

TEXT-TO-SPEECH SYNTHESIS FROM DARK DATA WITH EVALUATION-IN-THE-LOOP DATA SELECTION

Kentaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari

The University of Tokyo, Japan.

ABSTRACT

This paper proposes a method for selecting training data for text-to-speech (TTS) synthesis from dark data. TTS models are typically trained on high-quality speech corpora that cost much time and money for data collection, which makes it very challenging to increase speaker variation. In contrast, there is a large amount of data whose availability is unknown (a.k.a., “dark data”), such as YouTube videos. To utilize data other than TTS corpora, previous studies have selected speech data from the corpora on the basis of acoustic quality. However, considering that TTS models robust to data noise have been proposed, we should select data on the basis of its importance as training data to the given TTS model, not the quality of speech itself. Our method with a loop of training and evaluation selects training data on the basis of the automatically predicted quality of synthetic speech of a given TTS model. Results of evaluations using YouTube data reveal that our method outperforms the conventional acoustic-quality-based method.

Index Terms— text-to-speech synthesis, dark data, automatic speech quality evaluation, data selection, data cleansing, YouTube

1. INTRODUCTION

With large speech corpora and the development of sequence-to-sequence models, recent text-to-speech (TTS) models have achieved human-like speech synthesis at a level comparable to human speech utterances [1]–[3]. Multi-speaker TTS can synthesize speech with the desired speaker characteristics by inputting speaker information in addition to text [4]–[6]. TTS training typically requires pairs of text and speech. The speech data is recorded in a well-designed environment [7]–[9], e.g., a recording studio with a professional speaker. Furthermore, constructing a multi-speaker TTS corpus is even more burdensome. This fact significantly limits the variety of speakers that TTS can synthesize. Fig. 1 is an example of speaker distributions of a famous multi-speaker TTS corpus (JVS [9]) and one we deal with in this paper. This is a t-SNE [10] plot of x -vectors [11] of natural speech in the two corpora. Typical multi-speaker TTS corpora only cover limited speaker variation.

In contrast, there is a large amount of speech data (e.g., YouTube videos) with unknown potential for machine learning, known as *dark data* [12], [13], on the Internet. We expect to solve the small speaker variation problem if we can fully automate TTS model training from dark data. Related to this, methods have been proposed for building TTS corpora from datasets other than TTS corpora [14]–[17]. These methods determine a TTS corpus on the basis of the acoustic quality of each utterance. On the other hand, recent TTS models are becoming more robust to data noise than the old-fashioned parametric models [18]–[20]. In other words, low-quality speech data does not necessarily have a negative impact on TTS model training. In light

This work is supported by JSPS KAKENHI 22H03639 and Moonshot R&D Grant Number JPMJPS2011. We also appreciate Kenta Udagawa of the University of Tokyo for his help.

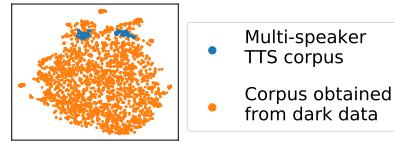


Fig. 1: Distributions in speaker space for multi-speaker TTS corpus [9] and dark data (e.g., data from YouTube [21]). Compared with speakers in the massive dark data, speakers in TTS corpus are very limited and localized.

of the above, TTS corpus construction should be based on the importance of the speech data as training data for a given TTS model, not on the quality of speech data itself.

This paper proposes an evaluation-in-the-loop data selection method for TTS model training from dark data. Our method uses a pseudo score on the perceptual quality of synthetic speech output from a given TTS model and scores each of the training data candidates in terms of the importance of training data rather than acoustic quality. The training data candidates are filtered by the importance. The procedure automates the selection and training loop using a pre-trained model that automatically predicts the mean opinion score (MOS) of synthetic speech on naturalness. This paper also proposes a method for pre-screening dark data for TTS use, achieving a fully automated process from dark data collection to TTS model training. We conduct experiments with dark data, in particular, actual data obtained from YouTube. The results show that our evaluation-in-the-loop method achieves better than the conventional acoustic-quality-based method. The contributions of this work are as follows:

- We propose an evaluation-in-the-loop data selection method for TTS model training, that enables more efficient selection than the conventional acoustic-quality-based method.
- We conduct experiments using actual data downloaded from YouTube and demonstrate the validity of the proposed method.

2. RELATED WORK

Multi-speaker and noise-robust TTS. Multi-speaker TTS uses speaker representations to control the speaker of the synthetic speech [4], e.g., the x -vector [11] of automatic speaker verification (ASV); we follow this method in this paper. Also, a few-shot speaker adaptation, voice building of a new speaker from a small amount of speech data [6], is a promising application of multi-speaker TTS. To increase the number of speakers by using diverse training data, there exist methods for noise-aware TTS training [18]–[20].

Building TTS corpus on automatic speech recognition (ASR) corpus. The use of more speakers’ data in training increases the speaker variation in TTS [5]. Methods based on selecting data from ASR corpora have been proposed. The successful examples are based on acoustic noise [5] and text label noise (i.e., a mismatch between text and speech) [14]. However, they are based only on data quality and are performed independently from TTS models. They

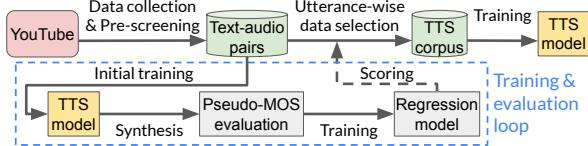


Fig. 2: Procedure of proposed method. We obtain dark data from YouTube and evaluate each utterance through loop of TTS model training and synthetic-speech evaluation. It finally builds TTS corpus from dark data by utterance-wise filtering.

are not appropriate considering the noise robustness of the models, which we described above.

In addition, dark data has the potential to build massive corpora almost fully automatically, and indeed, successful examples of this have been reported in the construction of ASR corpora [22], [23].

Prediction of perceptual quality in synthetic speech. Methods for automatically predicting the subjective evaluation score of synthetic speech have been studied to reduce the evaluation cost for TTS [24], [25]. The challenge in automatic prediction is generalization performance, i.e., the robustness of the prediction against divergence between training and evaluation data. The recent deep learning-based models (used in this paper) have relatively good accuracy in predicting the relative rank between speech samples, even if the prediction value itself is not accurate [26].

3. PROPOSED METHOD

3.1. Data collection and pre-screening

As shown in Fig. 2, the first step is to collect text-audio pairs and pre-screen dark data to filter too low-quality data for TTS training. We collect a dataset from YouTube and combine data screening methods for ASR and ASV. We briefly describe each method below, but see the paper [21] for more detail.

Cleansing based on text-audio alignment accuracy. Speech data in TTS training data must align well with its transcription. We calculate the connectionist temporal classification (CTC) score [27] to quantify how well speech utterances fit to text (YouTube subtitles in this case). Audio data is divided into each utterance by CTC segmentation, and utterances with lower scores are eliminated.

Cleansing based on speaker compactness. Having multiple utterances for each speaker is desirable for TTS training data. We calculate the x -vector [11] variance within each utterance group (utterances belonging to one video in the YouTube case) to quantify how stable the speaker representation is in the multiple utterances. Utterance groups with lower this scores are filtered out as well as above. After the filtering, utterances within one group are considered to be by a unique speaker. Note that this method implicitly rejects single speaker utterances with multiple styles or large fluctuations in style because the x -vector varies greatly depending on the speaking style [28], even for the same speaker. We used this method because we focus on speaker variation rather than speaking-style variation.

3.2. Data selection using automated evaluation-in-the-loop

Our method aims to estimate the quality of each data in terms of *training data for the given TTS model*. For each speech data, we predict the pseudo perceptual naturalness of the synthetic speech when the TTS model is trained on the speech data. We use this score for the quality score of each data. The TTS model is finally trained using the TTS training data obtained via the evaluation-in-the-loop selection.

3.2.1. Initial training using pre-screened data

We perform the initial training of the given TTS model, using all the data obtained from the pre-screening.

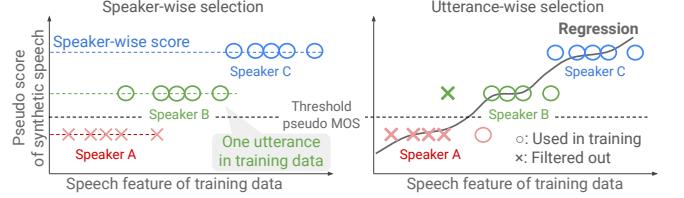


Fig. 3: Comparison of speaker-wise and utterance-wise selection. With regression, we filter out low-score utterances even if speaker’s pseudo MOS is high.

3.2.2. Evaluating synthetic speech quality

We evaluate the quality of speech synthesized by the initially-trained TTS model. A simple evaluation method is to calculate the value of the loss function for each utterance during training (e.g., the distance between the ground-truth and the predicted features). However, the distance does not necessarily correspond to the perceptual quality of the synthesized speech [29], [30].

Since MOS directly reflects perceptual quality, another method is to conduct subjective evaluation. However, it does not scale well for training with a massive amount of data. We leverage pseudo MOS predicted with an automatic MOS prediction model. As described in Section 2, current MOS prediction models have achieved a generalization ability. Therefore, we use a pseudo MOS score of subjective evaluation on naturalness predicted by a pre-trained MOS prediction model.

This evaluation is used to determine the score of each training data, as described in the following Section 3.2.3. In other words, this evaluation aims to estimate the difference in the effect of each data on the quality of the synthesized speech. The simplest method is to synthesize training data sentences and to filter out sentences with lower scores. However, this method uses different sentences among speakers. It is inappropriate because 1) the pseudo MOS score changes depending on the sentence to be synthesized [31] and 2) a sentence set greatly varies among speakers in dark data. Therefore, we evaluate the quality of synthesized speech for each speaker, using common sentences. In this way, it is possible to quantify the difference in synthetic speech quality of each speaker, without depending on the speakers’ utterances. Therefore, we evaluate the quality of synthesized speech for each speaker using common sentences not included in the training data, and averages of the values are used for each speaker.

3.2.3. Quantifying score as training data for each utterance

We filter the training data on the basis of the obtained speaker-wise pseudo MOS. The simplest method is to filter at the speaker level on the basis of the values, i.e., filtering out speakers with lower pseudo MOSs. However, since the data quality varies within the same speaker, filtering should be performed at the utterance level. In other words, we should filter out low-quality utterances even if speaker’s pseudo MOS is high, and vice versa as shown in Fig. 3.

For this purpose, we train a regression model that predicts the speaker-wise pseudo MOS from each utterance in the training data. We assume that acoustically similar training data will achieve similar naturalness in the synthesized speech. Furthermore, we assume that a regression model predicts close values for acoustically similar data. These assumptions motivate us to use a regression model.

We train this regression model on all the pre-screened data. We confirm the effect of using a regression model in the experiment described in Section 4.2.1.

3.3. Training data selection and re-training

We evaluate the training data at the utterance level with the regression model. Then, utterances with lower scores are filtered out. Finally, we retrain the TTS model with the filtered data.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

4.1.1. Dataset

We followed JTubesSpeech scripts [21] [32] to obtain dark data from YouTube; the amount was approximately 3,500 hours. Pre-screening with a CTC threshold of -0.3 and speaker compactness threshold of $[1, 7]^1$, we obtained approximately 66 hours (60,000 Japanese utterances) of 2,719 speakers as the pre-screened data. The sentences used for calculating the pseudo MOS were 100 phoneme-balanced sentences from the JVS corpus [9]. The test data used to evaluate the finally trained TTS models was 324 sentences from the ITA corpus [33]. There was no overlap in text among the pre-screened data, sentences for the pseudo MOS, and test data.

4.1.2. Model and training

We used FastSpeech 2 [34] for our multi-speaker TTS model and the pre-trained HiFi-GAN vocoder [35] UNIVERSAL_V1 [36]. We followed the model size and hyperparameters of the open-sourced implementation [37] except for the speaker representation. Instead of the one-hot speaker representation implemented in the repository, we used an open-sourced x -vector extractor [38], and a 512-dimensional x -vector was used to condition the TTS model. The x -vector was added to the output of the FastSpeech 2 encoder via a 512-by-256 linear layer. The x -vector was averaged for each speaker; one x -vector corresponded to one unique speaker. The TTS model was pre-trained using 10,000 utterances from the JVS corpus [9], the 100-speaker Japanese TTS corpus. We performed 300k steps with a batch size of 16 in this pre-training. TTS training in this paper started from this pre-trained model with 100k steps with a batch size of 16.

We used a pre-trained UTMOS [31] strong learner [39] to obtain a five-scale pseudo MOS on naturalness from synthetic speech. The regression model for predicting the pseudo MOS from the training data was 1-layer 256-unit bi-directional long short-term memory [40], followed by a linear layer, ReLU activation, and another linear layer. We used frame-level self-supervised learning (SSL) features² obtained with a wav2vec 2.0 model [42] [43], as the input. The frame-level outputs were aggregated to predict the pseudo MOS. The number of training steps, minibatch size, optimizer, and training objective were 10k, 12, Adam [44] with a learning rate of 0.0001, and mean squared error, respectively.

4.1.3. Compared methods

We compared the following data selection methods.

- **Unselected:** All the pre-screened data was used for the TTS training; the training data size was approximately 60,000 utterances.
- **Acoustic-quality (utterance-wise):** The training data was selected in terms of the acoustic quality of the data. We used NISQA [41], a recent deep learning-based model to predict scores on naturalness, noisiness, coloration, discontinuity, and loudness of the speech data. Each score takes $[1, 5]$, and we

¹These values are the same as the experiments in the JTubesSpeech paper [21].

²We compared the SSL features and the NISQA [41] features (used in the baseline) in the preliminary evaluation. The result demonstrated that the SSL features performed better.

set the threshold to 3.5, i.e., data for which all the scores were higher than 3.5 were selected. The TTS training data size is approximately 12,000.

- **Ours-Utt (evaluation-in-the-loop utterance-wise selection):** Our evaluation-in-the-loop data selection. For each data, we estimate the speech quality synthesized by the TTS model from an initial training with pre-screened data. The threshold for selecting training data was set to have the resulting training data be the same in size as “Acoustic-quality”.
- **Ours-Spk (evaluation-in-the-loop speaker-wise selection):** Our data selection, but the data selection was performed per speaker as described in Section 3.2.2. The threshold for selecting training data was set to have the resulting size of the training data be almost the same as “Acoustic-quality”.

4.1.4. Evaluation

We evaluated the selection methods to clarify the following:

- **Does our method obtain more “high-quality speakers?”: pseudo MOS comparison.** Our TTS model is expected to reproduce voices for a higher number of speakers. We define a “high-quality speaker” as a speaker with a higher pseudo MOS than the threshold. We in advance trained an high-quality multi-speaker TTS model using the JVS corpus [9] and calculated the speaker-wise pseudo MOS scores. We set the lowest score among the JVS speakers as the threshold³. Speakers with a higher score than the threshold were considered to be high-quality speakers. For each data selection method, we calculated 1) the distribution of the pseudo MOSs for synthetic speech by the trained TTS model and 2) the number of high-quality speakers.
- **Does our method work for unseen speaker?: pseudo MOS comparison.** The performance of the multi-speaker TTS model affects the synthetic speech quality of unseen speakers. Using the x -vector for unseen speakers, we counted the number of high-quality speakers among seen and unseen speakers, respectively.
- **Does our method increase speaker variation?** We evaluated whether our method obtains diverse (i.e., sounding different) high-quality speakers. To quantify the speaker variation, we calculated the cost of a Euclidean minimum spanning tree [45] of x -vectors of the high-quality speakers. The calculation is similar to the g2g (median of the distances to the nearest x -vector) [46], but we used summation instead of a representative value (i.e., median) because our purpose is to evaluate how widely speakers spread.
- **Does synthetic speech of so-called “high-quality speakers” truly sound natural?: actual MOS evaluation.** We evaluated whether our selection based on pseudo MOS is truly effective in synthesizing perceptually high-quality speech. We subjectively evaluated the synthetic speech quality for the data selection methods and the performance, including the relationship with pseudo MOS.

Note that seen and unseen speakers are different among the data selection methods. All the speakers were seen ones for “Unselected,” but only parts of them were seen in the other methods. Also, seen speakers were different among “Acoustic-quality,” “Ours-Utt,” and “Ours-Spk.” Speakers not used for training were considered to be unseen speakers for each method. Unless otherwise noted, we describe results aggregating those of both seen and unseen speakers.

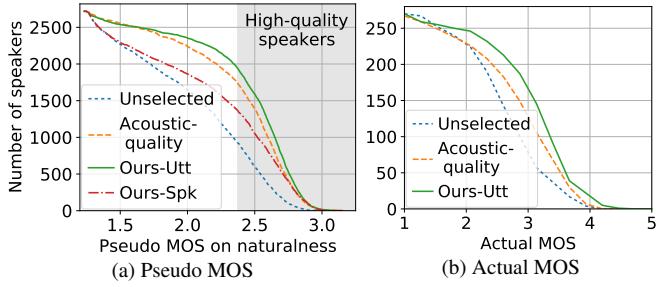


Fig. 4: Cumulative histograms of pseudo MOS and actual MOS. Y-axis value indicates number of speakers with higher score than x-axis value.

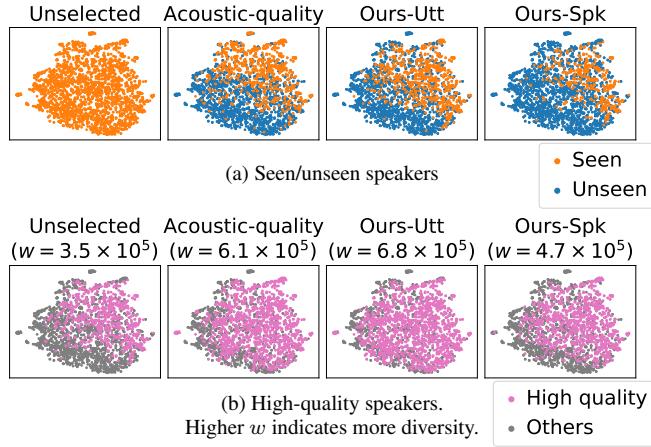


Fig. 5: Distributions of speakers by each data selection method.

4.2. Results

4.2.1. Number of high-quality speakers

Fig. 4a is a cumulative histogram of the pseudo MOSs. Our method had the highest values among the methods, demonstrating that a TTS model trained on our data selection can synthesize multi-speaker voices with higher quality than the other methods. The numbers of high-quality speakers for “Unselected,” “Acoustic-quality,” “Ours-Utt,” and “Ours-Spk” were 924, 1737, 1942, and 1367, respectively. We see that the proposed method increased the number of high-quality speakers, compared with the other methods. Specifically, the increment was approximately 1.2 times from that of “Acoustic-quality.” From these results, we can say that the proposed method work better than the conventional methods for the purpose of increasing the speaker variety of the multi-speaker TTS model. Note that this TTS corpus is much larger than the previous Japanese multi-TTS corpus (JVS) composed of 100 speakers.

In addition, comparing “Ours-Utt” and “Ours-Spk”, the proposed method was significantly better in terms of both the pseudo MOS distribution and the number of the high-quality speaker. This indicates that utterance-wise selection significantly contributes to enhancing the performance, rather than speaker-wise selection.

4.2.2. Performance of unseen speakers

Fig. 5a shows the distributions of the seen/unseen speakers to qualitatively compare the data selection methods. Compared with “Acoustic-quality,” “Ours-Utt” had a similar variation. Also, compared with “Ours-Spk,” “Ours-Utt” covered a wider range of speakers.

³The JVS corpus was constructed in a well-designed environment, and we confirmed that it was not an outlier.

Table 1: Number of seen and unseen speakers. Values of each cell are number of high-quality speakers, all speakers, and ratio of two values, respectively.

Method	Seen	Unseen
Unselected	924/2719(34.0%)	-
Acoustic-quality	731/912(80.2%)	1006/1807(55.7%)
Ours-Utt	811/882(92.0%)	1131/1837(61.6%)
Ours-Spk	468/505(92.7%)	899/2214(40.6%)

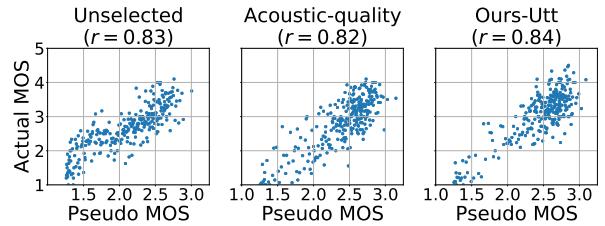


Fig. 6: Pseudo MOS vs. actual MOS. Each point is each speaker.

Table 1 lists the number of seen and unseen speakers for each data selection method. “Ours-Utt” outperformed “Acoustic-quality” in terms of the percentage and absolute value for both seen and unseen speakers; our method worked even for unseen speakers. “Ours-Spk” was good at the percentage of seen speakers but not at the others. We expect that the small size of high-quality seen speakers in “Ours-Spk” caused this result.

4.2.3. Evaluation of speaker variation

Fig. 5b shows the distributions of the high-quality speakers and speaker variation scores w . Both qualitatively and quantitatively, “Ours-Utt” increased the speaker variation compared with the other methods, indicating that our method contributes to speaker variation.

4.2.4. Comparison of pseudo MOS and actual MOS

We conducted a five-point MOS test on the naturalness of synthetic speech. 500 listeners participated, and each listener listened to 24 samples. The TTS models except for “Ours-Spk” data selection were used for synthesizing speech. To reduce the evaluation cost, we sampled speakers to be evaluated rather than using all the 2,719 speakers. For each data selection method, we divided the pseudo MOS in Fig. 4a into 272 intervals and randomly selected one speaker from each interval. Therefore, 272 speakers (10% of 2,719 speakers) were prepared for each method. The actual MOSs were aggregated for each speaker.

Fig. 4b shows the result. The proposed method had the highest values among all the methods. This result indicates that the proposed method is more effective than the conventional methods even in perceptual speech quality.

To further analyze these results, we investigated the relationship between the pseudo MOS and actual MOS. Fig. 6 shows scatter plots and correlation coefficients r . These results show that the correlation was always high ($r > 0.8$) despite our target language (Japanese) not being included in the training data (English and Chinese) of the pseudo MOS prediction model. This indicates that pseudo MOS is valid for languages not included in pseudo MOS training data, indicating a good capability for languages.

5. CONCLUSION

We proposed an evaluation-in-the-loop data selection method for TTS from dark data. Experimental results using YouTube videos showed that our method significantly outperformed the conventional acoustic-quality-based method. Our future work includes the use of more recent TTS models.

References

- [1] J. Shen, R. Pang, R. J. Weiss, *et al.*, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018, pp. 4779–4783.
- [2] N. Li, S. Liu, Y. Liu, *et al.*, “Neural speech synthesis with Transformer network,” in *Proc. AAAI*, vol. 33, 2019, pp. 6706–6713.
- [3] X. Tan, J. Chen, H. Liu, *et al.*, “NaturalSpeech: End-to-end text to speech synthesis with human-level quality,” *arXiv:2205.04421*, 2022.
- [4] W. Ping, K. Peng, A. Gibiansky, *et al.*, “Deep Voice 3: 2000-speaker neural text-to-speech.,” in *ICLR*, vol. 79, 2018, pp. 1094–1099.
- [5] Y. Jia, Y. Zhang, R. Weiss, *et al.*, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Proc. NIPS*, vol. 31, 2018.
- [6] E. Cooper, C.-I. Lai, Y. Yasuda, *et al.*, “Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings,” in *Proc. ICASSP*, 2020, pp. 6184–6188.
- [7] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, “Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” 2016.
- [8] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis,” *arXiv:1711.00354*, 2017.
- [9] S. Takamichi, K. Mitsui, Y. Saito, *et al.*, “JVS corpus: Free Japanese multi-speaker voice corpus,” *arXiv:1908.06248*, 2019.
- [10] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [11] D. Snyder, D. Garcia-Romero, G. Sell, *et al.*, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proc. ICASSP*, 2018, pp. 5329–5333.
- [12] D. Trajanov, V. Zdravesci, R. Stojanov, *et al.*, “Dark data in Internet of things (IoT): Challenges and opportunities,” in *7th Small Systems Simulation Symposium*, 2018, pp. 1–8.
- [13] B. Schembera and J. M. Durán, “Dark data as the new challenge for big data science and the introduction of the scientific data officer,” *Philosophy & Technology*, vol. 33, no. 1, pp. 93–115, 2020.
- [14] H. Zen, V. Dang, R. Clark, *et al.*, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in *Proc. INTERSPEECH*, 2019, pp. 1526–1530.
- [15] E. Bakhturina, V. Lavrukhin, B. Ginsburg, *et al.*, “Hi-Fi multi-speaker English TTS dataset,” in *Proc. INTERSPEECH*, 2021, pp. 2776–2780.
- [16] P. Puchtler, J. Wirth, and R. Peinl, “HUI-Audio-Corpus-German: A high quality TTS dataset,” in *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, 2021, pp. 204–216.
- [17] A. Rousseau, P. Deléglise, and Y. Esteve, “TED-LIUM: An automatic speech recognition dedicated corpus.,” in *Proc. LREC*, 2012, pp. 125–129.
- [18] C. Zhang, Y. Ren, X. Tan, *et al.*, “DenoiSpeech: Denoising text to speech with frame-level noise modeling,” in *Proc. ICASSP*, 2021, pp. 7063–7067.
- [19] T. Saeki, K. Tachibana, and R. Yamamoto, “DRSpeech: Degradation-robust text-to-speech synthesis with frame-level and utterance-level acoustic representation learning,” in *Proc. INTERSPEECH*, 2022, pp. 793–797.
- [20] K. Nikitaras, G. Vamvoukakis, N. Ellinas, *et al.*, “Fine-grained noise control for multispeaker speech synthesis,” in *Proc. INTERSPEECH*, 2022, pp. 828–832.
- [21] S. Takamichi, L. Kürzinger, T. Saeki, *et al.*, “JTubeSpeech: Corpus of Japanese speech collected from YouTube for speech recognition and speaker verification,” *arXiv:2112.09323*, 2021.
- [22] D. Galvez, G. Diamos, J. M. C. Torres, *et al.*, “The People’s Speech: A large-scale diverse English speech recognition dataset for commercial usage,” in *Proc. NeurIPS Datasets and Benchmarks*, <https://openreview.net/forum?id=R8CwidgJ0yT>, 2021.
- [23] G. Chen, S. Chai, G. Wang, *et al.*, “GigaSpeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio,” *arXiv:2106.06909*, 2021.
- [24] S. Möller, F. Hinterleitner, T. H. Falk, *et al.*, “Comparison of approaches for instrumentally predicting the quality of text-to-speech systems,” in *Proc. INTERSPEECH*, 2010, pp. 1325–1328.
- [25] B. Patton, Y. Agiomrygiannakis, M. Terry, *et al.*, “AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech,” in *Proc. NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*, 2016.
- [26] W. C. Huang, E. Cooper, Y. Tsao, *et al.*, “The VoiceMOS Challenge 2022,” in *Proc. INTERSPEECH*, 2022, pp. 4536–4540.
- [27] L. Kürzinger, D. Winkelbauer, L. Li, *et al.*, “CTC-segmentation of large corpora for German end-to-end speech recognition,” in *Proc. SPECOM*, 2020, pp. 267–278.
- [28] J. Williams and S. King, “Disentangling style factors from speaker representations.,” in *Proc. INTERSPEECH*, 2019, pp. 3945–3949.
- [29] T. Hayashi, R. Yamamoto, T. Yoshimura, *et al.*, “ESPnet-TTS: Extending the edge of TTS research,” *arXiv:2110.07840*, 2022.
- [30] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, *et al.*, “Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis,” in *Proc. ICASSP*, 2021, pp. 5679–5683.
- [31] T. Saeki, D. Xin, W. Nakata, *et al.*, “UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022,” in *Proc. INTERSPEECH*, 2022, pp. 4521–4525.
- [32] *JTubeSpeech: Corpus of Japanese speech collected from YouTube*, <https://github.com/sarulab-speech/jtubespeech>.
- [33] *ITA corpus*, <https://github.com/mmorisie/ita-corpus>.
- [34] Y. Ren, C. Hu, X. Tan, *et al.*, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *Proc. ICLR*, 2021.
- [35] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Proc. NeurIPS*, vol. 33, pp. 17 022–17 033, 2020.
- [36] *HiFi-GAN*, <https://github.com/jik876/hifi-gan>.
- [37] *FastSpeech2 JSUT implementation*, <https://github.com/Wataru-Nakata/FastSpeech2-JSUT>.
- [38] *x-vector extractor for Japanese speech*, https://github.com/sarulab-speech/xvector_jtubespeech.
- [39] *UTMOS: UTokyo-SaruLab MOS Prediction System*, <https://github.com/sarulab-speech/UTMOS22>.
- [40] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [41] G. Mittag, B. Naderi, A. Chehadi, *et al.*, “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Proc. INTERSPEECH*, 2021, pp. 2127–2131.
- [42] A. Baevski, Y. Zhou, A. Mohamed, *et al.*, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.
- [43] *wav2vec 2.0*, <https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec>.
- [44] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *Proc. ICLR*, 2015.
- [45] W. B. March, P. Ram, and A. G. Gray, “Fast Euclidean minimum spanning tree: Algorithm, analysis, and applications,” in *Proc. ACM SIGKDD*, 2010, pp. 603–612.
- [46] D. Stanton, M. Shannon, S. Mariooryad, *et al.*, “Speaker generation,” in *Proc. ICASSP*, 2022, pp. 7897–7901.