Advancing Text-to-Speech Synthesis in Code-Switching Scenarios (Bachelor Thesis)

In the realm of human-robot interaction and multimedia content creation, text-to-speech (TTS) models play a pivotal role in crafting immersive and engaging experiences. Whether integrated into robotic systems or utilized for tasks like Face-dubbing and video generation, the quality of TTS significantly influences the perceived realism and overall effectiveness of the synthetic content.

This research endeavor seeks to delve into the landscape of state-of-the-art (SOTA) approaches [7],[1][5][6][2], examining their efficacy in the domain of Code-Switching (CSW) speech synthesis. Our primary objectives encompass:

- **Dataset Preparation:**
  - Investigate diverse SOTA models employing varied input-output configurations.
  - Implement data augmentation techniques to enhance dataset robustness.
- **Model Training:**
  - Implement and train models tailored for speech synthesis.
- **Performance Evaluation:**
  - Conduct quantitative assessments to gauge the models' performance.
  - Evaluate and identify the most promising model for this specific application.

Throughout the research journey, you will receive guidance from experts specializing in Code-Switching speech recognition and multilingual speech synthesis. Access to cutting-edge GPUs will provide a robust foundation for your experiments.

Successful outcomes of this study will be compiled into a research paper for submission to a prestigious conference. If you are intrigued by this research opportunity, please submit your application, including your CV and transcript of records, to:

Enes Ugan
Email: enes.ugan@kit.edu
Website: `https://isl.anthropomatik.kit.edu/english/21_9532.php`

Related research

[3] Improving Code-Switching and Named Entity Recognition in ASR with Speech Editing based Data Augmentation
[4] Zero-shot code-switching ASR and TTS with multilingual machine speech chain

References

[1] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *CoRR*, abs/2106.06103, 2021.

[2] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale, 2023.

[3] Zheng Liang, Zheshu Song, Ziyang Ma, Chenpeng Du, Kai Yu, and Xie Chen. Improving code-switching and named entity recognition in asr with speech editing based data augmentation. *arXiv preprint arXiv:2306.08588*, 2023.

[4] Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Zero-shot code-switching asr and tts with multilingual machine speech chain. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 964–971. IEEE, 2019.

[5] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017.

[6] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023.

[7] Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling, 2023.