Neutral TTS Female Voice Corpus in Brazilian Portuguese

Pedro H. L. Leite, Edmundo Hoyle, Álvaro Antelo, Luiz F. Kruszielski and Luiz W. P. Biscainho

***Abstract*— This paper introduces a new dataset designed to address the limitations in high-quality, diverse and representative datasets for training text-to-speech (TTS) models, specifically for female voices in Brazilian Portuguese. The dataset features a female voice recorded in a professional and controlled environment with neutral emotion and comprises more than 20 hours of recordings. The goal is to facilitate transfer learning and enable the development of more natural-sounding, high-quality, and gender-balanced TTS systems. Alongside the dataset, gender-aware voice transfer experiments are performed to understand the impact of utilizing gender-specific pretrained models for speech synthesis. The results obtained show that same-gender voice transfer yields better speech similarity and intelligibility when compared to cross-gender transfer, emphasizing the importance of gender-aware training procedures and highlighting the need for balanced gender data.**

***Keywords*— Text-to-speech, TTS, speech synthesis, Brazilian Portuguese, dataset, gender-aware, voice transfer.**

I. Introduction

Text-to-speech (TTS) systems have developed rapidly in the last few years, due to the success of some deep learning neutral network architectures/strategies such as Transformers, GANs and diffusion. These approaches have been the basis for the development of current state of the art TTS models (e.g. *Tacotron2* [1], *Fastspeech2* [2], *(Multiband)-MelGAN* [3], [4] and *FastDiff* [5]). Despite these advances, the availability of high-quality, diverse, and representative datasets for training TTS models remains a challenge for underrepresented languages. Although Brazilian Portuguese is one of the most widely spoken languages globally [6] and has been the focus of several TTS research efforts, there is still a gap in resources and specifically for female voices in professional recording conditions, which limits the development of more natural-sounding, high-quality and gender-balanced TTS systems.

As the human speech production is highly dependant on anatomy [7], there are significant differences for male and female speech. The fundamental frequency (usually called F0) trails generated by the vocal folds, for example, are generally higher for female than for male speakers because of strength, elasticity and size divergence. The so called formants (resonance frequencies of the vocal tract) also occur in different ranges, because the distances between the tract cavities are usually larger for men. In the face of those and other sources of variability, the development of gender-specific pre-trained models for speech synthesis is expected to help the processes of creating new voices in Brazilian Portuguese.

As shown in the literature review done in [8], there is an unbalance between male and female data for single-speaker voice datasets, specially in professional studios. Thus, we introduce in this paper a new dataset to address this issue: a Brazilian Portuguese TTS dataset featuring a female voice recorded with high quality in a controlled environment, with neutral emotion and more than 20 hours of recordings. Our dataset aims to facilitate transfer learning for researchers and developers working on TTS applications: a highly professional neutral female voice can serve as a good warm-up stage for learning language-specific structures, pronunciation and other non-individual characteristics of speech, leaving to further training procedures only to learn the specific adaptations needed (e.g. timbre, emotion and prosody). This can surely help enabling the accommodation of a more diverse range of female voices in Brazilian Portuguese. By doing so, we also hope to contribute to the development of accessible and high-quality TTS systems for several use cases such as virtual assistants, audiobooks, language learning tools and accessibility solutions. Considering the high quality and careful conditioning of the recorded dataset, we also anticipate its use for other audio applications, such as Automatic Speech Recognition (ASR) and Speech Enhancement.

Alongside the dataset, we propose voice transfer experiments to understand the impact of having both male and female datasets available to construct pre-trained TTS models. The goal is to show that female-to-female and male-to-male transfers yield faster and more reliable learning pipelines than the cross-gender case, and with better final synthesis quality.

In the next section, we present how the dataset was conceived and structured to be made available to users. Section III describes the voice transfer experiment, whose results are discussed in Section IV. Lastly, we make the final considerations and point out possible future contributions in Section V.

II. Dataset

To address TTS applications, the general data need is for prepared speech with corresponding transcripts, with the emotional, language and accent context that will be required at inference time. As this may not be achievable for the final target voices, a common approach is to gather a great collection of data with a voice that is not biased in any of those dimensions. Then, pretrained models can be developed to enable further voice transfer stages (generally short in time and data). As a way to develop such models in Brazilian

Pedro H. L. Leite, Audio Processing Group (GPA), Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro-RJ, e-mail: pedro.lopes@smt.ufrj.br; Edmundo Hoyle, Globo, Rio de Janeiro-RJ, e-mail: edmundo.hoyle@g.globo. This work was partially supported by National Council for Scientific and Technological Development (CNPq).

Portuguese and specifically for a female voice, we proceed to explain how the data distributed with this work is structured.

A. *Data Acquisition*

The written part of the dataset consists of short-duration excerpts of sentences sampled from the website of the main news program in Brazil (Jornal Nacional), virtually identical to the semantic contents of the dataset presented in [8] (except for a few sentences omitted for technical reasons), with a slightly longer total recording time. Audio content was recorded by a female professional narrator in a professional, noise-controlled studio environment. The speaker was asked to say the phrases out loud coherently, using neutral emotion and accent. As for the recording procedures, the signals were acquired using a Neumann TLM 103 cardioid microphone, then sampled at a rate of 96 kHz and stored in 24-bit PCM, reproducing the same conditions as [8]. The high sample and bit rate allows the acquisition of data to cover almost all speech phenomena we can hear with almost no distortion, ensuring noise and artifacts are not introduced at the earliest stage. The database is then structured as a set of *.wav* files with durations between 5.5 and 36 seconds (with an average of 15.4 seconds), each accompanied by a *.txt* file containing the matching transcript, also following the structure presented in [8].

The speech features as described above are important for the construction of a generalist TTS dataset in order to guarantee that the models are not affected by environmental noise or personal biases such as regionalisms and/or emotion, in a way that intelligibility is improved and linguistic attributes become clearer to learning procedures. In addition, its combination with [8] allows for a baseline comparison between female and male voices under similar conditions (recording procedure, total duration, texts and context).

B. *Pre-processing*

Both audio and text of the speech dataset were duly prepared prior to training procedures. Following the structuring of successful English datasets such as LJSpeech [9], VCTK [10] and LibriTTS [11], we segmented the signals to avoid large audio files (greater than 20 seconds) and to keep most speech excerpts between 5-10 seconds in duration. The intuition behind this procedure is that with smaller segments more data can fit into a single batch without allocation peaks, increasing the reliability of the statistics computed by learning models.

Firstly, the original text data was segmented according to commas and periods into smaller units. Next, the corresponding audio files were cut using Aeneas[1] software, which aligns the segmented text with the corresponding speech signal and provides the cutting instants. Then, the original audio files are divided into smaller cuts matched to the segmented texts.

After the signals had been aligned and segmented, a Voice Activity Detection (VAD) program[2] was used to remove breaths and to correct bad cuts, thereby improving the quality and accuracy of the speech data. The final formatted dataset contains 10333 speech segments whose duration distribution is depicted (in red) in Figure 1 along with the original duration histogram (in green), and the distribution of speech segments in the LJSpeech dataset (in blue) for comparison.

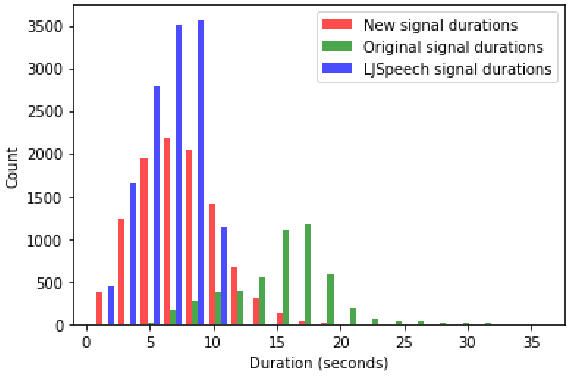

Fig. 1. Signal duration histograms of the original (in green) and highly segmented (in red) versions of the dataset, and of LJSpeech (in blue).

To increase the data similarity for the experiments that will be described in the next section and for possible further research purposes, we submitted the recordings in [8] through the same pre-processing stages, making the reformatted files available to the general public. Data access details are available at this work's companion website[3].

III. Gender Aware voice transfer experiment

In order to assess the impact of having a female specific dataset in the Brazilian Portuguese context, we conducted some voice transfer experiments involving (and not) gender exchanges. Initially, one male and one female TTS model were trained using *Tacotron2* for the generation of mel-spectrograms and *Multiband-MelGAN* for final waveform generation (neural vocoding), starting with a significant amount of recording hours (the dataset of [8] for the male voice and the dataset described in this work for the female voice). These pretrained models served as a starting point for training two target voices, again one male and one female, coming from the CETUC [12] dataset, which contains a shorter total duration (around 75 minutes) of recordings per speaker. After the text-to-mel models are trained on the larger datasets, they are fine-tuned to the target voices, having the waveform generation stage enabled with the pretrained vocoders. Using this setup, it is expected that adapting voices with similar vocal tracts (female to female or male to male) will be easier than when crossing their respective anatomies (as the learning procedures implicitly model physiology), yielding better final results in terms of quality and robustness. The speech synthesis procedures for the experiments are depicted in Figure 2.

To measure the similarity between the synthetic and real samples, we employ an embedding generator tool coming from a speaker verification task to create latent spaces that can represent individual speaker characteristics. The intuition is that after generating embeddings for the synthetic and real samples, the closer the synthetic embeddings are to the real

[1] `https://github.com/readbeyond/aeneas`
[2] `https://github.com/wiseman/py-webrtcvad`
[3] `https://www.smt.ufrj.br/gpa/sbrt2023/`

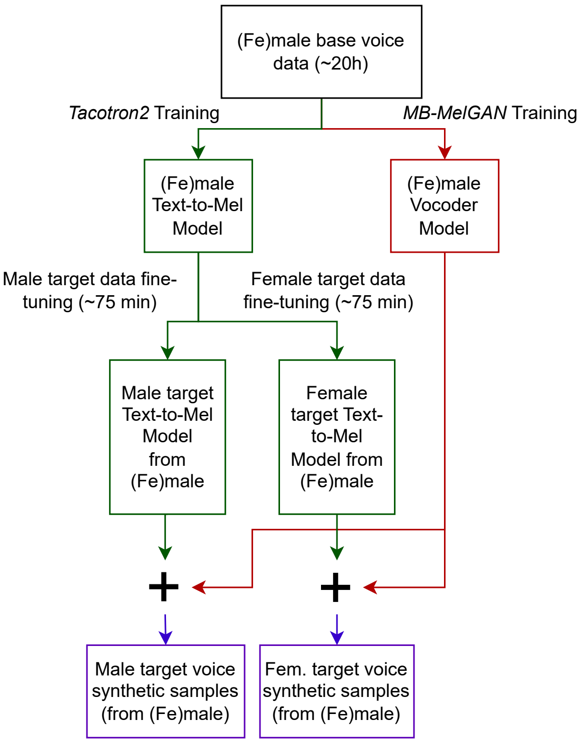
(Fe)male base voice data (~20h)
Tacotron2 Training
MB-MelGAN Training
(Fe)male Text-to-Mel Model
(Fe)male Vocoder Model
Male target data fine-tuning (~75 min)
Female target data fine-tuning (~75 min)
Male target Text-to-Mel Model from (Fe)male
Female target Text-to-Mel Model from (Fe)male
+
+
Male target voice synthetic samples (from (Fe)male)
Fem. target voice synthetic samples (from (Fe)male)

Fig. 2. Block diagram of the synthetic speech generation procedures. Data from the original voice with more recording time goes through two parallel training pipelines (Text-to-Mel and Vocoder), where it is converted into another voice with *Tacotron2* fine-tuning. The final synthesized samples are obtained combining the trained *Multiband-MelGAN* Vocoder with the fine-tuned *Tacotron2* model.

ones, the stronger the indication that the samples are alike. The embedding generator to be used in this work is part of the *dvector*[4] library, which implements the speaker verification procedure described in [13]. Thus, the procedure for computing the similarity measure is as follows: embeddings are generated from each audio signal, and normalized to have unit $L2$ norm; each normalized embedding is compared to the ground truth mean embedding using cosine similarity [14]–[16] (which is just the dot product in this case of normalized vectors). In this context, the cosine similarity metric gives us a 0 to 1 score of how similar the synthetic samples are to the mean real sample (in terms of the main abstract features used by our perception to distinguish between speakers).

In addition to the similarity assessment, the word error rate (WER) is used to compare the intelligibility of the produced samples, following the recent trend to use text-discrepancy metrics for objective evaluation of TTS systems [17], [18]. To compute it, a Speech-to-Text deep learning model with high margin of confidence [19] generates a concurrent transcript of the audio signals to be compared with the original text data, so that each discrepant word pair is counted as an error. With this test, we intend to indicate the ease of recognition of the semantic content of each speech sample.

Both metrics were calculated using unseen text and speech data from the target voice dataset, along with their paired synthesized version, as illustrated in Figure 3.

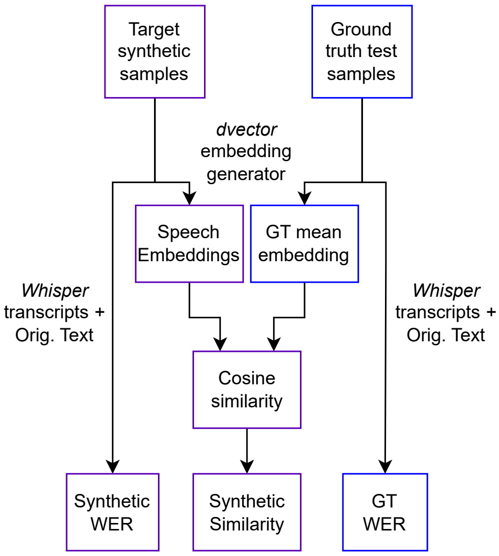
Target synthetic samples
Ground truth test samples
dvector embedding generator
Speech Embeddings
GT mean embedding
Whisper transcripts + Orig. Text
Whisper transcripts + Orig. Text
Cosine similarity
Synthetic WER
Synthetic Similarity
GT WER

Fig. 3. Block diagram of the metric calculations. Both natural and synthesized samples pass through two evaluations: WER is calculated comparing *Whisper* transcriptions with the original text present in the datasets; an embedding generator is applied to the signals to get corresponding speaker features. The synthetic embeddings are compared to the mean natural embedding with cosine similarity, providing a value for the proximity of speaker features.

IV. Voice Transfer Results

The *Tacotron2* model was initially trained for 100k steps with 32 samples per batch, in both female and male base voices, in $\approx$ 5 whole days on an NVIDIA QUADRO RTX 8000 GPU. For *Multi-band MelGAN*, 255k steps were run (200k steps with generator only and mixed precision and 55k steps with the discriminator training active and no mixed precision, to avoid instability) with 512 batch size, in $\approx$ 3 days on a NVIDIA GEFORCE RTX 3090 GPU. The fine-tuning stages took 40k steps (in $\approx$ 2 days) for *Tacotron2*, keeping the same configurations of the first training procedure.

In the similarity outcomes, shown in Figure 4, it can be observed that the voice transfer between speakers of the same gender (male to male and female to female) results in higher levels of similarity between the synthetic samples and the original voices. This suggests that employing gender-aware transfer learning can effectively help to preserve gender-specific characteristics, resulting in an easier replication of the target voice. On the other hand, the cross-gender cases present greater challenges to replicate the voices identities with the same level of fidelity, as the similarity scores for them are comparatively lower.

Regarding intelligibility results, summarized in Table I, no fundamental difference was found when assessing the female target voice case: the WER values for cross-gender and single-gender transfers are very close to each other. However, it is noticeable that for the male target voice there is a meaningful difference between the results in the cross-gender and single-gender scenarios.

A mel spectrogram comparison is presented in Figures 5 and 6 for both male and female target voices. In the upper map, a real recorded sample is shown, demonstrating the original frequency content to be compared with the subsequent

[4] https://github.com/yistLin/dvector

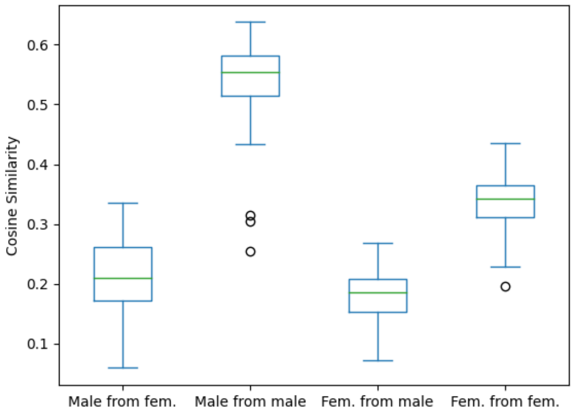

Fig. 4. Box plot for cosine similarity results.

TABLE I
INTELIGIBILITY RESULTS.

Experiment	WER (%)
Male (from fem.)	29.8
Male (from male)	12.5
Fem. (from male)	8.0
Fem. (from fem.)	8.1
Male (ground truth)	4.8
Fem. (ground Truth)	1.8

synthetic samples coming from the gender transfer experiment. We note that the models were free to create prosody and therefore not forced to generate aligned speech with the ground truth samples, meaning that the speech patterns can be disjoint in time.

Regarding the male target voice (in Figure 5), it is possible to note that there are significant differences in the replicas: although the general trails for F0 are in the right place, the ressonance patterns, duration of the speech segments and energy distribution over the frequency axis do not match with the real sample. Moreover, the cross-gender case seems to have worsened these discrepancies with less detail capabilities and the occurrence of some artifacts.

For the female target (in Figure 6), the original frequency shapes seem to have been better learned by both models, but with more (and sharper) details in the samples created by the single-gender modeling. The F0 curves and energy distribution are very well-defined and visually coherent with the reference in this case, contributing to the better intelligibility results.

Some of the audio samples used in the generation of the metrics and images are available on our website[5].

[5] `https://www.smt.ufrj.br/gpa/sbrt2023/audio_samples/`

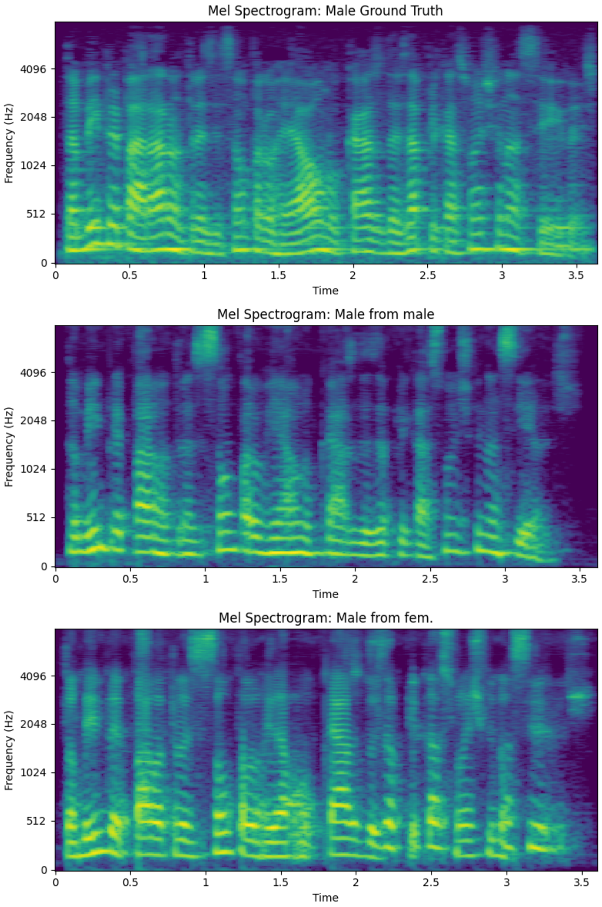

Fig. 5. Mel spectrograms for the male target voice. At the top, the ground truth real sample, followed by the male from male and male from female transfer experiments, respectively.

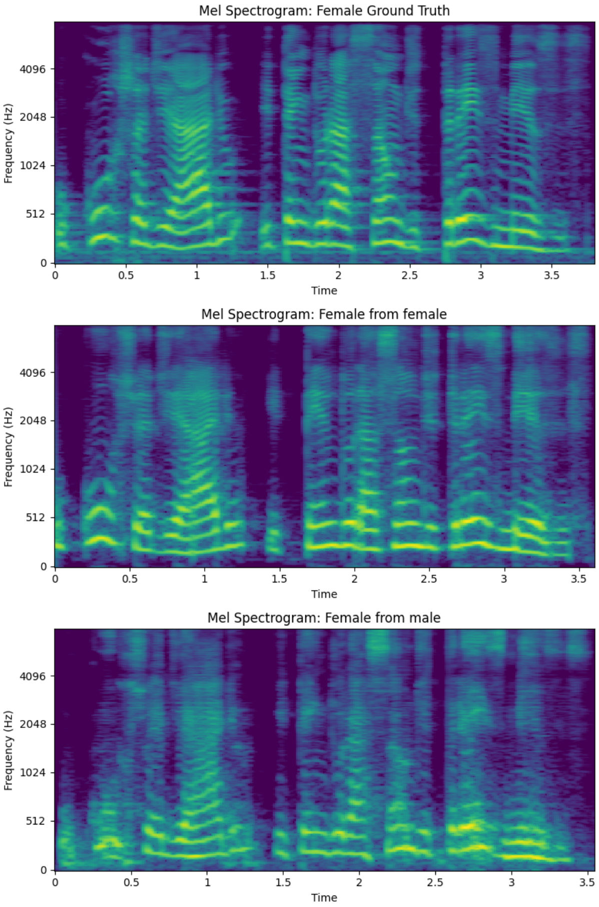

Fig. 6. Mel spectrograms for the female target voice. At the top, the ground truth real sample, followed by the female from female and female from male transfer experiments, respectively.

V. Conclusion and future works

After all considerations, the experiments conducted and the results provided in this paper emphasize the importance of considering gender-specific characteristics during voice transfer in TTS applications. The observations indicate that gender-aware learning procedures might positively impact the similarity and intelligibility of the final synthesized speech, which induces the need for balanced gender data. In this context, we hope that the dataset made public with this work can help structure more robust TTS models in Brazilian Portuguese, especially for female voices.

There are still gaps to be investigated regarding gender-aware TTS model training experiments. The measurement of the impact of fine-tuning the vocoder with target data and conditioning it to mel spectrogram predictions, the influence of gender balanced data in multi-speaker and multi-language simultaneous training possibilities and training single-stage (text-to-waveform) models are possible paths to be followed in future works, not to mention the need for subjective tests to better measure the perceptive impact of considering gender aspects in voice transfer procedures.

References

[1] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4779–4783.

[2] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, Sep. 2021, poster.

[3] K. Kumar, R. Kumar, T. de Boissiere, *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, vol. 32, Dec. 2019.

[4] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 492–498, May 2020.

[5] R. Huang, M. Lam, J. Wang, D. Su, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, Jul. 2022, pp. 4132–4138.

[6] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the world. twenty-sixth edition.* `https://www.berlitz.com/blog/most-spoken-languages-world`, Online, accessed in 08 Aug 2023, 2023.

[7] P. C. Loizou, "Speech production and perception," in *Speech Enhancement: Theory and Practice, Second Edition*. Boca Raton: CRC Press, 2007, pp. 45–65.

[8] P. H. L. Leite, E. Hoyle, Á. Antelo, L. F. Kruszielski, and L. W. P. Biscainho, "A corpus of neutral voice speech in Brazilian Portuguese," in *Computational Processing of the Portuguese Language*, 2022, pp. 344–352.

[9] K. Ito and L. Johnson, *The LJ speech dataset*, `https://keithito.com/LJ-Speech-Dataset/`, Online, accessed in 18 Apr 2023, 2017.

[10] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)*, `https://doi.org/10.7488/ds/2645`, Online, accessed in 18 Apr 2023, 2019.

[11] H. Zen, V. Dang, R. Clark, *et al.*, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019)*, Sep. 2019, pp. 1526–1530.

[12] V. Alencar and A. Alcaim, "LSF and LPC-derived features for large vocabulary distributed continuous speech recognition in Brazilian Portuguese," in *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, Sep. 2008, pp. 1237–1241.

[13] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, *Generalized end-to-end loss for speaker verification*, `https://arxiv.org/abs/1710.10467`, Online, accessed in 15 May 2023, 2017.

[14] H. Zeinali, A. Mirian, H. Sameti, and B. BabaAli, "Non-speaker information reduction from cosine similarity scoring in i-vector based speaker verification," *Computers & Electrical Engineering*, vol. 48, pp. 226–238, Oct. 2015.

[15] W. Ahmad, H. Karnick, and R. M. Hegde, "Cosine distance metric learning for speaker verification using large margin nearest neighbor method," in *Advances in Multimedia Information Processing – PCM 2014*, W. T. Ooi, C. G. M. Snoek, H. K. Tan, C.-K. Ho, B. Huet, and C.-W. Ngo, Eds., Cham: Springer International Publishing, Dec. 2014, pp. 294–303.

[16] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in *Proceedins of The Speaker and Language Recognition Workshop (Odyssey 2010)*, Jul. 2010, paper 16.

[17] C. Wang, S. Chen, Y. Wu, *et al.*, *Neural codec language models are zero-shot text to speech synthesizers*, `https://arxiv.org/abs/2301.02111`, Online, accessed in 15 May 2023, 2023.

[18] D. Lim, S. Jung, and E. Kim, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH 2022)*, Sep. 2022, pp. 21–25.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, `https://arxiv.org/abs/2212.04356`, Online, accessed in 15 May 2023, 2022.