State-of-the-Art Galician Speech Synthesis Using a Small-Scale Open-Source Dataset

Síntesis de voz en gallego de última generación utilizando un conjunto de datos de código abierto a pequeña escala

Antonio Moscoso-Sánchez[2], Carmen Magariños[1], Alberto Bugarín-Diz[1,2]
[1]Departamento de Electrónica e Computación
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes
Universidade de Santiago de Compostela
{antonio.moscoso.sanchez, mariadelcarmen.magarinos, alberto.bugarin.diz}@usc.gal

Abstract: Text-to-speech (TTS) synthesis has seen significant shifts with the advent of deep neural networks, yet achieving high-quality results remains a challenge for low-resource languages. This study investigates the development of state-of-the-art Galician TTS models using a small-scale, open-source dataset of approximately 4 hours. We evaluate two prominent paradigms: the end-to-end VITS architecture and the non-autoregressive Matcha-TTS. To overcome data scarcity, we explore both from-scratch training and transfer learning strategies, further augmenting the corpus through targeted synthetic data generation to address specific linguistic, phonetic, and prosodic shortcomings. The performance of the resulting models is validated through a comprehensive evaluation framework, including objective metrics (MCD, WER, UTMOSv2) and subjective listening tests. Our results demonstrate the feasibility of building high-quality, natural-sounding TTS systems for Galician even under significant data constraints.
Keywords: Speech Synthesis, Galician Language, Low-Resource Languages, Data Augmentation.

Resumen: La síntesis de voz (TTS) ha evolucionado significativamente con las redes neuronales profundas, aunque lograr alta calidad en lenguas de bajos recursos sigue siendo un reto. Este estudio investiga el desarrollo de modelos de última generación para el gallego utilizando un corpus de código abierto de solo 4 horas. Evaluamos dos paradigmas destacados: la arquitectura end-to-end VITS y el modelo no autorregresivo Matcha-TTS. Para mitigar la escasez de datos, exploramos el entrenamiento desde cero y el aprendizaje por transferencia, aumentando el corpus mediante generación de datos sintéticos para corregir deficiencias lingüísticas, fonéticas y prosódicas. La calidad se valida mediante un marco de evaluación integral que incluye métricas objetivas (MCD, WER, UTMOSv2) y pruebas subjetivas. Nuestros resultados demuestran la viabilidad de crear sistemas TTS de alta calidad y naturalidad para el gallego incluso bajo restricciones severas de datos.
Palabras clave: Síntesis de voz, Lengua Gallega, Lenguas con Recursos Limitados, Aumento de Datos.

1 Introduction

Text-to-speech synthesis (TTS), defined as the automatic conversion of text into human-like speech (Dutoit, 1997; Taylor, 2009), plays a central role in human-computer interaction. The ultimate goal of TTS systems is to produce intelligible and natural-sounding speech from the input text. Over the years, a variety of TTS methods have been proposed (Tabet and Boughazi, 2011), with concatenative unit selection (Black and Campbell, 1995; Hunt and Black, 1996) and statistical parametric synthesis (Black, Zen, and Tokuda, 2007; Zen, Tokuda, and Black, 2009) historically being among the most prevalent.

Recently, deep learning-based TTS systems have significantly advanced the field, achieving unprecedented levels of naturalness (Ning et al., 2019; Tan et al., 2021). Nonetheless, these models typically require

large single-speaker datasets (20–40 hours) to perform optimally. This high data demand is a significant bottleneck for low-resource languages, such as Galician, since obtaining the necessary training data usually involves professional speakers recording extensive specific corpora in a controlled environment. Several techniques have been explored to address this issue, such as transfer learning, multilingual training, and zero-shot learning (Casanova et al., 2022). However, their success heavily depends on data quality and the specific phonetic and morphological characteristics of the target language.

More recently, the TTS landscape has been further transformed by the emergence of generative Speech Language Models (Speech LMs) (Borsos et al., 2023; Wang et al., 2023; Anastassiou et al., 2024). By discretizing audio into quantized tokens, these systems achieve remarkable prosodic naturalness and zero-shot speaker cloning capabilities. However, deploying these massive models for low-resource languages introduces distinct bottlenecks, most notably prohibitive computational overhead for both training and inference, alongside a heavy reliance on vast multi-speaker pre-training corpora that are seldom available for languages such as Galician. Consequently, investigating more compact, data-efficient, and computationally viable architectures remains paramount for resource-constrained scenarios.

In this study, we leverage a small-scale open-source Galician TTS dataset to train and evaluate different neural voice models. Our investigation focuses on two prominent state-of-the-art paradigms: an end-to-end model (VITS (Kim, Kong, and Son, 2021)), which learns acoustic and vocoding tasks jointly, and a recent non-autoregressive architecture (Matcha-TTS (Mehta et al., 2024d)). We compare two training strategies: training from scratch and fine-tuning, allowing us to assess the effectiveness of transfer learning in low-resource scenarios. Our goal is to evaluate the feasibility of achieving high-quality synthesis under limited data constraints and to identify the most effective techniques for Galician, including the use of synthetic data generation to augment the available corpus.

The remainder of this paper is structured as follows. Section 2 situates our research within the current landscape of Galician speech synthesis by reviewing relevant related work. Section 3 delineates the experimental methodology, describing the corpora, neural architectures, and the targeted data augmentation strategies employed. Section 4 provides a comprehensive analysis of the results, discussing both objective metrics and subjective perceptual evaluations. Finally, Section 5 synthesizes the main findings and outlines potential avenues for future research.

2 Related work

Compared to widely spoken languages such as English and Spanish, the development of speech synthesis systems for the Galician language has been relatively limited, mainly due to the lack of publicly available corpora until recent years. For many years, the primary tool available was Cotovía (Rodríguez Banga et al., 2012), an open source TTS system for Galician and Spanish. As a unit selection TTS system, Cotovía boasts capabilities such as text normalisation, grapheme-to-phoneme conversion, morphosyntactic analysis, and Part-of-Speech (POS) tagging.

In 2021, the Galician Government (Xunta de Galicia) and the University of Santiago de Compostela (USC) launched the Proxecto Nós,[1] (Vladu et al., 2022; de Dios-Flores et al., 2022) which aims to generate the necessary resources to position Galician at the forefront of language technologies. This initiative is framed within broader national efforts—such as the recently concluded ILENIA[2] project (Külebi et al., 2024) and the ongoing ALIA[3] project—which seek to strengthen the digital presence of all of Spain's official languages.

Notable outcomes of the Nós initiative include two high-quality speech corpora: one female voice of approximately 25 hours (Vázquez Abuín et al., 2023; García Díaz et al., 2024) and one male voice of 18 hours (Vladu et al., 2025; Vladu et al., 2026). Furthermore, the project has released several voice models, including Sabela (Öktem, Magariños, and Moscoso, 2023), Celtia (Magariños, 2023), Icía (Moscoso, Magariños, and Bugarín-Diz, 2023), and Brais (Moscoso, 2023), all based on either the VITS (Kim, Kong, and Son, 2021) or Matcha-TTS architectures (Mehta et al., 2024a). Some of these models can be tested via the online

[1] https://nos.gal

[2] https://proyectoilenia.es

[3] https://alia.gob.es

demo Nós-TTS[4] (Magariños et al., 2024). Additionally, the Nós initiative has facilitated the public release of previously developed resources, such as the CRPIH UVigo-GL-Voices dataset, a multi-speaker speech dataset containing approximately 21 hours of audio (CRPIH and GTM, 2023).

Parallel to the Nós initiative, and also funded under the umbrella of the ILENIA project, the Aholab Signal Processing Laboratory developed the aHoTTS repository (HiTZ, 2026). This repository further expands the Galician synthesis ecosystem by providing alternative open-source TTS models trained on the same corpora mentioned above.

Historically, the only commercial synthetic voice for Galician was Nuance's Carmela (Harpo Software, 2023). This landscape shifted in 2022 when Microsoft introduced the neural voices Sabela and Roi to Azure (Microsoft, 2023). Recent expansions include Google's integration of Galician into Google Translate (Google, 2024) and NotebookLM (Google, 2025b), alongside the addition of Galician support in ElevenLabs (ElevenLabs, 2024). While these proprietary platforms have significantly expanded accessibility, they remain closed-source systems with limited flexibility for fine-tuning or specialized linguistic control. This underscores the need for open-source research and customizable models, such as those developed within the Proxecto Nós initiative, to ensure technological sovereignty for the Galician language.

3 Methodology

This section details the experimental framework designed to develop and evaluate high-quality Galician TTS models under low-resource constraints. To ensure a systematic and reproducible approach, our research pipeline is organized into four main stages. First, we describe the structural and acoustic properties of the utilized corpora, focusing on the diagnostic analysis of the target dataset. Next, we present the selected neural architectures (VITS and Matcha-TTS) and detail the training paradigms explored, comparing from-scratch learning against both cross-lingual and intra-lingual transfer learning. Central to our methodology is the targeted data augmentation process, implemented not only to mitigate data scarcity but specifically to overcome architectural limitations and correct deficiencies in the original corpus, where the speaker frequently deviated from normative oral realizations. Finally, we define the dual evaluation framework, which combines objective metrics with subjective perceptual tests to validate the performance and naturalness of the resulting models.

3.1 Datasets

3.1.1 Experimental Datasets

To evaluate TTS models under low-resource constraints and assess the impact of transfer learning, we utilize a primary target corpus alongside two auxiliary high-resource datasets for baseline pre-training:

- **Target Dataset (Icía):** The primary focus of this study is the Icía corpus from the UVigo-GL-Voices dataset (CRPIH and GTM, 2023). It comprises approximately 4 hours (2,950 sentences) of female speech recorded by an amateur speaker in a semi-professional environment. The dataset, containing newspaper extracts and manually created sentences of varying lengths (1-44 words) and types, provides audio at sampling frequencies of 16kHz/16-bit and 96kHz/24-bit. Text was normalized via Cotovia (Rodríguez Banga et al., 2012) to produce grapheme and phoneme transcriptions, with metadata following an adapted LJSpeech (Ito and Johnson, 2017) format.
- **Auxiliary Dataset 1 (Celtia):** A high-quality, 25-hour single-speaker Galician corpus recorded by a professional female speaker (Vázquez Abuín et al., 2023), featuring 20,000 phonetically, prosodically, and morphosyntactically rich sentences. Audio is provided in 48kHz/16-bit format, with a pre-trained model available on HuggingFace (Magariños, 2023). This dataset is employed as the source domain for intra-lingual fine-tuning experiments.
- **Auxiliary Dataset 2 (LJSpeech):** A widely used 24-hour English corpus containing 13,100 utterances from a single female speaker (Ito and Johnson, 2017). Audio is provided as 16-bit WAV

[4] https://tts.nos.gal

at 22.05 kHz. We include it to evaluate cross-lingual transfer learning from a high-resource language to Galician.

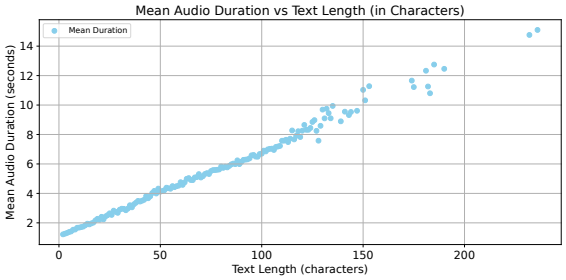
3.1.2 Target Dataset Analysis

Given the amateur nature of the target dataset, we conducted a comprehensive diagnostic analysis to assess its homogeneity and acoustic quality.

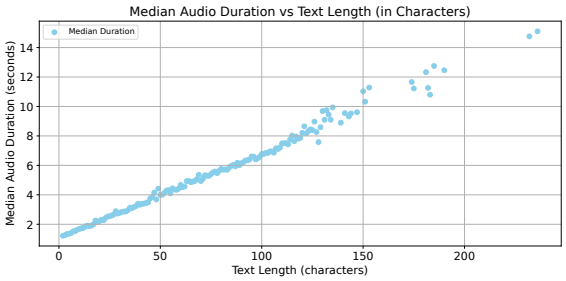
An initial textual analysis reveals a strong linear correlation between text length and both mean (Figure 1a) and median (Figure 1b) audio duration, indicating a consistent speaking rate throughout the recordings. Variability analysis (Figure 1c) demonstrates higher standard deviation for utterances between 40 and 100 characters, suggesting greater prosodic variation. Notably, the apparent stabilization of variance in texts exceeding 140 characters is primarily a statistical artifact due to the scarcity of samples at those extreme lengths. Furthermore, the text length follows a right-skewed distribution (Figure 1d), where sentences between 50 and 100 characters are the most frequent, while samples exceeding 150 characters remain rare.

Phonetically, the corpus exhibits robust coverage for a low-resource scenario. As expected for a Romance language, the distribution is heavily dominated by vowels (with /a/, /e/, and /o/ accounting for the highest frequencies). More importantly, as shown in Figure 2, all Galician phonemes are represented, with even the least frequent phoneme (/x/) appearing over 100 times. This baseline representation is crucial to ensure that the neural acoustic models learn stable embeddings for rare sounds, minimizing phonetic artifacts during synthesis.

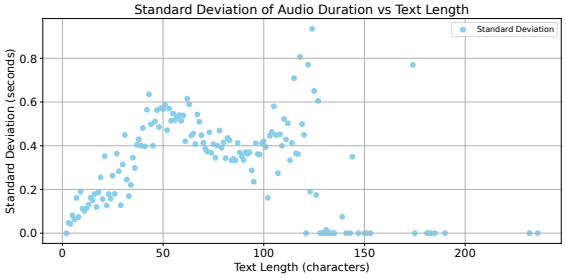
Finally, to quantify the acoustic clarity of the recordings, we estimated the Signal-to-Noise Ratio (SNR) using the WADA algorithm (Kim and Stern, 2008). Figure 3 illustrates that the SNR follows a unimodal distribution, with a median of roughly 30 dB. The vast majority of samples fall within the 20–40 dB range, indicating clean recordings with good speech intelligibility. Notably, all samples exhibit SNR values above 10 dB, comfortably exceeding the baseline for acceptable TTS training, with a long tail of high-clarity outliers surpassing 45 dB. Additionally, the isolated spike at 100 dB represents an algorithmic artifact typical of WADA when processing files containing segments of absolute digital silence.



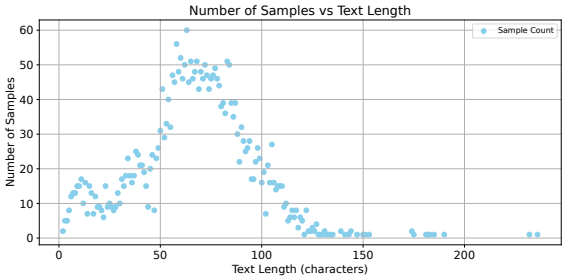
(a) Correlation between text length and mean audio duration.



(b) Correlation between text length and median audio duration.



(c) Correlation between text length and standard deviation of audio duration.



(d) Distribution of text length.

Figure 1: Text length and audio duration: distribution and correlation analysis.

3.2 Architectures

In this work, we explore two state-of-the-art paradigms for Galician speech synthesis: a fully end-to-end variational architecture and

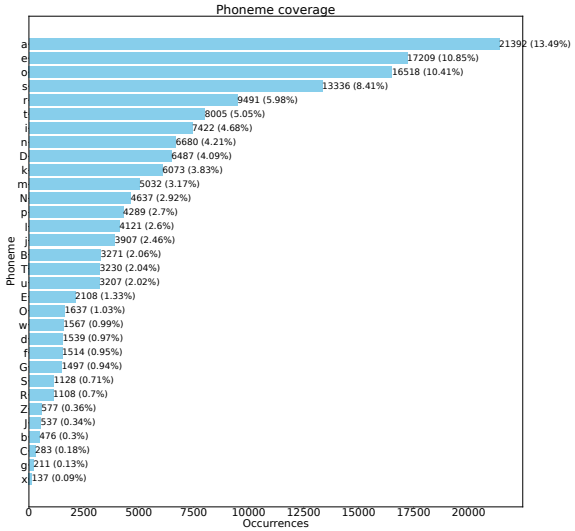


Figure 2: Phoneme coverage in the Icía dataset.

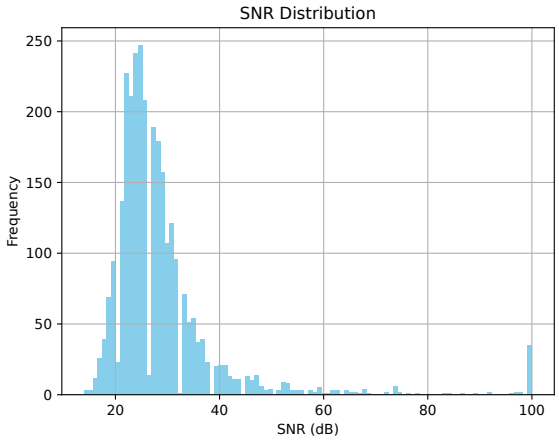


Figure 3: Signal-to-noise ratio distribution of the Icía dataset.

a non-autoregressive acoustic model paired with a neural vocoder.

VITS. VITS (Kim, Kong, and Son, 2021) is an end-to-end architecture that achieves near-human quality by integrating a GlowTTS (Kim et al., 2020) encoder and a HiFi-GAN (Kong, Kim, and Bae, 2020) vocoder into a single pipeline. By combining transformers, VAEs, and a stochastic duration predictor, it enables the synthesis of natural, expressive speech with diverse rhythms from the same input text.

Matcha-TTS. Matcha-TTS (Mehta et al., 2024a) is a non-autoregressive architecture using optimal-transport conditional flow matching to map Gaussian noise to mel-spectrograms via an ODE-based process. Utilizing a text encoder, duration predictor, and U-Net decoder, it produces high-quality speech more efficiently than diffusion models when paired with a HiFi-GAN vocoder.

HiFi-GAN. Since Matcha-TTS operates strictly as an acoustic model, it requires a neural vocoder to reconstruct the final audio signal. For this purpose, we employ HiFi-GAN (Kong, Kim, and Bae, 2020), a neural vocoder that generates high-fidelity waveforms from mel-spectrograms. It uses a GAN architecture to produce realistic and high-quality audio quickly, enhancing the naturalness of synthesised speech.

3.3 Experimental Setup

To ensure reproducibility, the models were trained using standard open-source frameworks. The VITS architecture was implemented via the Coqui-TTS library (Coqui GmbH, 2023), while Matcha-TTS was trained using its official implementation (Mehta et al., 2024b) based on the Lightning-Hydra-Template (Łukasz Zalewski and others, 2023). All experiments were conducted on an NVIDIA Hopper H100 80GB GPU. Audio was processed at a sampling rate of 16 kHz, a standard practice (Eren and The Coqui TTS Team, 2023) that balances high-quality synthesis with training efficiency. The primary hyperparameters remained consistent across experiments for each architecture: both models were optimized using a learning rate of 1e-4, with the batch size set to 16 for VITS and 32 for Matcha-TTS.

3.3.1 Training Strategies

To systematically evaluate the generation of Galician speech, we explored two main training paradigms across both architectures: training from scratch and fine-tuning (transfer learning).

As a baseline for our experiments, we utilized a publicly available VITS model (Moscoso, Magariños, and Bugarín-Diz, 2023) fine-tuned on the target Icía dataset, initialized from the Celtia base model (Magariños, 2023). This baseline relies exclusively on phonemic input representations. While it provides a functional starting point, we aimed to surpass its performance by evaluating Matcha-TTS as an alternative architecture, exploring different input representations (graphemes vs. phonemes), applying cross-lingual and intra-lingual transfer learning paradigms, and implementing targeted data augmentation techniques.

For the fine-tuning strategy, we leveraged pre-trained base models to accelerate convergence and improve acoustic stability. For Matcha-TTS, we explored two base models: one trained on the English LJSpeech corpus (Mehta et al., 2024c) (to test cross-lingual transfer) and another trained on the larger Galician dataset *Celtia* (Vázquez Abuín et al., 2023; García Díaz et al., 2024) (to test intra-lingual transfer). For VITS, we only utilized a base model trained on *Celtia*, as cross-lingual transfer did not yield satisfactory performance during preliminary informal testing. Following standard practice, in the case of Matcha-TTS, the HiFi-GAN vocoder provided by the official library was used directly for inference without specific fine-tuning for Galician, as its role is limited to reconstructing the audio signal from the generated mel-spectrograms (Kong, Kim, and Bae, 2020).

3.3.2 Preliminary Evaluation

To optimize computational resources and focus the formal evaluation on the most promising approaches, an exhaustive informal evaluation was conducted to prune the initial training configurations. This assessment yielded three key findings:

- **Input representation:** Phonemic representation consistently outperformed raw graphemes in terms of overall acoustic quality and perceived naturalness during informal evaluations. Consequently, phonemes were adopted as the standard input for all subsequent experiments, including all Matcha-TTS models.
- **Training paradigm:** Fine-tuning consistently outperformed training from scratch in both naturalness and stability across both architectures.
- **Base models:** The intra-lingual fine-tuning approach (using the Celtia base model) yielded the highest overall quality during preliminary evaluations. Meanwhile, cross-lingual fine-tuning (LJSpeech base) demonstrated surprisingly robust performance for Matcha-TTS, though it was discarded for VITS due to lower synthesis quality.

3.3.3 Data Augmentation

Despite the improvements achieved through fine-tuning, informal evaluations revealed specific shortcomings across all models. These issues were attributed either to the inherent learning limitations of certain architectures or to phonetic and prosodic gaps in the original Icía corpus. To address these deficiencies, we employed a data augmentation strategy: we took text transcriptions from the male *Brais* corpus (Vladu et al., 2025) and used our most stable available model (Matcha-TTS fine-tuned on the Celtia base model, *ori_matcha_base_celtia* in Table 2) to generate synthetic sentences in Icía's voice. This text-based data augmentation expanded the training corpus while strictly preserving speaker identity.

The main linguistic and architectural phenomena addressed were:

One-word mispronunciations. While phoneme pronunciation was generally stable in multi-word utterances, VITS models exhibited significant degradation in one-word sentences (e.g., substituting initial phonemes, such as pronouncing "gabán" with [b] or "gabardina" with [d] instead of [g]). In contrast, Matcha-TTS did not present this issue. To reinforce phonetic mapping for isolated terms, 500 phonetically balanced single-word samples were synthesized. These terms were sourced from the *Brais* corpus (Vladu et al., 2025) to ensure a diverse and robust phonetic distribution.

Preposition *para* and Verbal Homographs. In Galician, the preposition *para* is typically realized as *pra* or *pa* in spoken language, frequently contracting with definite articles (e.g., *pró*, *prás*) (Real Academia Galega and Instituto da Lingua Galega, 2012). Because the speaker adhered strictly to the literal written text instead of the prescribed oral norm, and our grapheme-to-phoneme converter generates only uncontracted forms, the model could not automatically learn these speech habits. To overcome this limitation, we manually modified the phonetic transcriptions of 255 sentences from the *Brais* corpus to reflect the contracted forms prior to synthesis. This ensures that normative writing generates natural pronunciation. However, this poses a risk for the third-person singular of the verb *parar* (to stop), which is written identically but must never contract. To provide the necessary acoustic contrast and prevent overgeneralization, 5 specific text samples where

para acts as a verb were also synthesized using their standard, uncontracted phonetic representation.

Article Allomorphs (*segunda forma do artigo*). Galician grammar dictates that definite articles (*o, a, os, as*) change to (*lo, la, los, las*) following certain triggers, such as infinitives (e.g., *come-lo caldo* instead of *comer o caldo*) (Real Academia Galega and Instituto da Lingua Galega, 2012). As this was inconsistently realized in the Icía dataset, we corrected 168 instances by enforcing the allomorph pronunciation during synthesis. A common side effect of this rule in TTS is the over-generation of allomorphs when an infinitive is followed by the preposition *a* (e.g., *ir a Madrid* erroneously becoming *i-la Madrid*). To mitigate this, 12 samples containing these specific syntactic structures from the male corpus were integrated into the target dataset to serve as negative constraints.

Prosody in Tag Questions. To improve the intonation of common conversational patterns, we addressed tag questions, e.g., *Oíanse voces, non?* (You could hear voices, right?), and exclamatory emphatics, e.g., *Iso non pode ser, non!* (That can't be right!). 101 samples featuring these patterns from the *Brais* corpus were incorporated into the training set to enhance the model's prosodic variety and naturalness.

As summarized in Table 1, this strategy resulted in an expansion of the dataset from 2,950 to 3,568 utterances. This hybrid approach (synthesizing corrective audio for existing sentences and adding new synthetic samples to handle exceptions) was crucial for ensuring that the models followed normative pronunciation rules without introducing new artifacts.

Type of Utterance	Samples	Action
One-word sentences	500	Added
Para (preposition)	255	Modified
Para (verb)	5	Added
Article allomorph	168	Modified
Inf. + a (preposition)	12	Added
Tag questions / Exclam.	101	Added

Table 1: Breakdown of the synthetic data augmentation strategy applied to the target dataset. The table details the number of samples added or modified to address specific phonetic, morphological, and prosodic phenomena, expanding the dataset to 3,568 utterances.

3.3.4 Final Model Selection

Following the preliminary pruning and the synthetic expansion of the corpus, six definitive models were selected for the formal subjective and objective evaluations. These models, detailed in Table 2, were devised to isolate the impact of the architecture (VITS vs. Matcha-TTS), the base model (Celtia vs. LJSpeech), and the quality of the dataset (Original vs. Extended). All selected models utilize phoneme representations.

Model	Corpus	Arch.	Base
ext_matcha_base _celtia	Ext.	VITS	Cel.
ext_matcha_base _ljspeech		Matcha	Cel.
ext_vits_base_celtia			LJS
ori_vits_base_celtia	Ori.	VITS	Cel.
ori_matcha_base _celtia		Matcha	Cel.
ori_matcha_base _ljspeech			LJS

Table 2: Models selected for the formal subjective and objective evaluations, categorized by training corpus, architecture, and pretrained base model.

3.4 Evaluation Framework

To comprehensively assess the performance of the selected TTS models, we employed a dual evaluation strategy combining objective metrics with subjective perceptual listening tests.

3.4.1 Objective Evaluation

We evaluate the models using a shared general test set of seven unseen sentences covering diverse linguistic structures (declarative, interrogative, exclamatory, and elliptical). To ensure a clean baseline, this set excludes the specific edge cases (e.g., single-word utterances or phonetic contractions) described in Section 3.3.3. Over this dataset, we compute three automatic metrics to quantify quality, intelligibility, and naturalness: Mel-Cepstral Distortion (MCD), Word Error Rate (WER), and UTMOSv2.

MCD. Mel-Cepstral Distortion is an objective metric that measures the spectral difference between synthesized and reference speech using mel-cepstral coefficients. Lower values (in dB) indicate greater similarity to the ground truth. It is computed as shown in Eq. 1, where $c_d^{(1)}$ and $c_d^{(2)}$ are the coefficients of the reference and synthesized signals, respectively, and D is the number of coeffi-

cients. We compute MCD using the *pymcd* library (Chen, 2022), applying Dynamic Time Warping (DTW) to align the sequences and obtain the minimum distortion.

$$\text{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}\left(c_d^{(1)} - c_d^{(2)}\right)^2} \quad (1)$$

WER. To evaluate the intelligibility of the generated speech, we measure the Word Error Rate (WER) (Jelinek, 1998). We employ Google's state-of-the-art ASR system, the Chirp v2 model (Google, 2025a), to transcribe the synthesized audio. By comparing these transcriptions against the original input text, we can robustly quantify any phonetic or linguistic errors introduced by the TTS models.

UTMOSv2. UTMOSv2 (Baba et al., 2024), originally submitted as the T05 system for Track 1 of the VoiceMOS Challenge 2024 (Huang et al., 2024), is a state-of-the-art model specifically designed to predict the naturalness of high-quality synthetic speech. Trained predominantly on English datasets, such as BVCC (Cooper and Yamagishi, 2021) and SOMOS (Maniati et al., 2022), it combines spectrogram-based representations obtained via a pre-trained image feature extractor with acoustic features derived from self-supervised learning (SSL) speech models. For our evaluation, we used the official implementation (Baba et al., 2025) fine-tuned on the Blizzard Challenge 2008 dataset (Karaiskos et al., 2008).

3.4.2 Subjective Evaluation

Subjective evaluation was conducted via Mean Opinion Score (MOS) tests (Ling, Zhou, and King, 2021) on a 5-point Likert scale (1: completely unnatural, 5: completely natural) using the same 7-sentence test set described in Section 3.4.1. Synthesizing these sentences across the six models and a natural reference yielded 49 unique audio samples. To avoid listener fatigue, participants evaluated a randomized subset of 28 audios (4 per model).

4 Results and Discussion

4.1 Objective Evaluation Results

MCD. The spectral distortion results comparing the six models against the ground truth are shown in Figure 4. The lowest (best) MCD scores were achieved by the Matcha-TTS models trained on the extended Icía corpus, particularly the variant fine-tuned on the Celtia base model. While all configurations yielded satisfactory results within the ranges reported in related literature (Weiss et al., 2021), the Matcha models trained on the original dataset exhibited the highest spectral distortion.

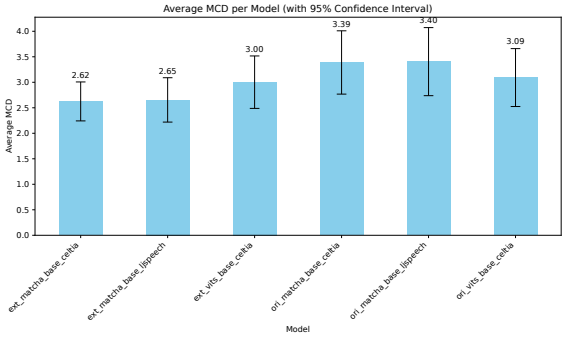


Figure 4: Objective MCD scores for the six selected models. The bar chart compares the spectral distortion against the ground truth, where lower values indicate greater acoustic similarity to the original human speech.

WER. Intelligibility across the board was remarkably high, with all systems achieving a very low WER of 1.75%, except for the Matcha-TTS model fine-tuned on LJSpeech (3.51%). Given that even the natural voice ground truth did not achieve a perfect 0% WER, this minor error rate is largely attributable to the inherent limitations of the ASR model (Google, 2025a), not to severe synthesis degradation.

UTMOSv2. The predicted MOS scores (Figure 5) indicate that the VITS model trained on the extended dataset and fine-tuned on Celtia outperformed the others, excluding the natural voice. This ranking differs slightly from the MCD results, highlighting the distinction between raw spectral similarity (MCD, lower-is-better) and predicted human perception (UTMOSv2, higher-is-better).

4.2 Subjective Evaluation Results

The MOS test was completed by 33 participants over one week, exceeding the 30-listener minimum established for reliable speech synthesis evaluation (Cooper et al., 2024). As detailed in Figure 6, the major-

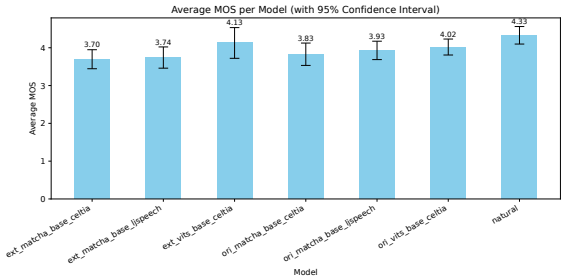


Figure 5: Objective naturalness scores predicted by UTMOSv2 for the six selected models and the natural reference. The bar chart compares the predicted MOS across the different training configurations, providing an automated estimation of human perception.

ity (25) were native Galician speakers. Most participants had prior experience in similar listening tests and used high-quality headphones, ensuring high reliability despite lacking formal linguistic training.

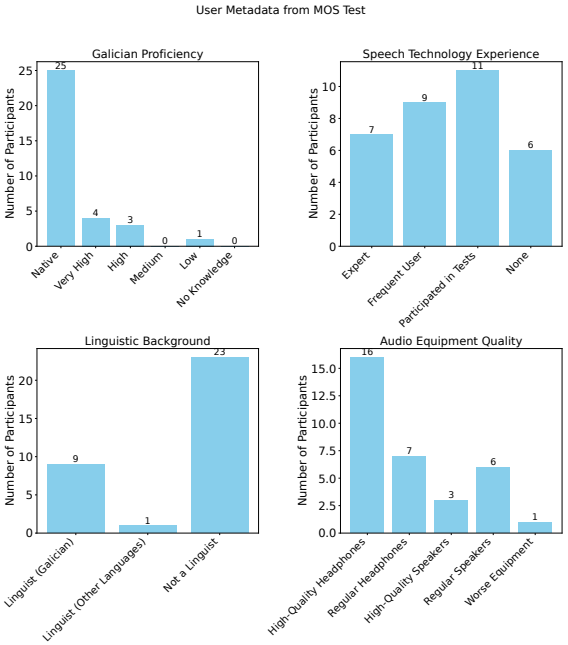


Figure 6: Demographic and technical metadata of the 33 participants in the subjective MOS evaluation. The four panels detail the evaluators' Galician language proficiency, prior experience with speech technologies, formal linguistic background, and the quality of the audio equipment used during the test.

Overall Performance. As expected, the natural voice anchor achieved the highest scores (4.23 for quality, 4.04 for naturalness). While lower than a theoretical perfect 5.0, which is common in MOS tests due to listener conservatism and recording artifacts, it provides a robust baseline (Figure 7). Among the synthetic voices, the Matcha-TTS model (Original dataset, Celtia fine-tuned) scored highest in quality, while the Matcha-TTS model (Extended dataset, LJSpeech fine-tuned) led in naturalness. Across all methods, quality scores consistently outpaced naturalness scores. Performance gaps among the top-tier models were marginal and did not reach statistical significance at the 95% confidence level, as indicated by the overlapping confidence intervals. However, these systems collectively demonstrated a substantial improvement over the baseline VITS configuration.

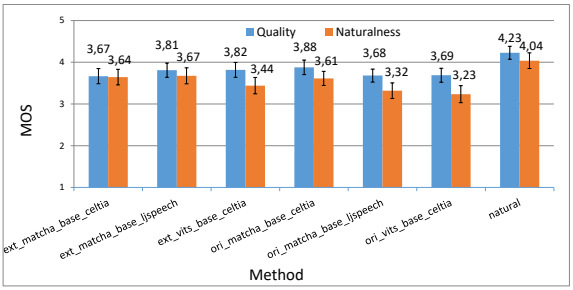


Figure 7: MOS results for quality and naturalness, including 95% confidence intervals, across the six selected models. The chart compares system performance using original and extended datasets, VITS and Matcha architectures, and different pre-training bases, with the original natural voice (*natural*) as a control.

Impact of the Extended Dataset. Comparing models trained on the original versus the extended dataset reveals a consistent trend: the augmented data reliably improved perceived naturalness across all three architectural configurations. This is likely due to the correction of article allomorphs and other linguistic phenomena. Quality scores also improved for the VITS model, though they remained relatively stable or slightly decreased for the Matcha-TTS variants.

Listener Bias Filtering. To ensure the robustness of our findings, we performed a secondary analysis (Figure 8) excluding 11 participants who rated the natural voice 2 points or lower. This filtering mitigates the impact of "harshness bias", where certain listeners predisposedly penalize speech samples. While the absolute scores across all models

increased after filtering, the relative ranking and relationships between the models remained identical, confirming the stability of our initial results.

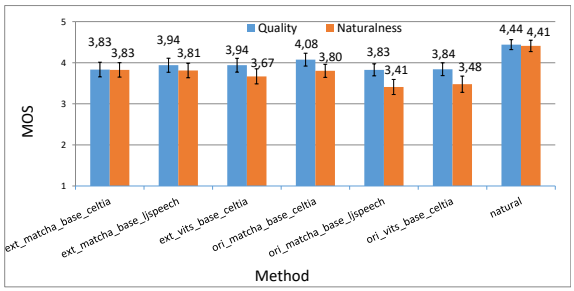


Figure 8: Filtered MOS results for quality and naturalness, including 95% confidence intervals, for the six selected models and the natural reference. This secondary analysis excludes participants who rated the natural voice 2 points or lower to mitigate harshness bias.

4.3 Qualitative Assessment

While the MOS test evaluated overall synthesis quality on generic unseen sentences, we conducted a targeted qualitative assessment to verify if the synthetic data augmentation successfully resolved the specific linguistic deficiencies identified in the original corpus.

We synthesized a separate set of test utterances heavily featuring the problematic phenomena: single-word sentences, the preposition *para*, and article allomorphs. Expert attentive listening confirmed that all three models trained on the extended corpus (VITS and Matcha-TTS fine-tuned on Celtia, and Matcha-TTS fine-tuned on LJSpeech) successfully acquired the corrected pronunciations. Notably, the phonetic mispronunciations in isolated words, which previously affected the VITS models, were completely resolved.

5 Conclusions and Future Work

In this work, we presented a comprehensive development and analysis of state-of-the-art TTS models for Galician, a low-resource language, using a small-scale dataset. A primary contribution of our study is demonstrating the effectiveness of targeted synthetic data augmentation. By extending the original corpus, we successfully corrected specific linguistic and prosodic shortcomings without compromising the naturalness or overall quality of the synthesized speech.

Regarding architectural performance, subjective evaluations revealed that both VITS and Matcha-TTS—particularly when leveraging intra-lingual transfer learning (Celtia)—achieve comparable, high-quality results. While these models collectively demonstrated a substantial improvement over the baseline, the differences among the top-performing configurations were not statistically significant at the 95% confidence level. Therefore, instead of declaring a single superior architecture, we conclude that both paradigms are highly viable under data constraints. Future work will employ more granular perceptual evaluations, such as AB preference tests or the MUSHRA framework (Union, 2015), alongside a broader set of test utterances and human-verified intelligibility scores, to capture finer distinctions between these advanced models.

Our dual evaluation framework also highlighted the current limitations of automated metrics. Although UTMOSv2 served as a robust proxy and generated scores that closely mirrored human perception, its final model ranking diverged slightly from the subjective MOS results. This underscores that, while helpful, automated predictors cannot yet fully replace human listeners in low-resource contexts. Training an objective MOS predictor fine-tuned specifically on Galician perceptual data would be an ideal solution, though this remains an open challenge due to the current lack of such datasets.

Ultimately, this study proves the feasibility of building high-fidelity TTS systems with limited open-source data. Looking forward, we plan to expand the synthetic data generation to cover a wider array of complex linguistic phenomena, such as all allomorphic variations of the Galician article, while exploring model robustness across multiple speakers, styles, and noisy environments, further advancing the state of the art for Galician speech technologies.

To ensure reproducibility and adhere to ethical open-science practices, all developed software artifacts—including training scripts, model checkpoints, and synthetic data configuration—have been made publicly available via a dedicated Hugging Face collection (Nós, 2026) under an open-source license.

Acknowledgements

This research was carried out within *Proxecto Nós*, funded by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the projects ILENIA (ref. 2022/TL22/00215337) and *Desarrollo Modelos ALIA*, and by QUANTUM_IBER_IA (ERDF/POCTEP).

The support of the Galician Ministry for Education, Universities and Professional Training and the "ERDF A way of making Europe" is also acknowledged through grants "Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04".

The work was also funded by QUANTUM_IBER_IA (ERDF/POCTEP) and by MCIU/AEI/10.13039/501100011033 (grant with reference AIA2025-163322-C62).

Finally, the authors would like to express their gratitude to all the listeners who participated in the perceptual tests for their valuable time and contributions to the evaluation of the models.

References

Anastassiou, P., J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao, M. Gong, P. Huang, Q. Huang, Z. Huang, Y. Huo, D. Jia, C. Li, F. Li, H. Li, J. Li, X. Li, X. Li, L. Liu, S. Liu, S. Liu, X. Liu, Y. Liu, Z. Liu, L. Lu, J. Pan, X. Wang, Y. Wang, Y. Wang, Z. Wei, J. Wu, C. Yao, Y. Yang, Y. Yi, J. Zhang, Q. Zhang, S. Zhang, W. Zhang, Y. Zhang, Z. Zhao, D. Zhong, and X. Zhuang. 2024. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models.

Baba, K., W. Nakata, Y. Saito, and H. Saruwatari. 2024. The T05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*.

Baba, K., W. Nakata, Y. Saito, and H. Saruwatari. 2025. UTMOSv2 Github repository. `https://github.com/sarulab-speech/UTMOSv2`. Accesed the 29th of June 2025.

Black, A. W. and N. Campbell. 1995. Optimising selection of units from speech databases for concatenative synthesis. In *Proc. 4th European Conference on Speech Communication and Technology (Eurospeech 1995)*, pages 581–584.

Black, A. W., H. Zen, and K. Tokuda. 2007. Statistical Parametric Speech Synthesis. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–1229–IV–1232.

Borsos, Z., R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour. 2023. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2523–2533, June.

Casanova, E., J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR, 17–23 Jul.

Chen, Q. 2022. pymcd GitHub repo: compute Mel Cepstral Distortion in python. `https://github.com/chenqi008/pymcd`. Accesed the 29th of June 2025.

Cooper, E., W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi. 2024. A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, advpub:e24.12.

Cooper, E. and J. Yamagishi. 2021. How do voices from past speech synthesis challenges compare today? In *11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 183–188.

Coqui GmbH. 2023. Coqui TTS manual page. `https://tts.readthedocs.io/en/latest/index.html`. Accesed the 9th of April 2026.

CRPIH and GTM. 2023. CRPIH_UVigo-GL-Voices: Galician TTS dataset. `https://doi.org/10.5281/zenodo.8027725`. Dataset. Available under CC BY 4.0 license.

de Dios-Flores, I., C. Magariños, A. I. Vladu, J. E. Ortega, J. R. Pichel, M. Garcia, P. Gamallo, E. F. Rei, A. Bugarín, M. G. González, S. Barro, and X. L. Regueira. 2022. The Nos' Project: Opening routes for the Galician language in the field of language technologies. In *Proceedings of the TDLE Workshop LREC2022*, pages 52–61, Marseille. European Language Resources Association (ELRA).

Dutoit, T. 1997. *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers.

ElevenLabs. 2024. Elevenlabs text-to-speech system. `https://elevenlabs.io/text-to-speech`. Accessed: 2026-04-08.

Eren, G. and The Coqui TTS Team. 2023. What makes a good TTS dataset. `https://github.com/coqui-ai/TTS/wiki/What-makes-a-good-TTS-dataset`. Accesed the 24th of June 2025.

García Díaz, N., M. Vázquez Abuín, C. Magariños, A. Vladu, A. Moscoso Sánchez, and E. Fernández Rei. 2024. Nos_CeltiaGL: an Open High-Quality Speech Synthesis Resource for Galician. In *Proc. IberSPEECH*.

Google. 2024. Google translate. `https://translate.google.com` [Accessed: (02/08/2024)].

Google. 2025a. Google Chirp v2 model. `https://cloud.google.com/speech-to-text/v2/docs/chirp_2-model`. Accesed the 28th of June 2025.

Google. 2025b. NotebookLM Audio. `https://blog.google/technology/google-labs/notebooklm-audio-overviews-50-languages/`. Accesed the 3th of July 2025.

Harpo Software. 2023. Carmela nuance voice. `https://harposoftware.com/en/galician/215-Carmela-Nuance-Voice.html`. Accesed the 24th of June 2025.

HiTZ. 2026. Aholab TTS Synthesis models. `https://github.com/hitz-zentroa/aHoTTS`.

Huang, W.-C., S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao. 2024. The voicemos challenge 2024: Beyond speech quality prediction. *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 803–810.

Hunt, A. J. and A. W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference*, volume 1, pages 373–376. IEEE.

Ito, K. and L. Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Jelinek, F. 1998. *Statistical methods for speech recognition*. MIT press.

Karaiskos, V., S. King, R. A. Clark, and C. Mayo. 2008. The blizzard challenge 2008. In *Proc. Blizzard Challenge Workshop*.

Kim, C. and R. M. Stern. 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Interspeech 2008*, pages 2598–2601.

Kim, J., S. Kim, J. Kong, and S. Yoon. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Kim, J., J. Kong, and J. Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Kong, J., J. Kim, and J. Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Külebi, B., I. Hernáez, E. Fernández Rei, A. Montoyo, S. Solito, C. Armentano-Oller, J. Hernando, E. Navas, C. Magariños, A. Vladu, I. Saratxaga, J. Sánchez, V. García Romillo, A. Herranz, C. Souganidis, N. García, A. Moscoso Sánchez, X. L. Regueira, F. Dubert, and Y. Gutiérrez. 2024. Speech Technologies in the ILENIA Project: Generating

Resources to Develop Voice Applications in the Official Languages of Spain. In *IberSPEECH 2024*, pages 289–292.

Ling, Z., X. Zhou, and S. King. 2021. The blizzard challenge 2021. In *In Proceedings of the Blizzard Challenge Workshop 2021*, October.

Magariños, C., A. Öktem, A. M. Sánchez, M. V. Abuín, N. G. Díaz, A. I. Vladu, E. F. Rei, and M. B. Vidal. 2024. NósTTS: a Web user interface for Galician text-to-speech. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 200–203, Santiago de Compostela, Galicia/Spain, March. Association for Computational Lingustics.

Magariños, C. 2023. Nos_TTS-gl-celtia-vits-graphemes. `https://huggingface.co/proxectonos/Nos_TTS-celtia-vits-graphemes`.

Maniati, G., A. Vioni, N. Ellinas, K. Nikitaras, K. Klapsas, J. S. Sung, G. Jho, A. Chalamandaris, and P. Tsiakoulis. 2022. SOMOS: The samsung open mos dataset for the evaluation of neural text-to-speech synthesis. In *Proc. Interspeech 2022*, pages 2388–2392.

Mehta, S., R. Tu, J. Beskow, É. Székely, and G. E. Henter. 2024a. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *Proc. ICASSP*.

Mehta, S., R. Tu, J. Beskow, É. Székely, and G. E. Henter. 2024b. Official implentation of Matcha-TTS. `https://github.com/shivammehta25/Matcha-TTS`. Accesed the 24th of June 2025.

Mehta, S., R. Tu, J. Beskow, Éva Székely, and G. E. Henter. 2024c. Matcha model trained with LJSpeech. `https://tinyurl.com/24ywuk4z`. Accesed the 24th of June 2025.

Mehta, S., R. Tu, J. Beskow, Éva Székely, and G. E. Henter. 2024d. Matcha-TTS: A fast TTS architecture with conditional flow matching.

Microsoft. 2023. Azure speech's neural TTS. `https://techcommunity.microsoft.com/t5/ai-cognitive-services-blog/azure-speech-s-neural-tts-empowers-organizations-to-serve-users/ba-p/2920882`.

Moscoso, A. 2023. Nos_TTS-gl-brais-vits-phonemes. `https://huggingface.co/proxectonos/Nos_TTS-brais-vits-phonemes`.

Moscoso, A., C. Magariños, and A. Bugarín-Diz. 2023. Nos_TTS-gl-icia-vits-phonemes. `https://huggingface.co/proxectonos/Nos_TTS-icia-vits-phonemes`.

Ning, Y., S. He, Z. Wu, C. Xing, and L.-J. Zhang. 2019. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19).

Nós, P. 2026. Nos_TTS. `https://huggingface.co/collections/proxectonos/tts-models`.

Real Academia Galega and Instituto da Lingua Galega. 2012. *Normas ortográficas e morfolóxicas do idioma galego*. Editorial Galaxia, 2 edition.

Rodríguez Banga, E., C. García-Mateo, F. Méndez-Pazó, M. González-González, and C. Magariños. 2012. Cotovía: an open source TTS for Galician and Spanish. In *VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH*, pages 308–315.

Tabet, Y. and M. Boughazi. 2011. Speech synthesis techniques. A survey. In *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, pages 67–70. IEEE.

Tan, X., T. Qin, F. K. Soong, and T.-Y. Liu. 2021. A Survey on Neural Speech Synthesis. *ArXiv*, abs/2106.15561.

Taylor, P. 2009. *Text-to-Speech Synthesis*. Cambridge University Press.

Union, I. T. 2015. Method for the subjective assessment of intermediate quality level of audio systems.

Vladu, A. I., N. García Díaz, X. L. Regueira Fernández, C. Magariños, A. Moscoso Sánchez, D. Fernández López, E. Fernández Rei, and F. Dubert-García. 2025. Nos_Brais-GL. `https://doi.org/10.5281/zenodo.14265241`. Dataset. Available under CC BY 4.0 license.

Vladu, A. I., I. de Dios-Flores, C. Magariños, J. E. Ortega, J. R. Pichel, M. Garcia, P. Gamallo, E. F. Rei, A. Bugarín, M. G. González, S. Barro, and X. L. Regueira. 2022. Proxecto Nós: Artificial intelligence at the service of the Galician language. In *SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations*, A Coruña, Spain.

Vladu, A. I., A. M. Sánchez, C. Magariños, M. P. Lago, and E. F. Rei. 2026. Nos_Brais-GL: A FAIR Galician TTS Corpus for Neural Speech Synthesis. In S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, and A. Toral, editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9841–9849, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).

Vázquez Abuín, M., N. García Díaz, A. I. Vladu, C. Magariños, A. Vidal Miguéns, and E. Fernández Rei. 2023. Nos_Celtia-GL: Galician TTS corpus, March.

Wang, C., S. Chen, Y. Wu, Z.-H. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Weiss, R. J., R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma. 2021. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5679–5683. IEEE.

Zen, H., K. Tokuda, and A. W. Black. 2009. Review: Statistical Parametric Speech Synthesis. *Speech Commun.*, 51(11):1039–1064, nov.

Öktem, A., C. Magariños, and A. Moscoso. 2023. Nos_TTS-gl-sabela-vits-phonemes. `https://huggingface.co/proxectonos/Nos\_TTS-sabela-vits-phonemes`.

Łukasz Zalewski et al. 2023. Lightning-hydra-template implementation on github. `https://github.com/ashleve/lightning-hydra-template`. Accesed the 9th of April 2026.