **GT-NLP** include these labels as punctuation.

Table 1 shows the performance of the models on TTS label annotation tasks. The findings are summarized as follows:

Baseline vs. Proposed model As shown in Table 1, the proposed model performed best in both metrics when a large amount of annotation data is available (i.e., **ANNT-LARGE**). Furthermore, all our proposed models outperformed the baseline methods on the prosodic label prediction tasks, even when the ground truth grapheme sequence is used in the latter (i.e., **GT-NLP**). The results imply that the utilization of audio information is significantly effective in TTS label prediction.

[2] https://github.com/sarulab-speech/jsut-label
[3] https://open-jtalk.sp.nitech.ac.jp/
[4] Larger models were not used due to the limitation of computational resources.

Table 1: *Objective evaluation results on each task. CER and Prosody F_1 are metrics for phonemic and prosodic label annotation tasks, respectively.*

Model	CER (↓)	Prosody F_1 (↑)
ASR-NLP	6.45%	68.51%
GT-NLP	2.53%	73.43%
ANNT-JSUT	6.12%	88.77%
ANNT-JSUT-TTSAUG	**2.44%**	**95.96%**
ANNT-LARGE	**0.54%**	**98.84%**

Effectiveness of data augmentation Table 1 also shows that the proposed model trained with the augmented data by our framework (i.e., **ANNT-JSUT-TTSAUG**) significantly outperformed the baseline methods and the model trained with limited-scale data (i.e., **ANNT-JSUT**). This confirms that the proposed TTS data augmentation method improves the performance of the annotation model, even if the augmented data is automatically generated from text-only corpora.

3.2. Application to text-to-speech

3.2.1. Experimental conditions

To investigate the robustness of the proposed method against dataset variation, three datasets were used for TTS experiments: JSUT, JVS [25], and the LARGE dataset described in 3.1.1. For JSUT, we split the data into 4,500, 250, and 250 samples for training, validation, and evaluation, respectively. Note that **ANNT-JSUT** and **ANNT-JSUT-TTSAUG** were excluded from the evaluation on JSUT as these models used the same dataset for the training of annotation models. For JVS, we split the samples of *parallel100* subset into 90 and 10 samples for each speaker for training and validation, respectively. For testing, *nonpara30* subset was used. For LARGE, the held-out data in 3.1.1 was used for TTS experiments. The 15,987 samples were split into 14,000, 1,000, and 987 samples for training, validation, and evaluation, respectively.

We adopted the Period VITS architecture for our TTS model due to its high-quality speech generation capability [23]. We followed the settings of the original paper with two exceptions: 1) we did not use an emotion encoder, since no emotional dataset was used in the TTS experiments; 2) the training step was set to 200k based on the results of preliminary experiments. Since Period-VITS requires duration information of each phoneme, we trained a forced alignment model based on Gaussian mixture model and hidden Markov model (GMM-HMM) [26] on ReazonSpeech dataset [27], and used it to obtain phoneme alignment.

In addition to the TTS labels generated by the models in 3.1, two types of TTS labels were used in TTS experiments:

ORACLE: This model uses manually annotated labels.

ORACLE-WO-ACC: This model uses manually annotated labels, but drops prosodic labels. This model was introduced to assess the importance of prosodic labels.

Since manual annotation data was unavailable for JVS dataset, **ORACLE** was not trained, and **ORACLE-WO-ACC** was substituted with the phoneme sequences from Open JTalk with ground truth text. This model is referred to as **GT-NLP-WO-ACC**.

3.2.2. Evaluation on Text-to-speech

To evaluate the effectiveness of our proposed method on TTS tasks, we conducted subjective listening tests on the generated samples. These tests were based on the mean opinion score (MOS) of a five-point scale: 1 = Bad; 2 = Poor; 3 = Fair; 4 = Good; and 5 = Excellent. We asked native Japanese raters to make a quality judgment in terms of prosodic naturalness and pronunciation correctness. We showed the grapheme text to the raters during the listening tests to help accurately judge the naturalness of the prosody and pronunciation. The number of raters was eleven. For each of the three datasets, 50 sentences were randomly chosen from the evaluation set. Then, ground truth labels were used[5] to generate speech samples for each system. Since ground truth labels for JVS dataset were unavailable, we manually annotated the evaluation set.

Table 2: *MOS test results on different datasets with 95% confidence intervals. Note that **Reference** denotes recorded speech samples.*

Model	JSUT	JVS	LARGE
GT-NLP-WO-ACC	-	$2.52_{\pm 0.11}$	-
ORACLE-WO-ACC	$2.79_{\pm 0.11}$	-	$3.36_{\pm 0.12}$
ASR-NLP	$3.65_{\pm 0.11}$	$3.43_{\pm 0.11}$	$4.04_{\pm 0.09}$
GT-NLP	$3.69_{\pm 0.10}$	$3.75_{\pm 0.10}$	$4.05_{\pm 0.09}$
ANNT-JSUT	-	$3.77_{\pm 0.10}$	$4.26_{\pm 0.08}$
ANNT-JSUT-TTSAUG	-	$\mathbf{3.95}_{\pm 0.09}$	$\mathbf{4.33}_{\pm 0.08}$
ANNT-LARGE	$\mathbf{4.11}_{\pm 0.09}$	$3.75_{\pm 0.10}$	$4.29_{\pm 0.09}$
ORACLE	$4.15_{\pm 0.09}$	-	$4.22_{\pm 0.09}$
Reference	$3.99_{\pm 0.10}$	$4.39_{\pm 0.09}$	$4.64_{\pm 0.07}$

Table 2 summarizes the results of subjective evaluation. Firstly, the MOS scores are significantly lower for the TTS models lacking accent information than the others. This confirms that prosodic labels are quite important in improving the naturalness of Japanese speech synthesis, as reported in previous works [6, 13, 24]. We can also see that the proposed methods constantly outperform the baseline methods on all datasets, which is consistent with the results of objective evaluation on annotation accuracy. Moreover, for JSUT and LARGE datasets, the TTS models trained on the labels generated by proposed methods perform comparable or slightly better than those trained on oracle labels. This result indicates that the proposed method has the capability to generate a sufficiently high-fidelity TTS system from unlabeled speech data. Interestingly, for JVS dataset, **ANNT-JSUT-TTSAUG** achieved a higher score than **ANNT-LARGE**, which performed the best in objective evaluation. One possible reason is that while the proposed TTS data augmentation method can generate consistent label-speech paired data through the auxiliary TTS model, manually annotated labels could be noisy due to the inconsistent annotation across multiple annotators, which made it difficult for the annotation model to learn the correct mapping. This result also suggests that the proposed TTS data augmentation method is still effective when applied to TTS tasks. For JSUT dataset, some TTS models got higher scores than the reference. This is likely due to the inclusion of unclear pronunciations and lip noise in some of the reference audio samples.

4. Conclusions

In this paper, we proposed an annotation model for building high-fidelity TTS systems from unlabeled speech data. The proposed model predicts phonemic and prosodic label sequences from speech input. To address the challenge of collecting a sufficient amount of labeled data for model training, a data augmentation method utilizing the TTS model was proposed. The proposed model generated accurate TTS labels, enabling high-quality TTS models even when the number of manually annotated data is limited. Future work includes applying our approach to more challenging speech samples, including those with emotional content or pronounced dialectal variations.

[5]Using labels from the text processing model would be another option, but we used ground truth labels to minimize the errors derived from the input labels and focus on the quality of the TTS models.

5. References

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

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[3] 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.

[4] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[5] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *Advances in neural information processing systems*, vol. 36, 2024.

[6] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. ICASSP*, 2019, pp. 6905–6909.

[7] J. Pan, X. Yin, Z. Zhang, S. Liu, Y. Zhang, Z. Ma, and Y. Wang, "A unified sequence-to-sequence front-end model for mandarin text-to-speech synthesis," in *Proc. ICASSP*, 2020, pp. 6689–6693.

[8] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[9] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[10] S. F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *Proc. Eurospeech*, 2003, pp. 2033–2036.

[11] M.-J. Chae, K. Park, J. Bang, S. Suh, J. Park, N. Kim, and L. Park, "Convolutional sequence to sequence model with non-sequential greedy decoding for grapheme to phoneme conversion," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2486–2490.

[12] A. Rosenberg, "AutoBI: A tool for automatic tobi annotation." in *Proc. Interspeech*, 2010, pp. 146–149.

[13] 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, pp. 1931–1935.

[14] Z. Dai, J. Yu, Y. Wang, N. Chen, Y. Bian, G. Li, D. Cai, and D. Yu, "Automatic prosody annotation with pre-trained text-speech model," *arXiv preprint arXiv:2206.07956*, 2022.

[15] X. Yuan, R. Feng, and M. Ye, "Low-resource mongolian speech synthesis based on automatic prosody annotation," *arXiv preprint arXiv:2211.09365*, 2022.

[16] 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, pp. 5998–6008.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[18] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[19] M. Omachi, Y. Fujita, S. Watanabe, and M. Wiesner, "End-to-end ASR to jointly predict transcriptions and linguistic annotations," in *Proc. NAACL*, 2021, pp. 1861–1871.

[20] K. Audhkhasi, B. Kingsbury, B. Ramabhadran, G. Saon, and M. Picheny, "Building competitive direct acoustics-to-word models for English conversational speech recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 4759–4763.

[21] L. E. Shafey, H. Soltau, and I. Shafran, "Joint speech recognition and speaker diarization via sequence transduction," in *Proc. Interspeech*, 2019, pp. 396–400.

[22] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[23] Y. Shirahata, R. Yamamoto, E. Song, R. Terashima, J.-M. Kim, and K. Tachibana, "Period VITS: Variational inference with explicit pitch modeling for end-to-end emotional speech synthesis," in *Proc. ICASSP*, 2023, pp. 1–5.

[24] K. Kurihara, N. Seiyama, and T. Kumano, "Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS," *IEICE Transactions on Information and Systems*, vol. E104-D, no. 2, pp. 302–311, 2021.

[25] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[26] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

[27] Y. Yin, D. Mori, and S. Fujimoto, "ReazonSpeech: A free and massive corpus for Japanese ASR," in *Association for Natural Language Processing annual meeting (in Japanese)*, vol. 2023, 2023.