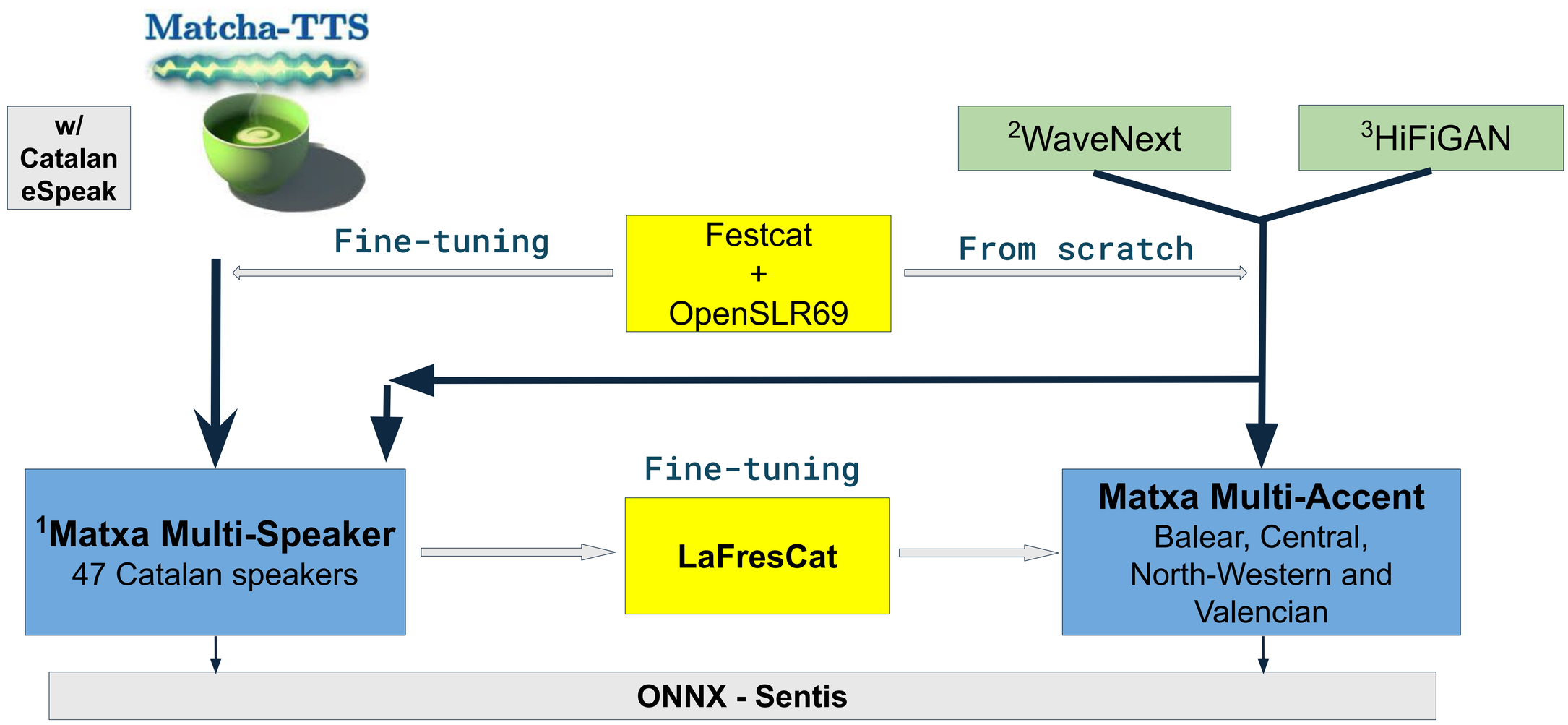
Multi-speaker and Multi-dialectal Catalan TTS Models for Video Gaming

Alex Peiró-Lilja[1,2], José Giraldo[1], Martí Llopart-Font[1], Carme Armentano-Oller[1], Baybars Külebi[1], Mireia Farrús[2]
[1] Barcelona Supercomputing Center (BSC), Spain,
[2]Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona, Spain

Abstract

Recently, we explored and trained different state-of-the-art text-to-speech (TTS) architectures for Catalan. We used existing datasets but also we produced a new Catalan multi-accent dataset to train these architectures. The objective of our work is to improve the quality of current TTS systems in Catalan and export the resulting models for potential interactive applications and video games. For this reason, our set of multi-speaker and multi-accent Catalan TTS models are presented within a demo made in Unity. The users are able to interact with game characters which are attached to our Catalan TTS. Generated Catalan speech inferences are played while execution time, real-time factor and translated transcripts are shown on screen.



Data

- **[4]Festcat:** Studio quality recordings of the Festival suite in Catalan.
- **[5]OpenSLR69:** A crowd-sourced collection of transcribed audio of Catalan sentences recorded by volunteers.
- **[6]LaFrescat:** We produced the first Catalan multi-accented and multispeaker dataset.

Setup

ONNX model sizes were reduced with onnx-simplifier[7]. Sentis open beta library to import deep learning models into Unity projects. We exported our models with Sentis 1.4.0-pre.2 version. Unity dev editor version that accepts this library is Unity 2023.3.0 Beta 10.

Evaluation

	Size (MB)	RTF (GPU)	RTF (CPU)
Mtch+HFG	123	0.013 (0.003)	0.089 (0.007)
Mtch+WN	**122**	**0.010 (0.001)**	**0.087 (0.010)**

Table 1: *TTS models comparison in terms of size and RTF.*

Demo

The game is controlled using the classic **WASD** keys for movement, allowing players to navigate the environment seamlessly. Press **K** to activate Matcha-WaveNext and **L** to activate Matcha-HiFiGAN, adding dynamic audio effects to the gameplay.

[1] S. Mehta, R. Tu, J. Beskow, Éva Székely, and G. E. Henter, "Matcha-tts: A fast tts architecture with conditional flow matching," 2024.

[2]J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," 2018.

[3]J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," 2020.

[4]https://huggingface.co/datasets/projecte-aina/festcat_trimmed_denoised

[5]https://huggingface.co/datasets/projecte-aina/openslr-slr69-ca-trimmed-denoised

[6]https://huggingface.co/datasets/projecte-aina/LaFrescat

[7]https://github.com/daquexian/onnx-simplifier

*https://bakudas.itch.io/generic-rpg-pack (sprites and tiles used in the demo)

This work has been promoted and financed by the Generalitat de Catalunya through the Aina project.