TTS applied to the generation of datasets for automatic speech recognition

Edresson Casanova[1,2]**, Sandra Aluísio**[1]**,** and **Moacir Antonelli Ponti**[1,3]
[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil;
[2] NVIDIA; [3] Mercado Livre, Brazil.

Abstract

Despite automatic speech recognition (ASR) systems evolution with deep learning methods, for languages with a shortage of open/public resources, the resulting systems still present low-quality performance. On the other hand, Text-to-Speech (TTS) has also evolved in the last decade, allowing for zero-shot multi-speaker TTS (ZS-TTS) models to generate speech of a target speaker using only a few seconds of its speech. These advances motivated the use of ZS-TTS in the training of ASR systems to improve the performance of the models. However, ZS-TTS models still require a large number of diverse speakers and hours of speech during training, thus hindering their practical use in languages with less accessible data. In this work, we explored ZS-TTS in scenarios with few available speakers. We proposed the use of flow-based models due to its state-of-the-art (SOTA) results and explored the use of multilingual models, seeking to leverage available data from languages with many available speakers. The results achieved by this work made possible the development of ZS-TTS and zero-shot voice conversion (VC) systems in languages with few available speakers. The approach proposed in this work was applied to improve ASR systems in two languages, simulating a scenario with only one speaker available for the training of the ZS-TTS model. Despite using only one speaker in the target languages, our data augmentation approach achieved results comparable to the SOTA in the English language.

1 Introduction

Text-to-Speech (TTS) systems have garnered significant attention in recent years due to the advancements in deep learning. These breakthroughs have enabled their widespread use in applications like virtual assistants, allowing TTS models to attain a level of naturalness akin to human speech (Shen et al., 2018; Valle et al., 2020; Kim et al., 2020). Nonetheless, the majority of TTS systems are designed for a single speaker, even though many applications could benefit from synthesizing new speakers not seen during training, utilizing only a few seconds of target speech. This approach is referred to as zero-shot multi-speaker TTS (ZS-TTS) (Jia et al., 2018).

Advances in TTS technology have also inspired research that leverages it to enhance Automatic Speech Recognition, as demonstrated in studies like Li et al. (2018); Rosenberg et al. (2019); Laptev et al. (2020). Most of these studies employ pre-trained TTS models to generate ASR data, using the LibriSpeech dataset (Panayotov et al., 2015) for ASR model training. While Li et al. (2018), used three speakers from the American English M-AILABS dataset (Solak, 2019) for TTS model training, Rosenberg et al. (2019) and Laptev et al. (2020) trained their TTS models with over 251 speakers from LibriSpeech. These papers showcased that ASR models trained with a combination of synthesized speech and human speech achieved relative improvements ranging from 0.79% to 4.56% when compared to models trained solely on human speech. However, a substantial disparity was observed between models trained with only human speech and those trained with only synthesized speech, with relative differences of 80.17% and 78.98%, respectively, in the case of Li et al. (2018), and Rosenberg et al. (2019). This stark contrast highlights the need for further research and enhancements in this field.

1.1 Gaps

Although previous work shows the potential of multi-speaker TTS models for ASR data augmentation, these models still require high-quality datasets with many speakers and hours of speech to converge (Laptev et al., 2020). Generally, such models are trained on English with big datasets such as LibriSpeech and LibriTTS[1], which is not suitable for

[1] https://www.openslr.org/60/

medium/low-resource languages that do not have a public multi-speaker TTS dataset.

Although some multilingual multi-speaker datasets were released in recent years (Pratap et al., 2020; Elizabeth et al., 2021), they just attend a small number of languages and for many applications, even these may not be sufficient to build a competitive ASR system. In addition, creating a high-quality multi-speaker dataset is hard, because it requires the effort of multiple target-language speakers. It is especially hard for languages with small populations, where recruiting participants is difficult, or in more extreme scenarios with languages that are almost extinct and have just a few speakers (e.g. indigenous languages). In a range of scenarios creating a high-quality multi-speaker dataset is not viable. In light of this, an approach that applies TTS/VC for ASR data augmentation that requires just a medium/low-quality few speakers dataset could make the application of this technology viable for languages that really need it, helping to preserve/protect nearly extinct languages, for example.

1.2 Research Question and Hypothesis

Given that ZS-TTS systems require datasets with a large number of speakers for its convergence, is it possible to overcome this limitation and obtain a ZS-TTS system in languages for which the number of available speakers tends to one?

The hypothesis is that a flow-based model, such as Glow-TTS (Kim et al., 2020), adapted for zero-shot multi-speaker training can achieve convergence with a smaller number of speakers. Also, it is possible to train by taking advantage of the number of speakers present in other languages and, in this way, reduce the number of speakers needed for training in the target language.

1.3 Main goal and specific objectives

The main goal of this work was to propose an approach for training a ZS-TTS model in languages where just a small number of speakers are available to make the use of TTS applied to the ASR task viable. In addition, to evaluate the behavior of these methods in languages other than English, in particular Portuguese, and to investigate methods that work with multiple languages.

To achieve the main goal, the following specific objectives were defined: (1) Develop and make publicly available a dataset for TTS in Brazilian Portuguese (Section 2); (2) Propose a new model SOTA ZS-TTS model that can achieve good results with a smaller number of speakers (Section 3); (3) Investigate and propose adaptations to the model proposed in (2) for training with multiple languages (Section 4); and (4) Exploration of the model proposed in (3) in ASR models training (Section 5).

This extended thesis abstract will be organized as follows. The next sections will introduce the main papers of the thesis including a small abstract describing the importance of the paper on the thesis scope. Section 6 presents a summary of the contributions of this Ph.D. research to the speech processing field and a list of all publications carried out during this thesis development. The full thesis is available at: https://doi.org/10.11606/T.55.2022.tde-02092022-142539

2 TTS-Portuguese Corpus

During the Ph.D. research, there were no publicly available datasets with a sufficient number of hours and audio quality to train deep learning-based TTS models in Brazilian Portuguese. For this reason, in Casanova et al. (2022a), we proposed and made publicly available the TTS-Portuguese Corpus. TTS-Portuguese Corpus consists of 10.5 hours of speech from a single native Brazilian Portuguese speaker. We did experiments with the novel dataset and we showed that it can be used to achieve SOTA results in Brazilian Portuguese. The obtained results using the Tacotron 2 model are comparable to the original work that was trained using the English language (Shen et al., 2018) and the current SOTA in European Portuguese (Quintas and Trancoso, 2020).

3 SC-GlowTTS

Despite recent advances, ZS-TTS is still an open problem, there is still a large voice similarity gap between speech generated for seen and unseen speakers. Furthermore, in 2020 normalizing flows (or flow-based models) have been successfully applied in the TTS field, achieving SOTA results (Valle et al., 2020; Kim et al., 2020). Despite this, ZS-TTS models were still heavily based on the Tacotron 2 model (Shen et al., 2018). Tacotron 2-based ZS-TTS models require a large number of speakers for training, making it impossible to obtain good-quality models in languages with few resources available. For these reasons, in (Casanova et al., 2021d), we proposed the flow-based model SC-GlowTTS. SC-GlowTTS is an efficient ZS-

TTS model that improves similarity for speakers unseen during training, achieving SOTA results. We showed that our model can be trained with only 11 speakers achieving results comparable to a Tacotron 2-based ZS-TTS model trained with 98 speakers. In addition, SC-GlowTTS is faster than previous ZS-TTS models and it achieves real-time in CPU. SC-GlowTTS implementation and checkpoints are open-source and it can be found at https://github.com/Edresson/SC-GlowTTS.

4 YourTTS

According to (Tan et al., 2021), the quality of current ZS-TTS models is not good enough, especially for target speakers with speech characteristics very different from those seen in training. Although SC-GlowTTS has achieved SOTA results, the gap between speakers seen in training and new ones is still an open research question. Furthermore, ZS-TTS still requires multi-speaker datasets, making it difficult to obtain high-quality models in really low-resource languages. Despite the promising results of SC-GlowTTS model using just 11 speakers, generally limiting the number of speakers in training makes it even more difficult to generalize the model to speakers with speech characteristics very different from those seen in training.

For these reasons, in Casanova et al. (2022b), we proposed YourTTS model. We explored the use of a multilingual approach, taking advantage of the number of speakers available in a language with many resources available (e.g. English) to help the convergence of the model in a low-resource language. YourTTS was trained with 1249 speakers in English from VCTK[2] and LibriTTS datasets, 5 speakers in French (Solak, 2019), and a single male speaker in Portuguese (Casanova et al., 2022a). YourTTS achieved SOTA results in ZS-TTS and results comparable to SOTA in zero-shot voice conversion in English. Additionally, our approach achieves promising results in the Portuguese language using only a single-speaker dataset, opening possibilities for ZS-TTS and zero-shot voice conversion systems in low-resource languages. Even more, the YourTTS model was able to produce female voices in Portuguese even though it was not trained with female voices in this language. To address the voice similarity gap for speakers who have voice or recording conditions that differ greatly from those seen in training we proposed a fine-tuning approach. We showed that it is possible to fine-tune the YourTTS model with less than 1 minute of speech and improved a lot the voice similarity for these speakers, in this way solving the gap. An interesting application for fine-tuning is for patients who have voice problems, such as aphonia and dysphonia, which in some cases can cause total loss of voice. YourTTS can be applied to improving the well-being of these patients, allowing “as far as possible" to preserve and recover their voices digitally.

Since its publication, YourTTS has been referred to as SOTA in the literature and it has been used as a baseline for several papers in the TTS (Wang et al., 2023; Le et al., 2023; Jiang et al., 2023; Liu et al., 2023) and voice conversion (Li et al., 2023a; Hussain et al., 2023; Li et al., 2023b,c) field.

5 ASR data augmentation in low-resource

In Casanova et al. (2023), we proposed a novel approach for ASR data augmentation. Our approach is based on cross-lingual multi-speaker TTS and cross-lingual voice conversion and it uses YourTTS model. Through extensive experiments, we showed that our approach permits the application of TTS and voice conversion to improve ASR systems using only one target-language speaker during the TTS model training. We also managed to close the gap between ASR models trained with synthesized versus human speech compared to other works that use many speakers. Finally, we showed that it is possible to obtain promising ASR training results with our data augmentation approach using only a single real speaker in two target languages. Figure 1 shows a full ASR data augmentation diagram pipeline using only a single real speaker in the target languages. The ASR model trained only with one real speaker using human and augmented data reached a Word Error Rate of 33.96% and 36.59%, respectively, for the test set of the Common Voice dataset in Portuguese and Russian. In this way, our approach makes possible the training of a competitive ASR system in a target language using only approximately 10 hours of speech from a single speaker. This advance can help to preserve almost extinct languages that have a small number of speakers available like indigenous languages. Currently, we are working together with the PROINDL[3] challenge team of the Center for Artificial Intelligence IBM/Fapesp on the application

[2] https://datashare.ed.ac.uk/handle/10283/3443

[3] https://c4ai.inova.usp.br/pt/pesquisas/#PROINDL_port

of this approach in Brazilian indigenous languages that have few or even only one single-speaker data available.

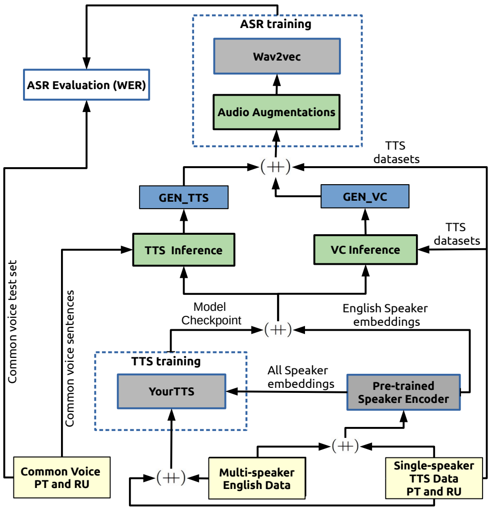

Figure 1: ASR data augmentation diagram pipeline, adopted from (Casanova et al., 2023)

6 Conclusions and Future Work

The main goal of this Ph.D. research was to propose an approach for training a ZS-TTS model in languages where just a small number of speakers are available to make the use of TTS applied to the ASR task viable. The main goal of this research was achieved after a series of studies. We also showed that it is possible to overcome the limitation and obtain a ZS-TTS system in languages for which the number of available speakers tends to one, confirming our hypothesis and answering our research question.

To make our main goal possible we needed to contribute to TTS and voice conversion fields by creating data resources (Section 2) and proposing new SOTA models (Sections 3 and 4). We also needed to propose a novel data augmentation approach for ASR, contributing directly to this field (Section 5). In addition, during this Ph.D., we also made other contributions in these fields and also in other speech fields that are not fully correlated to the thesis's main goal.

In Candido Junior et al. (2022), we contribute to the ASR field via the creation and release of a large Brazilian Portuguese dataset, called CORAA ASR. CORAA ASR is composed of 290.77 hours of spontaneous and prepared speech.

In Casanova et al. (2021b), we proposed a new method for speaker verification systems training, called Speech2Phone. Speech2Phone achieved results near the SOTA using almost 500 times less data during training.

During the COVID-19 pandemic, we participated in the SPIRA project, working on identifying respiratory failure through speech collected from COVID-19 patients (Casanova et al., 2021c). In the project, we developed a solid base of speech studies as a biomarker, thus allowing the faster development of identifiers through speech in future pandemics. Additionally, in Casanova et al. (2021a), we won the COMPARE (Schuller et al., 2021) COVID-19 identification through cough challenge that was organized at INTERSPEECH 2021.

Given that, this Ph.D. thesis had important contributions in the TTS, voice conversion, ASR, speaker verification, and illness identification fields. It also had a social impact because the methods developed in this thesis can be used as a tool to help preserve near-extinct languages and also to improve or create TTS, voice conversion, and ASR systems in all low-resource languages.

6.1 Publications

Table 1 presents in chronological order all the papers published during this Ph.D. research.

Papers
CABEZUDO, M. A. S.; INÁCIO, M.; RODRIGUES, A. C.; CASANOVA, E.; DE SOUSA, R. F.. **NILC at ASSIN 2: Exploring Multilingual Approaches**. In: ASSIN@STIL. 2019. p. 49-58.
CABEZUDO, M. A. S.; INÁCIO, M.; RODRIGUES, A. C.; CASANOVA, E.; DE SOUSA, R. F.**Natural Language Inference for Portuguese Using BERT and Multilingual Information**. In: Proceedings of The International Conference on the Computational Processing of Portuguese (PROPOR). Springer, Cham, 2020. p. 346-356.
CASANOVA, E.; TREVISO, M.; HÜBNER, L.; ALUÍSIO, S.. **Evaluating Sentence Segmentation in Different Datasets of Neuropsychological Language Tests in Brazilian Portuguese**. In: Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020). Marseille, France: European Language Resources Association (ELRA), 2020. p. 2605-2614.
GRIS, L. R. S.; CASANOVA, E.; DE OLIVEIRA, F. S.; SOARES, A. S.; CANDIDO Jr, A.. **Desenvolvimento de um modelo de reconhecimento de voz para o Português Brasileiro com poucos dados utilizando o Wav2vec 2.0**. In Anais do XV Brazilian e-Science Workshop. SBC., 2021. p. 129-136.
CASANOVA, E. ; GRIS, L. ; CAMARGO, A. ; SILVA, D. ; GAZZOLA, M.; SABINO, E.; LEVIN, A.; CANDIDO JR, A. ; ALUISIO, S.; FINGER, M.. **Deep learning against covid-19: Respiratory insufficiency detection in brazilian portuguese speech**. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. ACL, Aug. 2021.

Papers
CASANOVA, E.; CANDIDO JR, A.; FERNANDES JR, R. C.; Finger, M.; GRIS, L.; PONTI, M. A.. **Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021**. In: Proceedings of INTERSPEECH. ISCA, Aug. 2021.
CASANOVA, E.; SHULBY, C.; GÖLGE, E.; MÜLLER, N. M.; DE OLIVEIRA, F. S.; CANDIDO Jr, A. ; SOARES, A. S.; ALUISIO, S.; PONTI, M. A.. **SC-GlowTTS: an Efficient Zero-Shot Multi-Speaker Text-To-Speech Model**. In: Proceedings of INTERSPEECH. ISCA, Aug. 2021.
LEAL, S.; CASANOVA, E.; PAETZOLD, G.; ALUISIO, S.. **Evaluating Semantic Similarity Methods to Build Semantic Predictability Norms of Reading Data**. In: Proceedings of the 24th International Conference on Text, Speech and Dialogue, TSD 2021. ISCA, Sept. 2021.
CASANOVA, E.; CANDIDO JR, A.; SHULBY, C.; DE OLIVEIRA, F. S., GRIS, L. R. S., DA SILVA, H. P.; PONTI, M. A. **Speech2Phone: A Novel and Efficient Method for Training Speaker Recognition Models**. In: Brazilian Conference on Intelligent Systems. Springer, Cham, Dec. 2021. p 572-585.
CASANOVA, E.; CANDIDO JR, A.; SHULBY, C.; DE OLIVEIRA, F. S.; TEIXEIRA, J. P.; PONTI, M. A.; ALUISIO, S.. **TTS-Portuguese Corpus: a corpus for speech synthesis in Brazilian Portuguese**. In: Language Resources and Evaluation (LREV). Springer, 2022.
CASANOVA, E.; WEBER, J.; SHULBY, C.; JUNIOR, A. C.; GÖLGE, E.; PONTI, M. A.. **YourTTS: Towards ZS-TTS and Zero-Shot Voice Conversion for everyone**. In: Proceedings of International Conference on Machine Learning (ICML). PMLR, 2022.
CANDIDO JR, A.; CASANOVA, E.; SOARES, A.; DE OLIVEIRA, F. S.; OLIVEIRA, L.; JUNIOR, R. C. F.; ... ; ALUISIO, S.. **CORAA: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese**, In: Language Resources and Evaluation (LREV). Springer, 2022.
CASANOVA, E.; SHULBY, C.; KOROLEV, A.; CANDIDO JR, A.; SILVA, A.; ALUÍSIO, S.; PONTI, M. A.. **ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion**.In: Proceedings of INTERSPEECH. ISCA, Aug. 2023.

Table 1: List of published papers.

Acknowledgements

This study was funded by CAPES – Finance Code 001, CNPq grant 304266/2020-5, FUNAPE via CEIA, and FAPESP/IBM Corporation via C4AI-USP grant #2019/07665-4.

References

Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, et al. 2022. Coraa asr: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *Language Resources and Evaluation*, pages 1–33.

Edresson Casanova, Arnaldo Candido Jr., Ricardo Corso Fernandes Jr., Marcelo Finger, Lucas Rafael Stefanel Gris, Moacir Antonelli Ponti, and Daniel Peixoto Pinto da Silva. 2021a. Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021. In *Proc. Interspeech 2021*, pages 446–450.

Edresson Casanova, Arnaldo Candido Junior, Christopher Shulby, Frederico Santos de Oliveira, Lucas Rafael Stefanel Gris, Hamilton Pereira da Silva, Sandra Maria Aluísio, and Moacir Antonelli Ponti. 2021b. Speech2phone: a novel and efficient method for training speaker recognition models. In *Brazilian Conference on Intelligent Systems*, pages 572–585. Springer.

Edresson Casanova, Lucas Gris, Augusto Camargo, Daniel da Silva, Murilo Gazzola, Ester Sabino, Anna Levin, Arnaldo Candido Jr, Sandra Aluisio, and Marcelo Finger. 2021c. Deep learning against covid-19: respiratory insufficiency detection in brazilian portuguese speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 625–633.

Edresson Casanova, Arnaldo Candido Junior, Christopher Shulby, Frederico Santos de Oliveira, João Paulo Teixeira, Moacir Antonelli Ponti, and Sandra Aluísio. 2022a. Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese. *Language Resources and Evaluation*, pages 1–13.

Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Jr., Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti. 2021d. SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model. In *Proc. Interspeech 2021*, pages 3645–3649.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion. In *Proc. INTERSPEECH 2023*, pages 1244–1248.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022b. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Salesky Elizabeth, Wiesner Matthew, Bremerman Jacob, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W Oard, and Post Matt. 2021. The multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech 2021*, pages 3655–3659. ISCA - International Speech Communication Association.

Shehzeen Hussain, Paarth Neekhara, Jocelyn Huang, Jason Li, and Boris Ginsburg. 2023. Ace-vc: Adaptive and controllable voice conversion using explicitly disentangled self-supervised speech representations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*, pages 4480–4490.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Chen Zhang, Zhenhui Ye, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun Ma, et al. 2023. Mega-tts 2: Zero-shot text-to-speech with arbitrary length speech prompts. *arXiv preprint arXiv:2307.07218*.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *arXiv preprint arXiv:2005.11129*.

Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. 2020. You do not need more data: improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*.

Jason Li, Ravi Gadde, Boris Ginsburg, and Vitaly Lavrukhin. 2018. Training neural speech recognition systems with synthetic speech augmentation. *arXiv preprint arXiv:1811.00707*.

Jingyi Li, Weiping Tu, and Li Xiao. 2023a. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2023b. Slmgan: Exploiting speech language model representations for unsupervised zero-shot voice conversion in gans. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2023c. Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 920–927. IEEE.

Zhijun Liu, Yiwei Guo, and Kai Yu. 2023. Diffvoice: Text-to-speech with latent diffusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210. IEEE.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *Proc. Interspeech 2020*, pages 2757–2761.

Sebastião Quintas and Isabel Trancoso. 2020. Evaluation of deep learning approaches to text-to-speech systems for european portuguese. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 34–42. Springer.

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. 2019. Speech recognition with augmented synthesized speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002. IEEE.

Björn W Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, I Lefter, Heysem Kaya, Shahin Amiriparian, and LJM Rothkrantz. 2021. The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, volume 6. International Speech Communication Association.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4779–4783. IEEE.

Imdat Solak. 2019. The m-ailabs speech dataset.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Rafael Valle, Kevin J Shih, Ryan Prenger, and Bryan Catanzaro. 2020. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. In *International Conference on Learning Representations*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.