Enhancing Expressive TTS Synthesis for Multilingual Low-Resource Languages: Challenges and Applications

Rami Kammoun, Mohamed Salah Al-Radhi, Géza Németh
Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics, Budapest, Hungary

rami.kammoun@edu.bme.hu, {malradhi,nemeth}@tmit.bme.hu

Abstract

Text-to-speech (TTS) synthesis has undergone tremendous evolution over the past few decades, from early articulatory approaches to more sophisticated statistical methods, and now, cutting-edge deep learning-based solutions. These advancements have significantly improved the naturalness and intelligibility of synthesized speech. Autoregressive models played a fundamental role in this progression by maintaining accurate temporal alignments for generating long phrases. However, non-autoregressive models have further enhanced TTS systems by enabling faster inference through parallel computation, thus overcoming some limitations of autoregressive models. These modern systems, powered by attention mechanisms, have also been able to integrate additional acoustic features, such as duration, pitch, and energy, which further contribute to generating more natural and expressive speech.

Despite the significant progress in TTS technology for rich-resource languages, such as English, which benefits from a large quantity of high-quality audio and transcription datasets, there remain considerable challenges in developing expressive TTS for low-resource languages. These languages often lack sufficient, high-quality data, making it difficult to achieve the same level of naturalness and expressivity as rich-resource counterparts. This research focuses on addressing these challenges, particularly for languages such as the Arabic-Tunisian dialect, emotional French, Kazakhstani, and German.

Tunisian Arabic, in particular, is a dialect that has traditionally been classified as low-resource, with limited datasets available for TTS development. However, recent advancements have seen the release of new datasets, opening up opportunities for further research. In this work, we aim to fine-tune existing state-of-the-art models on these emerging datasets to improve expressivity in low-resource languages. Specifically, the ArTST system, which has already demonstrated superior results for Modern Standard Arabic, will be fine-tuned on an expressive mono-speaker dataset for Tunisian Arabic. This system will be tested to assess whether undiacritized speech could lead to better results, especially given the low-resource nature of the dialect. By applying these advanced models to expressive Tunisian speech, we aim to achieve the highest possible levels of subjective evaluation for naturalness and expressivity.

In addition, other state-of-the-art models, YourTTS and FastSpeech2 have shown promise in generating natural speech in rich-resource settings. Therefore, we aim to assess their performance in handling the complexities of expressive speech synthesis for underrepresented languages.

Special emphasis will be placed on understanding how self-supervised models like VITS can help bridge the gap between low and high-resource languages. While such models have shown remarkable capabilities in some domains, we expect that the small amount of available data for our target languages may limit the improvements achieved through self-supervision alone.

In conclusion, this study aims to push the boundaries of expressive TTS synthesis for low-resource languages by leveraging state-of-the-art models and exploring their applicability in multilingual contexts. By addressing the challenges posed by limited data availability, this research seeks to enhance the naturalness and expressivity of TTS systems, opening new paths for practical applications in a variety of domains.

Keywords

Multilingual TTS, low-resource languages, emotional speech synthesis, speech naturalness

Acknowledgements

This paper is supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) and by the Ministry of Innovation and Culture and the National Research, Development and Innovation Office of Hungary within the framework of the National Laboratory of Artificial Intelligence. M.S.Al-Radhi's research was supported by the EKÖP-24-4-II-BME-197, through the National Research, Development and Innovation (NKFI) Fund.

References

[1] Toyin, H. O., Djanibekov, A., Kulkarni, A., & Aldarmaki, H. ArTST: Arabic text and speech transformer, Proceedings of ArabicNLP, Singapore, pp. 41-51, 2023.

[2] Laouirine, I., Kammoun, R., & Bougares, F. TunArTTS: Tunisian Arabic text-to-speech corpus. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pp. 16879–16889, 2024.

[3] Diatlova, D., & Shutov, V. EmoSpeech: Guiding FastSpeech2 towards emotional text to speech, 12th ISCA Speech Synthesis Workshop (SSW), Grenoble, France, 2023.

[4] Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., & Ponti, M. A. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone, Proceedings of the 39th International Conference on Machine Learning (PMLR) 162:2709-2720, 2022.