Interspeech 2024
1-5 September 2024, Kos, Greece

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

{alexandre.peiro,jose.giraldo,marti.llopart,carme.armentano,baybars.kulebi}@bsc.es, mfarus@ub.edu

Abstract

Recently, we explored and trained different state-of-the-art text-to-speech (TTS) architectures for Catalan. We used existing datasets but also 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. While generated Catalan speech replies are reproduced, execution time, real-time factor and transcription are shown on screen. Exported weights, such as data and demo source code, are released to the public.

Index Terms: text-to-speech, human-computer interaction, videogaming, Catalan

1. Introduction

Architectures based on deep learning have evolved drastically in a short time for all technological branches. Recently, with the optimization of generative models, their integration into different applications is increasingly feasible. In video gaming, specifically, the inclusion of language and speech models is getting more attention for a deeper interaction between the player and characters. However, the most recent models and resources are always focused on the majority languages, such as English. Our goal is to bring Catalan, a language spoken by some 9 million speakers and with more limited resources, to the state of the art in speech synthesis and its inclusion into potential interactive applications such as video games. Latest open-source neural TTS in Catalan was *Catotron* [1] –based on Tacotron2 [2] and MelGAN [3]– which brought the synthesis of Catalan speech closer to the human level. However, newer architectures like diffusion-based offer a faster and lighter performance, keeping or even improving quality and naturalness of speech synthesis. In this light, we present our released models through an interactive demo developed in Unity[1]. With this we not only show that the most recent TTS architectures can be competent on performing in video games, but also we open the Catalan to the community of developers and researchers to take advantage of our models. Exported models, data and code are public.

[1]https://unity.com

2. Implementation

2.1. Data

2.1.1. Festcat and OpenSLR69

To train the model with a central Catalan accent we used two multi-speaker datasets: Festcat and OpenSLR-69 Catalan subset[4, 5]. The former focused on creating Catalan speech for the Festival suite. And the latter contains transcribed high-quality audio of Catalan sentences recorded by volunteers. A pre-processing pipeline was applied to these datasets, which includes resampling, de-noising with CleanUnet[6] and the trimming of start and end silences with WebRTC VAD. The processed data is publicly available[2][3].

2.1.2. LaFresCat multi-accent dataset

We produced LaFresCat dataset (to be published soon) to finetune multi-accent models. This dataset consists of a total of 3.5 hours of studio recordings. For this dataset, two voices, one female and one male, were chosen for each of the four main accents of Catalan (Balearic, Central, North-Western and Valencian), and between 20 and 30 minutes of each were recorded.

2.2. Models

On the acoustic feature prediction side of the TTS, we chose Matcha-TTS [7], a non-autorregressive encoder-decoder model trained with optimal-transport conditional flow matching. The encoder part is based on a text encoder and a phoneme duration prediction that together predict averaged acoustic features. And the ODE-based decoder is a U-Net capable of generating high output quality in few synthesis steps. On the vocoder side, we trained HifiGAN [8] and WaveNext [9]. The former uses as generator a block of transposed convolutions that upsamples an acoustic or hidden latent representation. The latter is a variant of Vocos[10] architecture that replaces the ISTFT operation with a linear layer to predict directly the waveform samples instead of the fourier coefficients. Both Vocos and WaveNext use ConvNeXt blocks inheriting the advantages in performance of this architecture. We choose WaveNext over Vocos as it can be fully exported to ONNX.

2.3. Training

Matcha multispeaker was finetuned from an English multispeaker pre-trained checkpoint. The embedding layer was reset and adjusted to our number of speakers –in total 47–. Vocoders

[2]https://huggingface.co/datasets/projecte-aina/festcat_trimmed_denoised,
[3]https://huggingface.co/datasets/projecte-aina/openslr-slr69-ca-trimmed-denoised

	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.*

were trained from scratch using 30.5 hours, WaveNext was modified to use the same acoustic features of HifiGAN. In total 5 versions of Matcha-TTS were trained: multi-speaker, balear, valencian, north-western and central. Since our produced multi-accent dataset has few samples per speaker, they were finetuned from the multi-speaker version.

3. Demo

3.1. Setup

The demo is developed and tested on an MSI Intel Core i7 12th Gen with 64GB of RAM and a GPU NVIDIA RTX 3070 Laptop. We opted for Unity game engine because of its accessibility to resources created by other developers. The trained TTS models were integrated using Sentis[4], a library for neural networks inference for Unity which is still an open beta. We worked under the Unity 2023.3.0 Beta 10 and the latest Sentis 1.4.0-pre.2 versions. Due to publication reasons, Matcha and the vocoders were exported separately to ONNX and then merged as end-to-end models. By using onnx-simplifier[5] the size of ONNX models was reduced. Finally, they were imported to Unity and serialized into sentis format. In Table 1 we show a comparison in terms of model size and real-time factor (RTF) of the two TTS configurations: Matcha+HiFiGAN (Mtch+HFG) and Matcha+WaveNext (Mtch+WN)[6]. WaveNext seems to perform a little bit faster and also it is slightly lighter than HiFiGAN.

3.2. Functionalities

The user will be able to move and interact with game objects –pixel art characters–. Each character stores a specific speaker ID and a sentence. When the user clicks on any of these characters another invisible game object that contains a TTS *Manager* launches the inference. Thus, the same TTS model can be shared by different characters using different speaker IDs. Different scenes are created to separate multi-accent and multi-speaker models. The user can choose between CPU or GPU backend and also the vocoder to test inference speeds and quality. At each interaction, the user can check the execution time, real-time factor and transcriptions in English. Note that we are currently working on the Catalan phonemizer for Unity applications, so a set of input sentences were pre-processed.

3.3. Interface

The interface is shown in Figure 1. There are always two boxes: on the bottom and on the top left side of the screen. The former shows the transcription in English –as we expect most of the users will not be Catalan speakers– when the TTS is activated. And the latter depicts the activated model name, the execution time and the real-time factor. Sprites and tiles used to build scenarios were taken from *Generic RPG Pack* created by *Estúdio Vaca Roxa* under a Creative Commons 0 (CC0) license[7].

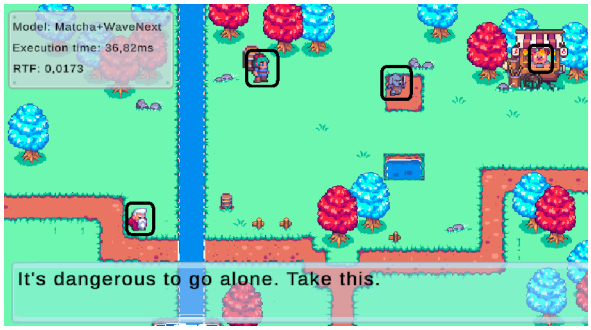

Figure 1: *Frame of a scene with characters marked with a black square.*

4. Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina project. Part of the training of the model was possible thanks to the compute time given by Galician Supercomputing Center CESGA (Centro de Supercomputación de Galicia), and also by Barcelona Supercomputing Center in MareNostrum 5.

5. References

[1] B. Külebi, A. Öktem, A. Peiró-Lilja, S. Pascual, and M. Farrús, "Catotron – a neural text-to-speech system in catalan," in *In: Interspeech 2020, Shangai, China. (Online)*, 2020.

[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] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.

[4] A. Cávez, I. Esquerra, L. Aguilar, S. Martínez, and A. Moreno, "Recent work on the festcat database for speech synthesis," 01 2009.

[5] O. Kjartansson, A. Gutkin, A. Butryna, I. Demirsahin, and C. Rivera, "Open-Source High Quality Speech Datasets for Basque, Catalan and Galician," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. Marseille, France: European Language Resources association (ELRA), May 2020, pp. 21–27.

[6] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," 2022.

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

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

[9] T. Okamoto, H. Yamashita, Y. Ohtani, T. Toda, and H. Kawai, "Wavenext: Convnext-based fast neural vocoder without istft layer," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[10] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," 2023.

[4] https://docs.unity3d.com/Packages/com.unity.sentis@1.4/manual/index.html

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

[6] https://huggingface.co/BSC-LT/sentis-matxa-tts-wavenext-multispeaker-ca

[7] https://bakudas.itch.io/generic-rpg-pack