

CIDAI Centre of Innovation
for Data tech
and Artificial Intelligence

MATXA

The first speech open-source
solution to support the
different Catalan varieties



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



**Generalitat
de Catalunya**



CIDAI-POC-2024-07

Some rights reserved. This work is available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International ([CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)).

Under the terms of this licence, you may copy, redistribute and adapt the work for non-commercial purposes, provided the work is appropriately cited, as indicated below.

In any use of this work, there should be no suggestion that CIDAI endorses any specific organization, products or services. The use of the CIDAI logo is not permitted.

If you adapt the work, then you must license your work under the same or equivalent Creative Commons licence.

If you create a translation of this work, you should add the following disclaimer along with the suggested citation: “This translation was not created by the Centre of Innovation for Data tech and Artificial Intelligence (CIDAI). CIDAI is not responsible for the content or accuracy of this translation. The original English edition shall be the binding and authentic edition”.

Any mediation relating to disputes arising under the licence shall be conducted in accordance with the mediation rules of the [World Intellectual Property Organization](https://www.wipo.int/mediation/).

Suggested citation. CIDAI-POC-2024-02 // MATXA - The first speech open-source solution by Baybars Külebi, Jose Omar Giraldo Valencia, Alexandre Cristian Peiro Lilja, Carme Armentano Oller, Marti Llopart Font, Ignasi Esquerra - BSC, CIDAI, 2024. Licence: [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)



1.	ABSTRACT	4
2.	DESCRIPTION OF THE PROBLEM SOLVED	7
3.	IMPLEMENTATION OF THE SOLUTION	8
4.	IMPACT OF THE SOLUTION	12
5.	CITED WORKS	13
6.	BIBLIOGRAPHY	14



1. Abstract

1.1. Summary

Matxa is a groundbreaking text-to-speech technology developed in Catalan, employing deep learning to revolutionise administrative tasks for the visually impaired. It is the first multispeaker, multidialectal neural TTS model, and comes together with the vocoder model alVoCat to generate high quality and expressive speech efficiently in four Catalan dialects:

- Balear
- Central
- North-Occidental
- Valencian

Both models are trained with open data; Matxa models are free (as in freedom) to use for non-commercial purposes, but for commercial purposes it needs licensing from the voice artist.

The significance of open-source text-to-speech (TTS) technologies for minority languages cannot be overstated. These technologies democratise access to TTS solutions by providing a framework for communities to develop and adapt models according to their linguistic needs. This is why we have developed different open-source TTS solutions in Catalan, using an ensemble of technologies. One of them is that Matxa synthesizes natural-sounding voices suitable for administrative texts and tasks, being its primary objective to assist visually impaired individuals in comprehending administrative content more effectively. The development of Matxa represents a signif-

icant advancement in assistive technology, addressing the specific needs of the visually impaired community in accessing administrative information. By leveraging the capabilities of deep learning, Matxa can accurately interpret and vocalise complex texts, ensuring that users receive accurate and comprehensive information.

Looking to the future, our goal with Matxa is to expand its capabilities even further by incorporating additional dialects and officially standardised versions of certain dialects. This expansion will enable users to choose voices that closely reflect their regional and cultural backgrounds, enhancing their sense of identity and connection with the synthesized speech. Additionally, we aim to introduce more expressive voices that convey emotions and nuances, making the interaction with administrative tasks even more engaging and personalized for visually impaired individuals. By continuously improving and diversifying the voices available in Matxa, we strive to ensure that all users can access administrative information in a manner that is both inclusive and reflective of their linguistic and emotional preferences.

In addition, we've created a [Hugging Face space](#) that provides a user-friendly interface where users can easily test and download models for natural language processing tasks. With intuitive features, users can explore various models, assess their performance, and seamlessly download them for use in their projects.

Matxa and alvoCat models have been promoted and financed by the Government of Catalonia through the [Aina project](#).

1.2. Objectives

The Matxa project has three primary objectives:

1- Developing Natural-Sounding TTS Synthesis for Catalan Dialects:

The project aims to create the most natural-sounding TTS synthesizer for various Catalan dialects, leveraging advanced linguistic and phonetic modelling. It will implement deep learning techniques to capture the nuances of each dialect, ensuring accurate representation in synthesised speech. As we have shown, the combination of Matcha with the Vocos vocoder achieves high levels of intelligibility and expressiveness, catering to the diverse linguistic landscape of Catalonia.

2- Seamless Integration with the administration for Visually Impaired Assistance:

Integration of the TTS synthesizer seamlessly with administrative systems will aid visually impaired individuals in their tasks. The project has developed interfaces that allow easy integration with administrative software, potentially enabling efficient access to documents, forms, and other materials. Compatibility with assistive technologies and standards will be ensured to maximise accessibility and usability for all users.

3- User-Friendly Interface and Model Download:

The project will provide the most user-friendly interface for interacting with the TTS synthesizer, with intuitive controls and clear documentation. It will enable users to test and download models easily, facilitating experimentation and customization for specific needs. Ensuring that the interface is accessible to users of all levels of technical expertise will promote widespread adoption and usage of the synthesizer.

1.3. Challenge that is being addressed

The development of Matxa presents several challenges, particularly concerning data quality, the type of data used, and representation of speakers and dialects. While the technology aims to provide natural-sounding voices for administrative tasks in Catalan, ensuring the accuracy and quality of the synthesized speech across multiple dialects poses a significant challenge. Variations in pronunciation, intonation, and vocabulary among different dialects require comprehensive data representation to capture the nuances effectively. Moreover, maintaining data quality for training the neural TTS models is crucial, as inaccuracies or biases in the data can lead to poor synthesis results, especially for minority languages like Catalan where available datasets may be limited.

Additionally, capturing the rich diversity of Catalan dialects through just two speakers, one female and one male, is a formidable task. Even within a single dialect, there can be significant variations in pronunciation, rhythm, and intonation depending on factors such as age, social background, and geographical location. Therefore, representing the full spectrum of linguistic diversity within each dialect requires a more extensive range of speakers. However, sourcing and recording data from a wide array of speakers, each representing different linguistic nuances, poses logistical challenges and may not always be feasible. Thus, the technology must strike a balance between the breadth of representation and the practical constraints of data collection and processing, ensuring that the synthesized voices remain authentic and comprehensible across dialectal variations.



2. Description of the problem solved

In today's societies, great importance is given to the written word, and it is considered common to provide important information in writing. While public administration has many platforms of open access, making it increasingly common for citizens to access official information and carry out procedures comfortably from home via the internet, people with visual impairments have been excluded from this trend. It forces them either to lose their autonomy by asking for help from a close person or to travel to official offices.

Our new TTS models are designed to read web pages and documents from official bodies (municipalities and social security). They are initially intended to help people with visual impairments, but they can also be useful for citizens with reading comprehension difficulties, who find it easier to understand texts read aloud. Secondly, they will also help alleviate congestion in public service offices, as many people who were previously obliged to travel there to carry out official procedures, will now be able to do them directly from home.

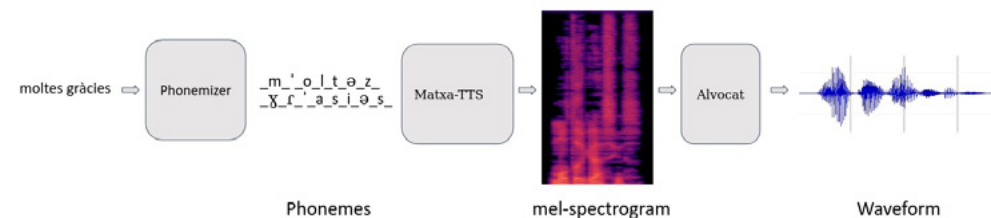
Our TTS models have three features that make them stand out from the rest: the first is that they are distributed freely, making them accessible to anyone who wants to use them. The second is that they can be installed on the user's computer or phone, ensuring their privacy. Finally, they allow you to choose between the 4 main accents of the Catalan language. This serves a dual purpose: on the one hand, the user can choose the accent they like the most or find most familiar. On the other hand, it helps to dignify the plurality of accents in the language, as often only the majority accent is taken into consideration.



3. Implementation of the solution

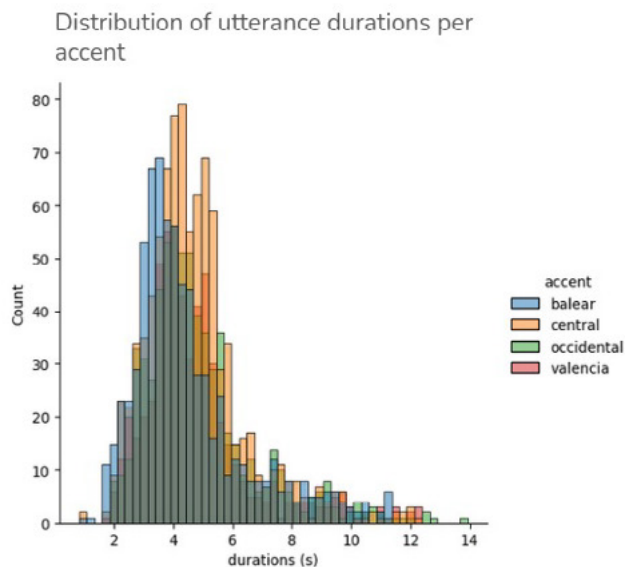
Our text-to-speech model tailored for Catalan employs a multi-step process to convert written text into spoken words with accurate pronunciation. These are the steps:

1. Initially, the model analyzes the input text, breaking it down into smaller linguistic units such as words and sentences while identifying any special characters. It then utilizes our version of eSpeak, a speech phonemizer, to generate phonemes based on the Catalan language's phonetic rules. For each Catalan accent, certain specifically adapted eSpeak rules apply.
2. The matcha-TTS model converts these phonemes into a mel spectrogram, a visual representation of the spectrum of frequencies of a sound over time.
3. This spectrogram is then fed into our adaptation of the Vocos vocoder, which synthesizes the speech waveform.

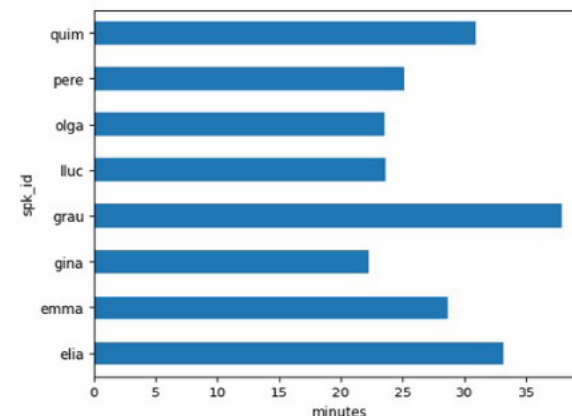


3.1. Multi-accent Data Produced

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.



Total durations per speaker



In the following example extracted from our eSpeak, clear phonetic differences can be observed:

Original sentence: “Volem sentir la teva veu perquè és molt important”

Balearic: vol'ɛn sɛnt'i lə t'evə v'ɛw pɛrk'ə 'əz m'olt import'ant

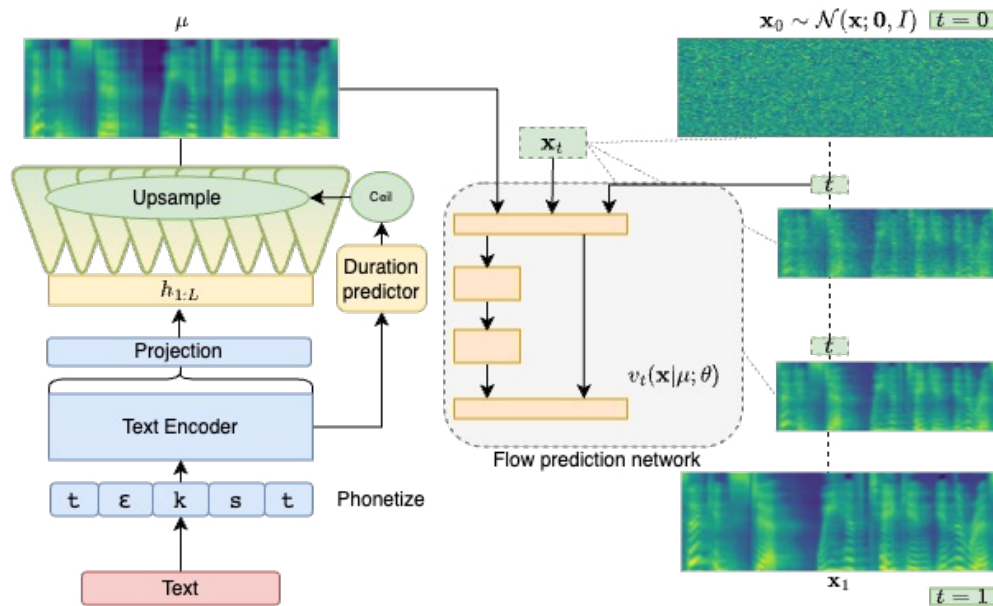
Central: bul'ɛm sɛnt'i lə t'eβə β'ɛw pɛrk'ɛ 'ez m'ol impurt'an

North-Western: bol'em sɛnt'i la t'ewɛ β'ew pɛrk'e 'ez m'ol import'an

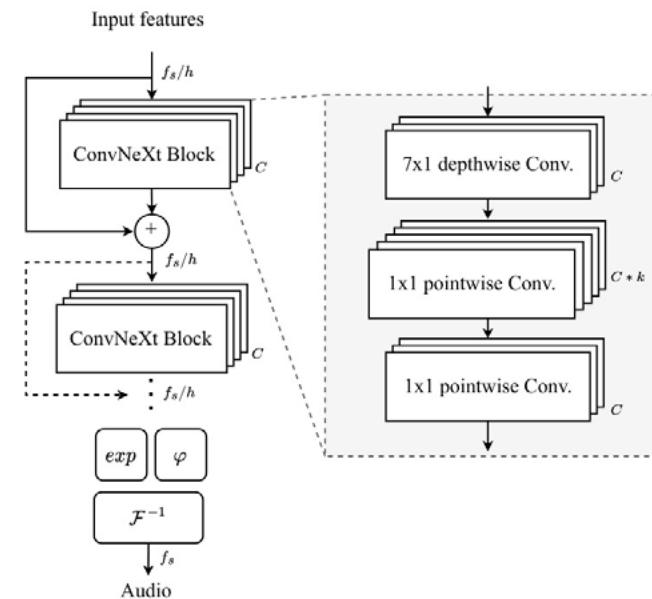
Valencian: vol'em sɛnt'ir la t'ewa v'ew pɛrk'e 'ez m'olt import'ant

3.2. Model Architectures

Matxa is based on Matcha-TTS[1], a non-autoregressive encoder-decoder model designed for fast acoustic modelling. The encoder processes input sequences of phonemes and, together with a phoneme duration predictor, predicts averaged acoustic features. And the decoder, whose backbone is essentially a U-Net based on the Transformer architecture [2], predicts the refined spectrogram. The model is trained with optimal-transport conditional flow matching. This yields an ODE-based decoder capable of generating high output quality in fewer synthesis steps.



AlVoCat is an adapted version of the recently published vocoder named Vocos[3]. It is a fast neural vocoder designed to synthesise audio waveforms from acoustic features. Unlike other typical GAN-based vocoders [4, 5], Vocos does not model audio samples in the time domain. Instead, it generates spectral coefficients, facilitating rapid audio reconstruction through inverse Fourier transform. The goal of this model is to provide an alternative to hifi-gan that is faster and compatible with the acoustic output of several TTS models. This version is tailored for the Catalan language, as it was trained only on Catalan speech datasets.



One of the main innovations of Vocos that allows a fast performance is the use of ConvNeXt [6] blocks as can be seen in the above figure, coupled with ISTFT (inverse short time fourier transform) algorithm.

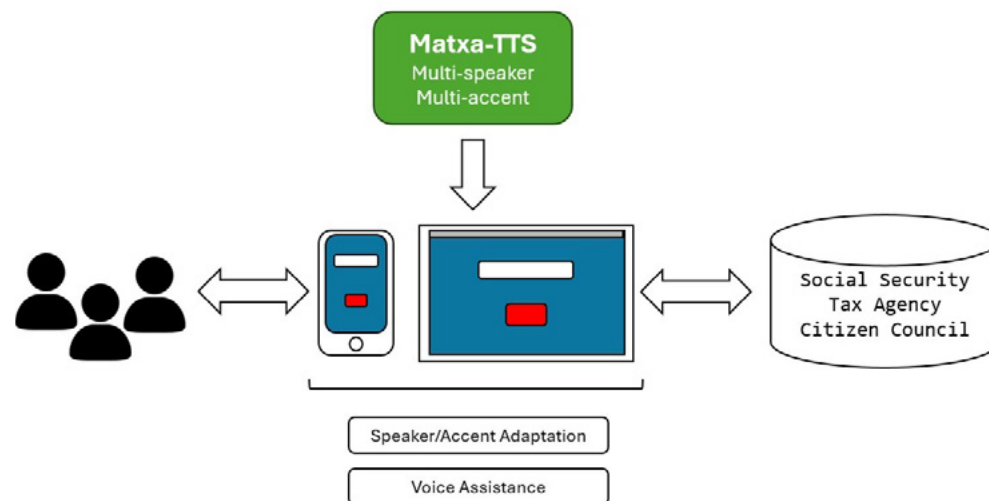
3.3. Implementation

The main advantages of these models used –Matxa and AlVoCat– are their lightness and fast response. Moreover, both models can be merged and exported to the ONNX format in a unified single architecture. Then its integration to a C++ API or a Swift-based app is perfectly feasible. We aim to adapt the system to all platforms and devices:

1. Smartphones (Android, iOS, Windows Phone, KaiOS...)
2. Web browsers (Chrome, Mozilla, Microsoft Edge, Safari, Opera...)

The application must meet the needs of visually impaired people as well as users with full visual ability. Its interface will present only the essential action buttons, and a “search” box to look for specific information that the user wants to consult and listen to. Each possible action will be expressed by a Matxa voice when the user places their finger or mouse pointer on the corresponding button (similar to the iOS’s VoiceOver). The user will be able to change Matxa’s speaker and accent at any time.

The application would be connected to the public administration database in order to bridge users to the requested information. Default speaker and accent would be female central, but both attributes can be changed at any time.





4. Impact of the solution

Thanks to the flexibility of the model in terms of speaker and accent choice, our trained TTS system can be distributed almost throughout the following Catalan-speaking territory:

- Catalonia
- Valencia
- Balearic archipelago
- Andorra

In the future, to complete the whole official Catalan areas, we plan to extend Matxa to Aranese and Alghero. This application can have a considerable impact, since only in Catalonia around 35.000 people (in 2022) were registered as blind or partially sighted. On the other hand, the inclusion of Catalan accents can offer many people from specific localities a greater understanding of the information required. Also by offering an intuitive interface and interaction with administrative information facilitates user accessibility.

Matxa can be integrated in a more complex system that may include a speech recognition, a large language model (LLM) or a dialogue system that allows a close-to-natural human-machine interaction.



5. Acknowledgements

The development of LaFresCat dataset, and the neural network models Matxa and alvoCat has been possible thanks to the financing by the Government of Catalonia through the [Aina project](#).



6. Cited works

- [1] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, 'Matcha-TTS: A fast TTS architecture with conditional flow matching', in Proc. ICASSP, 2024.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need", in Proc. NeurIPS, 2017.
- [3] H. Siuzdak, "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis," arXiv:2306.00814, 2023.
- [4] C. Donahue, J. McAuley, M. Puckette, 'Adversarial Audio Synthesis'. arXiv:1802.04208, 2018.
- [5] J. Kong, J. Kim, J. Jaekyoung Bae, 'HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis'. Advances in Neural Information Processing Systems, 33:17022-17033, 2020.
- [6] L. Zhuang, M. Hanzi, W. Chao-Yuan, C. Feichtenhofer. D. Trevor, X. Shining, 'A ConvNet for the 2020's'. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976-11986, 2022.

Socis



Membres

