Can we use Common Voice to train a Multi-Speaker TTS system?

Sewade Ogun, Vincent Colotte, Emmanuel Vincent

CAN WE USE COMMON VOICE TO TRAIN A MULTI-SPEAKER TTS SYSTEM?

Sewade Ogun, Vincent Colotte, Emmanuel Vincent

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sewade.ogun@inria.fr

ABSTRACT

Training of multi-speaker text-to-speech (TTS) systems relies on curated datasets based on high-quality recordings or audiobooks. Such datasets often lack speaker diversity and are expensive to collect. As an alternative, recent studies have leveraged the availability of large, crowdsourced automatic speech recognition (ASR) datasets. A major problem with such datasets is the presence of noisy and/or distorted samples, which degrade TTS quality. In this paper, we propose to automatically select high-quality training samples using a non-intrusive mean opinion score (MOS) estimator, WV-MOS. We show the viability of this approach for training a multi-speaker GlowTTS model on the Common Voice English dataset. Our approach improves the overall quality of generated utterances by 1.26 MOS point with respect to training on all the samples and by 0.35 MOS point with respect to training on the LibriTTS dataset. This opens the door to automatic TTS dataset curation for a wider range of languages.

***Index Terms*—** Multi-speaker text-to-speech, Common Voice, crowdsourced corpus, non-intrusive quality estimation

1. INTRODUCTION

Research on text-to-speech (TTS) is increasingly focusing on multi-speaker TTS as it is more challenging and often requires explicit modelling of speaker characteristics. This interest has helped improve the performance of multi-speaker TTS in terms of prosody [1], expressiveness [2], new speaker generation [3], zero-shot training [4], and synthetic data generation for downstream tasks like automatic speech recognition (ASR) [5], among others. Depending on the application, different characteristics of speech need to be modeled. For example, for synthetic ASR training data generation, it is necessary for the model to have seen diverse speakers and accents.

Datasets currently used for TTS system training fall into two categories, namely studio-quality TTS datasets such as VCTK [6], and TTS datasets curated from audiobooks such as LibriTTS [7]. However, the VCTK dataset includes only 110 speakers, and LibriTTS has a concentration of US English accents, which are not representative of the entire spectrum of speakers and accents. On top of that, the collection of datasets such as VCTK may be too expensive for some languages.

Large, crowdsourced ASR datasets are good candidates for driving TTS research in future directions, as they inherently exhibit the larger speaker variability (in terms of accent, speaking style, speaking rate, etc.) required for TTS systems to model diverse speakers. However, problems such as noise, low bandwidth, mispronunciation, variation in recording conditions, etc., hinder their usability for TTS training.

In this paper, we focus on automatically selecting high-quality training samples from a crowdsourced dataset, using Common Voice English [8] as an example. In this context, quality cannot be estimated via subjective listening tests, that are intractable with 1.4 M utterances, or objective metrics like PESQ [9] that require a reference signal. Instead, we leverage the increasing accuracy of deep learning based, non-intrusive quality estimators. Specifically, we use a self-supervised model fine-tuned for mean opinion score (MOS) estimation, WV-MOS [10], and select the speakers whose average WV-MOS score across all utterances is above a threshold. We evaluate the intelligibility, audio quality and speaker similarity of the utterances generated by a multi-speaker GlowTTS model trained on the resulting dataset, and also briefly explore the other factors not captured by WV-MOS.

Section 2 describes related works on TTS dataset creation, TTS training on noisy speech, and MOS estimation. Section 3 describes the Common Voice dataset, its properties and limitations. Sections 4 and 5 describe our method and the experiments performed to validate it. We conclude in Section 6.

2. RELATED WORK

Several multi-speaker datasets have been collected in recent years for TTS applications [6, 7]. To create these corpora, researchers either record utterances in semi-anechoic chambers for good signal quality, or utilise various methods to select utterances from audiobooks, as this is less cumbersome. For example, the LibriTTS dataset was derived from the popular LibriSpeech ASR dataset [11] by trimming silences and filtering out utterances with low estimated signal-to-noise ratio (SNR). Although this filtering step is not perfect and allows a few noisy samples to remain uncaught, the resulting dataset is believed to be good enough for TTS since the original LibriSpeech is higher-quality than Common Voice on average.

A few works have used other metrics such as the word er-

ror rate or the Mel cepstral distortion to automatically select good training utterances from noisy speech datasets [12–14], however they have only been demonstrated on small datasets so far. Research on multi-speaker TTS training using noisy speech has also focused on directly modelling the noise in order to factor it out during inference [15, 16], and on encoding all the environmental characteristics of speech for novel speech generation in different conditions [17].

Recently, several methods have been proposed to automatically measure the quality of speech utterances [18,19] but they do not always generalise well outside of the training corpus. Self-supervised pretrained models followed by a shallow MOS regression head result in higher correlation with human evaluators' scores [20] than previous architectures. They also generalise better to unseen speakers and utterances, and can be used to evaluate the performance of speech processing systems for a variety of tasks [10].

3. THE COMMON VOICE DATASET

Common Voice [8] is a crowdsourced, Creative Commons Zero licensed, read speech dataset currently available in over 93 languages. It contains recordings from volunteers who read a text transcript sourced from public domain text. Each utterance is up-voted or down-voted by volunteers according to a list of criteria.[1] These criteria are not very restrictive, e.g., various kinds of background noises are allowed. Utterances with more than two up-votes are marked as validated. The validated utterances are then split into train, development and test sets, with non-overlapping speakers and sentences.

3.1. Analysing Common Voice Dataset Quality for TTS

Although the validated set has been widely exploited for ASR [21], we observe some undesirable properties for TTS:

- Noise: Speech quality may be degraded by electromagnetic noise or acoustic noise such as mouse clicks, low frequency noise, background speakers and background music, among others. Since the utterances are stored as mp3, quantization noise can also sometimes be heard.
- Low bandwidth: Due to recording choices or high compression, some audio files are low-pass filtered, with a cutoff frequency that varies from one file to another.
- Mispronunciation: We observe mispronunciations of "unfamiliar" words, variations in the pronunciation of certain other words, and some utterances in other languages (e.g., German utterances in the English corpus).
- Unavailable speaker metadata: Age, gender and accent information are not available for all speakers, while some TTS systems require this information as input.
- Other factors include variable recording characteristics (microphone, room, recording device), speaking rate, and volume. These recording characteristics must be ignored by models, and the speaking rate and volume, while being inherent characteristics of the speaker, can enlarge the space of variables to be considered.

These characteristics are generally not a hindrance for ASR training, and they can even be desirable for robustness. However this is not the case for TTS training [7, 12].

3.2. Dataset Preparation

In the following, we use the English subset of Common Voice (version 7.0). We exclude the predefined development and test sets, and utterances longer than 16.7 s to allow large batch sizes. We consider all other utterances in the 2015 h validated set as candidate TTS training samples. The samples are preprocessed by resampling from 32 or 48 kHz to 16 kHz, and removing beginning and end silences using pydub[2] with a threshold of -50 dBFS. The range of speaker duration is also limited to between 20 min and 10 h by randomly selecting a 10 h subset of utterances for speakers with longer duration and discarding speakers with less than 20 min total duration.

Furthermore, the training utterances are denoised using the pretrained DPTNet model of Asteroid [22]. We run separate experiments for the original and denoised utterances to evaluate the impact of denoising on the resulting TTS model.

4. METHODOLOGY

We filter the dataset by only selecting speakers with high automatically estimated MOS scores. We believe that utterances from these speakers are of high quality and devoid of noise and missing frequency bands. To ascertain this, we train different TTS models on the same dataset filtered at different estimated MOS thresholds.

4.1. MOS Estimation

MOS estimation is performed using WV-MOS [10], a pretrained MOS estimation model.[3] The model combines a pretrained wav2vec2.0 feature extractor and a 2-layer multi-layer perceptron (MLP) head, which are jointly fine-tuned on the subjective evaluation scores of the Voice Conversion Challenge 2018 using a mean squared error loss. It was shown to correlate well with human quality judgment regarding noise and low bandwidth [10, App. C].

Every speaker is assigned a single, speaker-level WV-MOS score by averaging the estimated utterance-level scores. We assume that recording and environmental conditions for each speaker remain relatively constant.

We select all utterances from those speakers whose speaker-level WV-MOS score is above a threshold of 4.0, 3.8, 3.5, 3.0, or 2.0, and compare the resulting TTS systems

[1] https://commonvoice.mozilla.org/en/criteria

[2] https://github.com/jiaaro/pydub

[3] https://github.com/AndreevP/WV-MOS

with a baseline trained on all available data.[4] Table 1 shows the training data duration and number of speakers corresponding to each WV-MOS threshold. The thresholds were selected so as to balance data size and number of speakers.

Table 1: Training data duration and number of speakers for various selected WV-MOS thresholds.

WV-MOS threshold	Duration (h)	Number of speakers
Baseline	636.27	633
WV-MOS ≥ 2.0	620.14	623
WV-MOS ≥ 3.0	532.14	537
WV-MOS ≥ 3.5	310.40	337
WV-MOS ≥ 3.8	187.40	183
WV-MOS ≥ 4.0	86.05	88

4.2. TTS Model

We evaluate the dataset quality for TTS training at each WV-MOS threshold by training a multi-speaker GlowTTS model [23], conditioned on an external speaker embedding, similar to [4]. This model uses a Transformer encoder and a flow-based decoder, along with a phoneme duration prediction network. The model choice was influenced by its quality and its relatively short training time compared to other TTS models. Each utterance's speaker embedding and the corresponding sentence converted into phonemes are used as inputs to the model during training. The output is a Mel-spectrogram.

For each utterance, we pre-compute a speaker embedding from a speaker verification model[5] trained on Voxceleb. The embeddings are l2-normalised, 256-dimensional vectors.

Lastly, a 16 kHz HiFi-GAN $V1$ vocoder [24] was trained on the LibriTTS dataset to convert the generated Mel-spectrograms into audio signals. The vocoder was fixed and used to evaluate all the TTS systems considered, as the relative trends were true for vocoders trained on different datasets in our preliminary experiments. As such, the trained vocoder was not finetuned on Mel-spectrograms generated by GlowTTS so as to objectively evaluate the generated Mel-spectrograms. All experiments were carried out using the NeMo toolkit [25].

5. EXPERIMENTAL EVALUATION

5.1. Training Hyper-Parameters

All TTS models are trained using a global batch size of 128 or 256 on 4 GPUs (the noisier datasets only learn when the batch size is large). The model is optimised using the RAdam optimizer, a learning rate of 0.001, and a cosine-annealing scheduler with linear warm up steps of 6,000. Each model is trained until the validation loss stops decreasing for more than 10 epochs. We select 504 utterances randomly from each dataset for validation. At inference, the generation is done with a noise scale of 0.667 and length scale of 1.0, which are the best multi-speaker inference parameters reported in [23].

5.2. Objective Evaluation

To evaluate the TTS systems objectively, we generate utterances for speakers seen and speakers unseen at training time. 80 speakers are randomly selected from the smallest subset (WV-MOS ≥ 4.0) of Common Voice[6] and from the VCTK corpus to represent seen and unseen speakers, respectively. For each speaker, a single speaker embedding is extracted from a reference utterance of that speaker, that is either a randomly selected utterance with duration longer than 2 s for seen speakers or the fifth utterance (SpeakerID_005) as in [4] for unseen speakers. The embeddings are used to generate 25 utterances for each speaker, using text sentences from the VCTK corpus with more than 20 words. This results in a total of $2,000$ test utterances for both seen and unseen speakers.

To measure the audio quality, speaker similarity, and intelligibility of the generated utterances, we compute the average WV-MOS score **(WV-MOS)**, the cosine similarity between the speaker embeddings of the generated and the reference utterances **(cos-sim)**, and the character error rate **(CER)**, respectively. The pretrained QuartzNet15x5 model from the NeMo toolkit was used to compute the CER.

5.3. Subjective Evaluation

We also evaluate the TTS systems subjectively in terms of overall quality **(MOS)** (whether audio sounds natural, non-robotic, and noiseless), speaker similarity between the generated and reference utterances **(S-MOS)**, and intelligibility. In total, 24 volunteers participated in the evaluation.

For MOS and S-MOS, we selected 2 unseen male speakers (p245 and p254) and 2 unseen female speakers (p231 and p250) from the VCTK corpus, and generated 5 utterances per speaker for all TTS models. Volunteers were asked to listen to the utterances carefully and give an MOS score for each utterance on a 1–5 scale, then an S-MOS score on a 1–5 scale with respect to a VCTK reference utterance. They were asked to ignore the speaking rate in their scoring, as the VCTK reference speaker often speaks faster than the generated utterances.

To measure intelligibility, we apply the minimal pair approach [26]. A minimal pair is a pair of words that vary by only one phoneme, e.g., sea/she.[7]. This approach helps identify deficiencies in phoneme generation by the TTS systems, and it is widely used by phoneticians. In the evaluation, a carrier audio containing one of the pairs is presented to the evaluator and the evaluator has to select the word heard among

[4] In the case of denoised training samples, the WV-MOS score is computed before denoising so that the list of selected speakers is not affected.

[5] `https://github.com/resemble-ai/Resemblyzer`

[6] By design, all speakers in that subset are also included in the lower WV-MOS threshold training subsets.

[7] We select minimal pairs from `https://www.englishclub.com/pronunciation/minimal-pairs.htm`

three options: the correct word (e.g., sea), its minimal pair (e.g., she) or none of these. The proportion of correctly associated words is then computed as the **intelligibility score**. 25 minimal pairs were evaluated for each model considered.

5.4. Results

5.4.1. Impact of the WV-MOS Threshold

Figure 1 shows the objective evaluation plots. The WV-MOS scores in Fig. 1a follow an increasing trend from the baseline to the WV-MOS ≥ 4.0 dataset. Subjective MOS results in Table 2 also corroborate this. This indicates that filtering based on WV-MOS scores improves the quality as expected.

In Fig. 1b, for models trained on original data, the cos-sim for seen speakers shows an increasing trend, while the cos-sim for unseen speakers is lowest at the two ends of the plot. We see a similar trend in Table 2 for the S-MOS of unseen speakers. This indicates that more training speakers are key for modelling speaker variability.

In Fig. 1c, the CER decreases steadily from the noisy baseline dataset to the less noisy WV-MOS datasets for both seen and unseen speakers. We see a reduction of more than 50 % in the CER from the baseline to the WV-MOS ≥ 4.0 dataset. This indicates that intelligibility is mostly affected by the quality of the dataset, not the size.

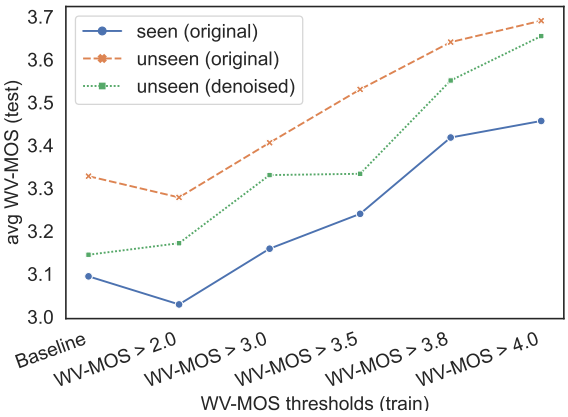
The lowest quality datasets (Baseline vs. WV-MOS $\geq$ 2.0) do not follow these trends due to the higher variance in utterance-level WV-MOS scores for lower-quality speakers.

Table 2: Subjective / objective quality (MOS / WV-MOS) and speaker similarity (S-MOS / cos-sim) of utterances generated for 4 unseen speakers by TTS models trained on the baseline dataset and the WV-MOS ≥ 3.0 and WV-MOS ≥ 4.0 datasets. Bold numbers denote the best system in each row and the systems statistically equivalent to it.

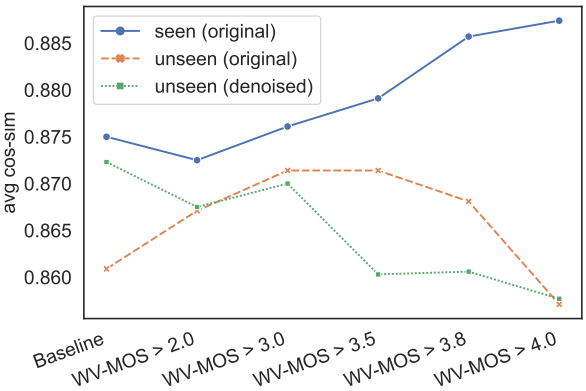
	Baseline	WV-MOS ≥ 3.0	WV-MOS ≥ 4.0
MOS	2.35	3.12	**3.69**
WV-MOS	3.63	3.59	**4.09**
S-MOS	2.69	**2.90**	**2.79**
cos-sim	**0.831**	**0.845**	**0.832**

5.4.2. Impact of Denoising Training Utterances

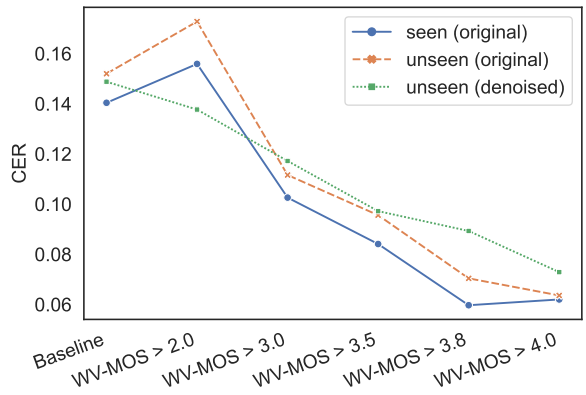
As seen in Figs. 1a and 1c, the WV-MOS scores and CER for unseen speakers follow the same trend, whether the training data are denoised or not. Furthermore, denoising degrades the WV-MOS score and the CER, except for low quality datasets in the case of the CER. In Fig. 1b, we see that denoising also degrades the cos-sim score, except for the baseline. This shows that some speaker information may be lost during denoising. We conclude that denoising is not beneficial and automatically selecting high-quality samples is the best strategy.



(a) WV-MOS scores for seen and unseen speakers.



(b) Cosine similarity between speaker embeddings.



(c) CER for seen and unseen speakers.

Fig. 1: Objective quality (WV-MOS), speaker similarity (cos-sim) and intelligibility (CER) of utterances generated for 80 seen and 80 unseen speakers by TTS models trained on original or denoised samples above a given WV-MOS threshold.

5.4.3. Common Voice vs. LibriTTS

We compare our dataset curation method to a standard TTS dataset: LibriTTS. Since LibriTTS has more speakers (2,484) and more data (492.68 h after discarding utterances longer

Table 3: Subjective quality (MOS), speaker similarity (S-MOS) and intelligibility of utterances generated for 4 unseen speakers by TTS models trained on the baseline dataset, LibriTTS, and the WV-MOS ≥ 4.0-all dataset. Corresponding objective scores (WV-MOS and cos-sim) are included. Bold numbers denote the best system in each row and the systems statistically equivalent to it. Copy-synthesis on VCTK speech (VCTK-copy) provides an upper bound on the achievable speaker similarity.

		Baseline	LibriTTS	WV-MOS ≥ 4.0-all	VCTK-copy
MOS	Male	2.44	3.15	**3.70**	-
	Female	2.26	**3.38**	**3.52**	-
	Total	2.35	3.26	**3.61**	-
WV-MOS		3.63	**3.75**	**3.80**	-
S-MOS	Male	**2.92**	**2.95**	**3.02**	4.50
	Female	2.46	2.53	**2.73**	4.73
	Total	2.69	2.74	**2.88**	4.61
cos-sim		0.831	**0.861**	**0.861**	0.869
Intelligibility score		0.72	**0.82**	**0.82**	-

than 16.7 s) than the WV-MOS ≥ 4.0 dataset, we select all speakers with WV-MOS above 4.0, without setting a 20 min lower bound on total speaker duration. We call the resulting 4,469-speaker, 230.75 h dataset WV-MOS ≥ 4.0-all. Table 3 evaluates the utterances generated by the TTS models trained on this dataset vs. LibriTTS for unseen speakers.

Training on WV-MOS ≥ 4.0-all results in a similar intelligibility score to training on LibriTTS.

The S-MOS score is highest when training on WV-MOS ≥ 4.0-all, however we still see a large gap in S-MOS between this model and the VCTK-copy topline, which leaves room for further improvement in speaker modelling. We notice that male S-MOS scores are higher than female S-MOS scores for both datasets, which indicates that male speakers are better modelled by models trained on either dataset.

Finally, while the WV-MOS scores for utterances generated by training on WV-MOS ≥ 4.0-all vs. training on LibriTTS are not statistically different, volunteers consistently gave higher MOS scores to the former, with an average improvement of 0.35 MOS point. Male speakers are assigned higher MOS scores, unlike LibriTTS where female voices are given higher MOS scores (in line with [7]).

5.4.4. Other factors not captured by WV-MOS

We performed experiments on a medium quality dataset, WV-MOS ≥ 3.5, to assess the impact of other factors. First, we removed sentences in a foreign language by using a language identification tool, LangID [8], to filter out sentences with English language probability lower than 0.8. Second, using the same pretrained ASR model as above, we dropped utterances with a CER above 0.4 for better alignment of training text to utterances. Third, we filtered utterances according to their WV-MOS at the utterance level, instead of the speaker level. Finally, we removed pauses longer than 180 ms inside utterances using a voice activity detector.[9] Each of these experiments resulted in discarding less than 1.5 % of the initial dataset. Informal listening tests did not show any improvement in quality and intelligibility. Although this would have required more formal tests to validate, we conclude that the other factors not captured by WV-MOS are not critical.

6. CONCLUSION

In this paper, we successfully improved the overall quality, speaker similarity, and intelligibility of utterances generated by a multi-speaker TTS model trained on the Common Voice English dataset. This was achieved by selecting high-quality training samples using a non-intrusive MOS estimator. Furthermore, we showed that denoising reduces the CER and increases the speaker similarity score (cos-sim) of generated utterances when the dataset is noisy, but degrades performance otherwise. The resulting automatically curated dataset shows promise for future TTS experiments, as it outperforms LibriTTS in terms of both subjective quality and speaker similarity. The applied approach is generic and could enable the creation of TTS training datasets for languages for which manual curation is not financially viable. In future work, we will report the impact of vocoder training data quality on the absolute performance of the system.

7. ACKNOWLEDGMENTS

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

[8] `https://github.com/saffsd/langid.py`

[9] `https://github.com/wiseman/py-webrtcvad`

8. REFERENCES

[1] G. Sun, Y. Zhang, Weiss, et al., "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *ICASSP*, 2020, pp. 6264–6268. 1

[2] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP*, 2020, pp. 6189–6193. 1

[3] D. Stanton, M. Shannon, S. Mariooryad, et al., "Speaker generation," in *ICASSP*, 2022, pp. 7897–7901. 1

[4] E. Casanova, C. Shulby, E. Gölge, et al., "SC-GlowTTS: An efficient zero-shot multi-speaker text-to-speech model," in *Interspeech*, 2021, pp. 3645–3649. 1, 3

[5] T.-Y. Hu, M. Armandpour, A. Shrivastava, et al., "SYNT++: Utilizing imperfect synthetic data to improve speech recognition," in *ICASSP*, 2022, pp. 7682–7686. 1

[6] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019. 1

[7] H. Zen, V. Dang, R. Clark, et al., "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Interspeech*, 2019, pp. 1526–1530. 1, 2, 5

[8] R. Ardila, M. Branson, K. Davis, et al., "Common Voice: A massively-multilingual speech corpus," in *LREC*, 2020, pp. 4211–4215. 1, 2

[9] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, vol. 2, pp. 749–752. 1

[10] P. Andreev, A. Alanov, O. Ivanov, et al., "HiFi++: A unified framework for neural vocoding, bandwidth extension and speech enhancement," *arXiv:2203.13086*, 2022. 1, 2

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *ICASPP*, 2015, pp. 5206–5210. 1

[12] P. Baljekar and A. W. Black, "Utterance selection techniques for TTS systems using found speech," in *ISCA Speech Synthesis Workshop*, 2016, pp. 184–189. 2

[13] F.-Y. Kuo, S. Aryal, G. Degottex, et al., "Data selection for improving naturalness of TTS voices trained on small found corpuses," in *SLT*, 2018, pp. 319–324. 2

[14] E. Cooper, X. Wang, A. Chang, et al., "Utterance selection for optimizing intelligibility of TTS voices trained on ASR data," in *Interspeech*, 2017, pp. 3971–3975. 2

[15] C. Zhang, Y. Ren, X. Tan, et al., "Denoispeech: Denoising text to speech with frame-level noise modeling," in *ICASSP*, 2021, pp. 7063–7067. 2

[16] W.-N. Hsu, Y. Zhang, R. J. Weiss, et al., "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *ICASSP*, 2019, pp. 5901–5905. 2

[17] J.-H. R. Chang, A. Shrivastava, H. S. Koppula, et al., "Style Equalization: Unsupervised learning of controllable generative sequence models," *arXiv:2110.02891*, 2021. 2

[18] C.-C. Lo, S.-W. Fu, W.-C. Huang, et al., "MOSNet: Deep learning-based objective assessment for voice conversion," in *Interspeech*, 2019, pp. 1541–1545. 2

[19] B. Patton, Y. Agiomyrgiannakis, M. Terry, et al., "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," in *NIPS workshop*, 2016. 2

[20] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *ICASSP*, 2022, pp. 8442–8446. 2

[21] M. Riviere, A. Joulin, P.-E. Mazaré, et al., "Unsupervised pretraining transfers well across languages," in *ICASSP*, 2020, pp. 7414–7418. 2

[22] M. Pariente, S. Cornell, J. Cosentino, et al., "Asteroid: The pytorch-based audio source separation toolkit for researchers," in *Interspeech*, 2020, pp. 2637–2641. 2

[23] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *NIPS*, vol. 33, pp. 8067–8077, 2020. 3

[24] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *NIPS*, vol. 33, pp. 17022–17033, 2020. 3

[25] O. Kuchaiev, J. Li, H. Nguyen, et al., "NeMo: A toolkit for building AI applications using neural modules," *arXiv:1909.09577*, 2019. 3

[26] M. M. Hodge and C. L. Gotzke, "Minimal pair distinctions and intelligibility in preschool children with and without speech sound disorders," *Clinical linguistics & phonetics*, vol. 25, no. 10, pp. 853–863, 2011. 3