INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, Yonghui Wu

Google AI
heigazen@google.com

Abstract

This paper introduces a new speech corpus called "LibriTTS" designed for text-to-speech use. It is derived from the original audio and text materials of the LibriSpeech corpus, which has been used for training and evaluating automatic speech recognition systems. The new corpus inherits desired properties of the LibriSpeech corpus while addressing a number of issues which make LibriSpeech less than ideal for text-to-speech work. The released corpus consists of 585 hours of speech data at 24kHz sampling rate from 2,456 speakers and the corresponding texts. Experimental results show that neural end-to-end TTS models trained from the LibriTTS corpus achieved above 4.0 in mean opinion scores in naturalness in five out of six evaluation speakers. The corpus is freely available for download from http://www.openslr.org/60/.

Index Terms: text-to-speech; neural network; corpus;

1. Introduction

The introduction of deep learning-based, neural end-to-end approaches has lowered the barrier to develop high-quality text-to-speech (TTS) systems [1–4]. With a sufficient amount of studio-quality recorded speech data from a single professional speaker, one can train a generative neural end-to-end TTS model capable of synthesizing speech in a reading style almost as natural as the training speaker [4,5]. Thus, the focus of TTS research is shifting toward more challenging tasks, such as multi-speaker TTS [6–8], building TTS systems from small amounts of data [9], few shot voice adaptation [8, 10, 11], unsupervised prosody and speaking style modeling [12, 13], and the use of found data [13, 14]. Having appropriate data is essential to explore these new tasks.

The LibriSpeech corpus [15] is derived from audiobooks that are part of the LibriVox project [16]. There are 982 hours of speech data from 2,484 speakers in this corpus. It is designed to be reasonably balanced in terms of gender and per-speaker duration. Furthermore, as it is released under a non-restrictive license, it can be used for both non-commercial and commercial purposes. Although this corpus was originally designed for automatic speech recognition (ASR) research, it has been used in various text-to-speech (TTS) research projects [7, 8, 11] thanks to its attractive properties, such as a non-restrictive license, a large amount of data, and wide speaker diversity.

However, it also has a number of undesired properties when considering its use for TTS . The properties which are addressed in this paper are as follows:

- *The audio files are at 16 kHz sampling rate*; 16 kHz sampling is high enough for the ASR purpose but too low to achieve high quality TTS. Modern production-quality TTS systems often use 24, 32, 44.1, or 48 kHz sampling rate [17, 18].
- *The speech is split at silence intervals*; the training data speech is split at silences longer than 0.3 seconds. To learn long-term characteristics of speech such as the sentence-level prosody for given a text, it is necessary to split speech only at sentence breaks.
- *All letters are normalized into uppercase, and all punctuation is removed*; capitalization and punctuation marks are useful features to learn prosodic characteristics such as emphasis and the length of pauses.
- *The position of sentences within paragraphs is discarded*; to learn inter-sentence prosody it is desirable to access neighbouring sentence text or audio, but this information is missing.
- *Some audio files contain significant background noise even within its "clean" subsets*; in the LibriSpeech corpus, speakers with low word error rates (WERs) using the Wall Street Journal (WSJ) acoustic model were designated as "clean". Therefore, the "clean" subset can contain noisy samples.

To address these undesired properties while keeping the desired properties (unrestricted license, large speaker inventories, and gender balance) of the LibriSpeech corpus as much as possible, this paper introduces a new corpus called "LibriTTS". The LibriTTS corpus is derived from the original materials (MP3 from LibriVox and texts from Project Gutenberg) of the LibriSpeech corpus and is distributed under the same non-restrictive license. It has the same speakers and subset split as the LibriSpeech corpus, offering similar gender balance. The above-mentioned undesired properties in the LibriSpeech corpus are also addressed:

- *The audio files are at 24kHz sampling rate*; as most of the original material is recorded at 44.1 or 32kHz sampling rate, all audio with a sampling rate of less than 24kHz were excluded (two 16kHz files and six 22.05kHz files).
- *The speech is split at sentence breaks*; book-level texts are split into sentences using Google's proprietary sentence splitting engine then the audio was split at these sentence boundaries.
- *Both original and normalized texts are included*; the text has been normalized using Google's proprietary text normalization engine.
- *Contextual information (e.g., neighbouring sentences) can also be extracted*; additional text files provide easy access to neighbouring sentences.
- *Utterances with significant background noise are excluded*; utterance-level signal-to-noise ratio (SNR) is estimated and used to filter out noisy utterances.

The rest of this paper is organized as follows. Section 2 summarizes existing corpora used for TTS research. Section 3 describes the data processing pipeline used to produce the LibriTTS corpus from the original materials. Section 4 presents overall statistics of the corpus. Section 5 shows the experimental results of building TTS models with this corpus. Concluding remarks are given in the final section.

http://dx.doi.org/10.21437/Interspeech.2019-2441

Table 1: *List of publicly available English corpora which are often used in recent TTS research papers and their attributes.*

Corpus	Domain	License	Hours	Sampling rate (kHz)	Total speakers
ARCTIC [19]	Reading	BSD-style	7	16	7
VCTK [20]	Reading	ODC-By v1.0 [21]	44	48	109
BC2011 [22]	Reading	Non-commercial	16.6	16	1
BC2013 [23]	Audiobook	Non-commercial	300	44.1	1
LJSpeech [24]	Audiobook	CC-0 1.0 [25]	25	22.05	1
M-AILABS [26]	Audiobook	BSD-style	75	16	2
LibriSpeech [15]	Audiobook	CC-BY 4.0 [27]	982	16	2,484
LibriTTS	Audiobook	CC-BY 4.0 [27]	586	24	2,456

2. Related work

Having appropriate data facilitates exploration of new tasks and research ideas. Table 1 lists publicly available English speech corpora which are used in recent TTS research papers.

The CMU ARCTIC corpus [19] has been used for many years in statistical parametric speech synthesis research [28]. However, it is too small to train neural end-to-end TTS models. The VCTK corpus [20] is popular for experimenting with multi-speaker TTS [6, 8, 11, 14, 29–31] as it contains studio quality speech data from multiple speakers. The Blizzard Challenge 2011 (BC2011) [22] corpora provides a relatively large amount of reading speech from a single professional speaker. It is distributed under a non-commercial license. The LJspeech [24], M-AILABS [26], and Blizzard Challenge 2013 (BC2013) [23] corpora include tens of hours of audio and text from audiobooks read by single speakers. The LJspeech and M-AILABS corpora are comprised of non-professional audiobooks from the LibriVox project [16] and distributed under a non-restrictive license, whereas the BC2013 is read by a professional speaker but distributed under a non-commercial license like BC2011. As they are audiobook recordings, they contain expressive lines and a wide range of prosodic variation. They are often used for building single-speaker TTS voices [12, 13, 32–35].

3. The data processing pipeline

We align the long-form audio recordings with the corresponding texts, and split them into sentence-level segments. We also need to exclude utterances with audio/text mismatches which can be caused by inaccuracies in the text, reader-introduced insertions, deletions, substitutions, and transpositions, disfluencies, and text normalization errors. This section describes the pipeline which we developed to produce the LibriTTS corpus.

3.1. Text pre-processing

The first step in the pipeline is to split the book-level text into paragraphs/sentences and perform text normalization.

1. Book-level texts are first split into paragraphs at consecutive newlines.
2. Each paragraph text is further split into sentences by the proprietary sentence splitter.
3. Non-standard words (e.g., abbreviations, numbers and currency expressions) and semiotic classes [36] (text tokens and token sequences that represent particular entities that are semantically constrained, such as measure phrases, addresses and dates) in the sentences are detected and normalized by a weighted finite state transducer (WFST)-based text normalizer [37].

3.2. Extracting multi-paragraph text

In the original, unprocessed audio and text materials released from the LibriSpeech site, each audiobook consists of chapter-level audio files (in 128kbps MP3 format) whereas each text is a single file of the entire book. The second step extracts the partial text corresponding to each chapter-level audio file.

1. Run ASR (Google Cloud Speech-to-Text API [38]) on the chapter-level audio and get its transcription.
2. Extract chapter-level text from the book-level text by matching the transcription with the book-level text.

3.3. Align the audio and text

The third step is to align the audio with the extracted text. This is done by the engine used for YouTube's "auto-sync" feature. This feature allows video owners to upload a simple text transcript of the spoken content of a video as an alternative to automatically-created closed captions from ASR [39]. Auto-sync is also used to generate data for ASR acoustic model training [40]. Here we used a modified version of Auto-sync to generate the LibriTTS corpus. The uploaded transcript, containing no timing information, is force-aligned ("auto-sync'ed") using standard ASR algorithms to generate start and end times of each word [41].

1. A miniature tri-gram language model (LM) is generated using only the concatenated normalized sentences.
2. The audio is recognized using a decoder graph derived from the mini-LM in combination with a bidirectional long short-term memory (LSTM) acoustic model [42].
3. The decoding result is then edit-distance aligned to the normalized sentences. A sentence is marked as "aligned" if all words are matching (with edit-distance of zero).
4. Then start and end times for the sentences are generated from the decoding result.

3.4. Post processing

The final step is post-processing. We filter out possibly problematic lines based on heuristics found with other corpora and perform normalization.

- Filter out sentences with more than 71 words, which are likely to be affected by sentence splitting errors.
- Filter out utterances with a large average word duration, which is possibly the result of severe audio/text mismatch.

Table 2: *Data subsets in LibriTTS.*

Subset	Hours	Female speakers	Male speakers	Total speakers
dev-clean	8.97	20	20	40
test-clean	8.56	19	20	39
dev-other	6.43	16	17	33
test-other	6.69	17	16	33
train-clean-100	53.78	123	124	247
train-clean-360	191.29	430	474	904
train-other-500	310.08	560	600	1,160
Total	585.80	1,185	1,271	2,456

Table 3: *The numbers of sentences in the original partial text, filtered sentences, and final output.*

	train-clean-360	train-other-500
Original	262,107	332,816
Not aligned	78,058	125,806
Too long	1,805	1,409
Ave. word dur.	26	72
WADA-SNR	65,718	485
Final	116,500	205,044

- Normalize the polarity of audio by flipping up-side-down waveforms by ensuring DC offsets are positive.
- Run a silence end-pointer to trim start and end silences.
- Compute SNR of the audio using waveform amplitude distribution analysis (WADA) [43]. Audio with WADA-SNR < 20dB and < 0dB are filtered out from the "clean" and "other" subsets, respectively.

After the post-processing step, pairs of audio and texts (original and normalized) are generated to form the final corpus. The next section describes the statistics of the generated corpus.

4. Statistics

Table 2 provides a summary of all subsets in the LibriTTS corpus. The amount of yielded audio was significantly lower than that of the LibriSpeech corpus (about 60%). This is due to 1) stricter requirement in the alignment step (all words in a sentence must have confidence one) and 2) SNR-based filtering. Table 3 shows the numbers of sentences in the original partial text, filtered sentences, and the final output. It can be seen from the table that about 25% of sentences were filtered via the SNR threshold (20dB) for the "clean" subset. As the SNR threshold for the "other" subset is lower (0dB), the number of filtered sentences is less significant.

Figure 1 shows violin plots [44] of the number of character per sentence in the LibriSpeech and LibriTTS corpora and that of the original LibriVox materials. The figure shows that the distribution of the sentence lengths in the LibriTTS corpus is similar to that of the original LibriVox materials, whereas that of the LibriSpeech corpus is significantly different. Possibly splitting speech at silence intervals rather than sentence boundaries causes this mismatch; after a short sentence, as readers didn't need to take a breath, the lengths of pauses could be shorter than the threshold (0.3 seconds). Due to the heuristics to filter-out suspicious long sentences, the distribution for the LibriTTS corpus has shorter tail than that of the LibriVox materials.

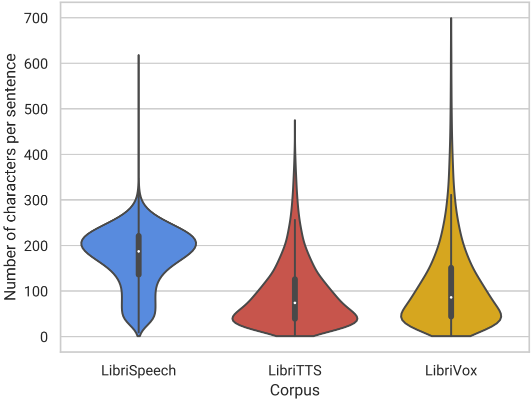

Figure 1: *Violin plots [44] of the number of characters per sentence on the LibriSpeech and LibriTTS corpora and that of the original LibriVox texts. The thick black bar in the center represents the interquartile range, the thin black line extended from it represents the 95% confidence intervals, and the white dot is the median. The width of the density plot indicates frequency, while each violin is normalized to have the same area.*

One side-effect of the more strict filtering applied in the LibriTTS data creation pipeline is the imbalance in terms of per-speaker duration. Figure 2 shows violin plots [44] of per-speaker audio duration on the LibriTTS and LibriSpeech corpora. It can be seen from the figure that the distribution of per-speaker audio duration on the LibriSpeech corpus has a sharp peak at its median (about 1,500 seconds). On the other hand, that of the LibriTTS corpus has a wider variance and lower median value (about 900 seconds). Furthermore, its diversity of total audio duration per speaker is much larger than that of the LibriSpeech corpus.

5. Experiments

This section presents TTS experimental results using models trained from the LibriTTS corpus to give baselines.

5.1. Experimental conditions

Gaussian mixture variational auto-encoder (GMVAE)-Tacotron models [13] were trained from the train-clean subsets of the LibriSpeech and LibriTTS corpora. The sizes of the latent attribute representation (size of latent vector) and the number of latent attribute classes (the number of mixture components in the Gaussian mixture prior) were 32 and 16, respectively. A speaker embedding table was used to give speaker identity conditioning. Each model was trained for at least 200k steps using the Adam optimizer [45]. Character sequences with punctuation marks from normalized texts were used as inputs of the network. WaveRNN [46]-based neural vocoders at 16kHz and 24kHz sampling rates were trained from the audio in the train-clean subsets of the LibriSpeech and LibriTTS corpora, respectively. At synthesis time, first a latent attribute class of the Gaussian mixture prior was randomly selected. Second, a mean vector associated with the selected Gaussian prior was used as the latent attribute representation. Third, a sequence of log-mel spectrogram was predicted from the normalized input text and

Table 4: *Five-scale subjective mean opinion scores with confidence intervals of the GMVAE-Tacotron models trained from the LibriSpeech and LibriTTS corpora. Total audio duration per speaker in seconds are are included at the last row.*

	Female			Male		
Dataset (sampling rate)	19	103	1841	204	1121	5717
LibriSpeech (16kHz)	4.03 ± 0.08	4.28 ± 0.07	4.07 ± 0.07	3.98 ± 0.07	3.95 ± 0.08	3.63 ± 0.08
LibriTTS (16kHz)	4.19 ± 0.07	4.35 ± 0.06	4.23 ± 0.07	4.01 ± 0.07	3.92 ± 0.08	3.55 ± 0.09
LibriTTS (24kHz)	4.39 ± 0.06	4.51 ± 0.06	4.40 ± 0.06	4.11 ± 0.08	4.05 ± 0.08	3.72 ± 0.09
Natural (24kHz)	4.57 ± 0.06	4.65 ± 0.09	4.57 ± 0.05	4.64 ± 0.10	4.59 ± 0.08	4.46 ± 0.06
Total audio duration (sec.)	64.5	182.2	1486.1	611.8	906.1	1360.5

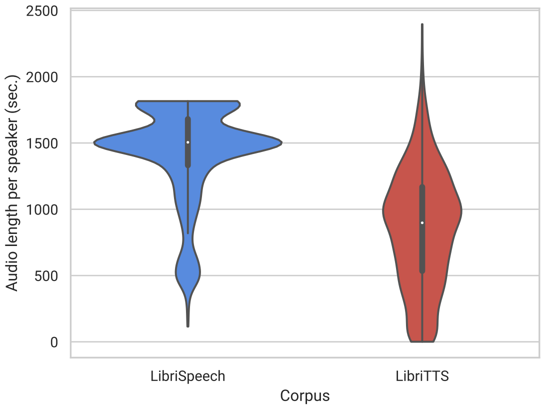

Figure 2: *Violin plots [44] of the total audio duration per speaker on the LibriTTS and LibriSpeech corpora.*

the latent attribute representation. Finally, a speech waveform was synthesized by driving the WaveRNN neural vocoder given the sequence of log-mel spectrogram. Please refer to [13] for details of the hyper-parameters.

Six readers (three female and three male) were selected from the train-clean subsets for evaluation. The female and male reader IDs were (19, 103, 1841) and (204, 1121, 5717), respectively. 620 evaluation sentences were randomly selected from the test subsets of the corpus. This set of evaluation sentences is also included in the release.

Quantitative subjective evaluations relied on crowd-sourced mean opinion scores (MOS) rating the naturalness of the synthesized speech by native speakers using headphones. After listening to each stimulus, a subject was asked to rate the naturalness of the stimulus in a five-point Likert scale score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent) in increments of 0.5. Each sample was rated by a single rater. To evaluate the effect of sampling rate, we also down-sampled the synthesized speech samples from the LibriTTS model to 16kHz and included them as test stimulus.

5.2. Results

Table 4 shows the experimental results. It can be seen from the table that LibriTTS (24kHz) achieved the best subjective scores with all speakers. The gaps in MOS between LibriTTS (16kHz) and LibriTTS (24kHz) for female and male speakers were 0.175 and 0.133, respectively. It clearly shows the benefit of having audio at higher sampling rate. On the other hand, those between LibriSpeech (16kHz) and LibriTTS (16kHz) were not statistically significant (0.15 for female and -0.03 for male). Although this result can suggest that preserving capitalization and punctuation marks were less important, this hypothesis is not fully confirmed as the size of the LibriTTS corpus is about half of that of LibriSpeech (245 hours vs. 460 hours).

It is interesting to note that subjective scores for male speakers were significantly lower than those for female speakers; the gaps between female and male were 0.48 and 0.26 for LibriTTS (24kHz) and LibriSpeech (16kHz), respectively. It indicates that the current model configuration is sub-optional for male speakers (e.g., filter-bank configuration, modeling dependency of time-domain signals by WaveRNN). Further experiments are required to fully understand the effect of these configurations.

Finally, there are still significant gap in MOS between natural and synthesized speech (0.16 for female, 0.61 for male). We need further work to improve the naturalness of synthesized speech on this task.

6. Conclusions

This paper introduced the LibriTTS corpus, which was designed for TTS use. It was derived from the original audio and text materials of the LibriSpeech corpus, by automatically aligning audiobooks and their texts, segmenting them into utterances, and filtering noisy transcripts and audio recordings. The corpus consists of 585 hours of speech data at 24kHz sampling rate from 2,456 speakers and its corresponding texts. To our knowledge this is the largest TTS-specific corpus. We demonstrated that Tacotron models trained from this corpus produced naturally sounding speech. This corpus is released online for public use; it is freely available for download from `http://www.openslr.org/60/`. We hope that the release of this corpus accelerates TTS research.

Future work includes evaluating the impacts of speaker imbalance, preserving punctuation marks and capitalization, and the relationship between amount of training data and the naturalness of synthesized speech. We also plan to expand this corpus by adding more speakers and languages.

7. Acknowledgements

The authors would like to thank Hank Liao and Hagen Soltau for their helpful comments about YouTube's auto-sync feature. We also thank Wei-Ning Hsu, Daisy Stanton, RJ Skerry-Ryan, Yuxuan Wang, and Markus Becker for helpful comments. We would like to express our gratitude to Guoguo Chen, Sanjeev Khudanpur, Vassil Panayotov, and Daniel Povey for releasing the LibriSpeech corpus, and to the thousands of Project Gutenberg and LibriVox volunteers.

8. References

[1] J. Sotelo, S. Mehri, K. Kumar, J. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-End speech synthesis," in *Proc. ICLR workshop*, 2017.

[2] S. Arik, C. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li *et al.*, "Deep Voice: Real-time neural text-to-speech," in *Proc. ICML*, 2017, pp. 195–204.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[4] J. Shen, R. Pang, J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[5] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," *arXiv:1809.08895*, 2018.

[6] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman *et al.*, "Deep Voice 2: Multi-speaker neural text-to-speech," in *Proc. NIPS*, 2017, pp. 2962–2970.

[7] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman *et al.*, "Deep Voice 3: 2000-speaker neural text-to-speech," in *Proc. ICLR*, 2018.

[8] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arxiv:1806.04558*, 2018.

[9] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," *arXiv:1808.10128*, 2018.

[10] S. Arik, J. Chen, P. W. Peng, Kainan and, and Y. Zhou, "Neural voice cloning with a few samples," *arXiv:1802.06006*, 2018.

[11] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, A. Wang *et al.*, "Sample efficient adaptive text-to-speech," *arXiv:1809.10460*, 2018.

[12] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, pp. 5167–5176.

[13] W.-N. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao *et al.*, "Hierarchical generative modeling for controllable speech synthesis," *arXiv:1810.07217*, 2018.

[14] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voice synthesis for in-the-wild speakers via a phonological loop," *arxiv:1707.06588*, 2017.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[16] "LibriVox – Free public domain audiobooks," https://librivox.org/.

[17] T. Capes, P. Coles, A. Conkie, L. Golipour, A. Hadjitarkhani, Q. Hu, N. Huddleston *et al.*, "Siri on-device deep learning-guided unit selection text-to-speech system," in *Proc. Interspeech*, 2017, pp. 4011–4015.

[18] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," *arXiv:1711.10433*, 2017.

[19] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, Tech. Rep. CMU-LTI-03-177, 2003.

[20] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR) http://dx.doi.org/10.7488/ds/1994, 2012.

[21] "Open Data Commons Attribution License (ODC-By) v1.0," https://opendatacommons.org/files/2018/02/odc_by_1.0_public_text.txt.

[22] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Blizzard Challenge Workshop*, 2011.

[23] ——, "The Blizzard Challenge 2013," in *Blizzard Challenge Workshop*, 2013.

[24] K. Ito, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[25] "CC0 1.0 Universal Public Domain Dedication," https://creativecommons.org/publicdomain/zero/1.0/.

[26] Munich Artificial Intelligence Laboratories GmbH, "The M-AILABS speech dataset," https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/.

[27] "Creative Commons Attribution 4.0 License (CC-BY 4.0)," https://creativecommons.org/licenses/by/4.0/.

[28] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commn.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[29] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," *arXiv;1802.06984*, 2018.

[30] Y. Lee, T. Kim, and S. Lee, "Voice imitating text-to-speech neural networks," *arXiv:1806.00927*, 2018.

[31] D.-R. Liu, C.-Y. Yang, S.-L. Wu, and H.-Y. Lee, "Improving unsupervised style transfer in end-to-end speech synthesis with end-to-end speech recognition," in *Proc. SLT*, 2018, pp. 640–647.

[32] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proc. ICML*, 2018, pp. 4700–4709.

[33] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," *arxiv:1808.01410*, 2018.

[34] K. Kastner, J. F. Santos, Y. Bengio, and A. C. Courville, "Representation mixing for TTS synthesis," *arXiv:1811.07240*, 2018.

[35] "GitHub – mozilla/TTS: Deep learning for Text to Speech," https://github.com/mozilla/TTS.

[36] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[37] P. Ebden and R. Sproat, "The Kestrel TTS text normalization system," *Journal of Natural Language Engineering*, vol. 21, no. 3, pp. 333–353, 2015.

[38] https://cloud.google.com/speech-to-text/.

[39] C. Alberti and M. Bacchiani, "Automatic captioning in YouTube," https://ai.googleblog.com/2009/12/automatic-captioning-in-youtube.html, 2009.

[40] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *Proc. ASRU*, 2013, pp. 368–373.

[41] P. J. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proc. ICASSP*, 2009, pp. 4869–4872.

[42] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014, pp. 338–342.

[43] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech*, 2009, pp. 2598–2561.

[44] J. L. Hintze and R. D. Nelson, "Violin plots: A box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[46] N. Kalchbrenner, E. Erich, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg *et al.*, "Efficient neural audio synthesis," *arXiv:1802.08435*, 2018.