Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno5 n.3 (2025)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Enabling Sardinian Language Learning via Word Translation and Speech Synthesis based on Artificial Intelligence

Salvatore Mario Carta [1,2], Gianni Fenu [1], Alessandro Giuliani [1], Marco Manolo Manca [1,*], Mirko Marras [1], Alessandro Sebastian Podda [1], Livio Pompianu [1]

[1] Department of Mathematics and Computer Science, University of Cagliari; {salvatore, fenu, alessandro.giuliani, marcom.manca, mirko.marras, sebastianpodda, livio.pompianu}@unica.it
[2] VisioScientiae S.r.l., Cagliari; salvatore.carta@visioscientiae.com
* Correspondence: marcom.manca@unica.it;

Abstract: The rapid advancement of digital technologies has reshaped language education, creating opportunities for innovative teaching practices that can be extended to minority and endangered languages. Sardinian, officially identified as one of Italy's historical minority languages, presents a highly complex linguistic landscape due to its internal variations, uneven geographical distribution, and fragile vitality. Such factors led to significantly weakened intergenerational transmission, primarily due to the language's marginalization within formal schooling and the scarcity of digital learning resources. From this perspective, integrating advanced linguistic-technological tools into the classroom represents a strategic action of cultural preservation and revitalization, enabling both the safeguarding and promotion of an intangible linguistic heritage and a higher engagement of learners with their local identity. In such a context, text-to-speech technologies, initially developed for accessibility and assistive purposes, are being increasingly valued as pedagogical tools for fostering linguistic and phonological competence. Building on this potential, this work addresses the limited intergenerational transmission and the lack of modern teaching resources for Sardinian by introducing a digital educational environment based on a generative text-to-speech model. This initiative is expected to enhance learners' phonological and metalinguistic skills while supporting the revitalization and digital vitality of the Sardinian language.

Keywords: multilingual teaching; minority languages; voice technologies.

1. Introduction

The rapid development of digital technologies has significantly transformed the landscape of language education, providing innovative resources to support teaching and learning across diverse linguistic contexts. However, the current digital era presents several challenges to the preservation and promotion of linguistic and cultural diversity. Although the world hosts thousands of languages, most online content is dominated by English and often reflects American perspectives. This imbalance threatens not only cultural heritage but also economic opportunity, as artificial intelligence (AI) systems trained on limited data may underperform in less represented

languages, creating inequities in access and accuracy. This issue is also crucial in the context of educational technologies, where the main challenge lies in the limited availability of pedagogical and technological resources tailored to minority and low-resource languages, often excluded from mainstream digital innovation.

To confirm that, many studies in the literature focused on such a scenario. Existing studies highlighted that only a handful of the world's languages benefit from robust language technologies, while the majority remain unsupported (Helm et al., 2023). Furthermore, recent reports from UNESCO (2024) and analyses in the field of language technologies emphasize how this digital language gap can hamper inclusion and perpetuate inequities. The phenomenon by which computational systems preferentially support a subset of languages has been termed language inequality in AI (Joshi et al., 2020). Addressing this issue requires an interdisciplinary approach that combines computational, sociolinguistic, and pedagogical perspectives to ensure that minority languages gain a place in digital ecosystems (D'Angelo, 2024; Brookings, 2024).

In this context, bridging the language gap in AI is presented as a cultural imperative from several perspectives, including the need for innovation and digital inclusivity. The integration of advanced AI-driven tools into language education can help address this imbalance, offering learners opportunities to engage with their heritage languages in interactive and multimodal ways. Among these tools, text-to-speech (TTS) systems, originally conceived to facilitate accessibility for visually impaired users, have progressively emerged as a valuable way to foster linguistic, phonological, and metalinguistic skills. In particular, recent advances in generative AI have improved the naturalness, prosodic richness, and adaptability of synthetic voices, opening new possibilities for their integration into formal and informal educational environments.

From an educational perspective, the integration of voice-based technologies is part of a larger movement towards AI literacy (Long & Magerko, 2020) and critical media education (Rivoltella, 2017). These approaches show the importance of a collaborative learning process in which educators and learners work together to understand how technologies encourage cultural meaning, identity, and social relations. Within this framework, speech synthesis is not just a tool for pronunciation and listening practice, but a cultural interface that fosters a dynamic engagement with the sound, rhythm, and emotional dimensions of their linguistic heritage. This aligns with the concept of multimodal learning ecologies, where learning is a collaborative interplay among human actors, technological artefacts, and cultural contexts.

The present study aims to address the case of Sardinian, one of Italy's historical minority languages. Despite its cultural and linguistic richness, Sardinian faces serious challenges, including reduced intergenerational transmission, limited representation in institutional education, and a lack of modern pedagogical resources. This situation is further complicated by the internal diversity of Sardinian, which encompasses several varieties (e.g., Campidanese, Logudorese, Nuorese) and numerous localized forms.

Our work serves as an ideal benchmark for exploring the intersection of AI, language revitalization, and education. As a low-resource language, Sardinian poses challenges from a twofold viewpoint: on the one hand, from a technological perspective, the main issues stem from the scarcity of annotated data; on the other hand, from a pedagogical perspective, the teaching of Sardinian and its varieties often relies on informal or community-based practices. Developing speech technologies that re-

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

flect the internal diversity means not only preserving phonetic and prosodic authenticity but also recognizing the coexistence of multiple linguistic identities within the same community. In this sense, technology becomes an opportunity to negotiate linguistic variation rather than impose standardization.

In detail, the study proposes the development of a digital educational framework that integrates a state-of-the-art TTS model tailored to Sardinian. By combining educational research, computational linguistics, and heritage pedagogy, the tool would support promoting the use of Sardinian in primary and secondary schools and contribute to the broader digital revitalization of minority languages. The initiative aims not only to provide teachers and learners with accessible technological support but also to highlight the social, political, and cultural significance of incorporating minority languages into advanced AI infrastructures.

2. Materials and Methods

The underlying methodological framework integrates computational linguistics and educational design research, aiming to develop a digital learning environment that authentically reproduces Sardinian speech while remaining pedagogically usable in educational contexts. Figure 1 provides an overview of the entire pipeline encompassing the fine-tuning process, translation module, and graphical user interface.

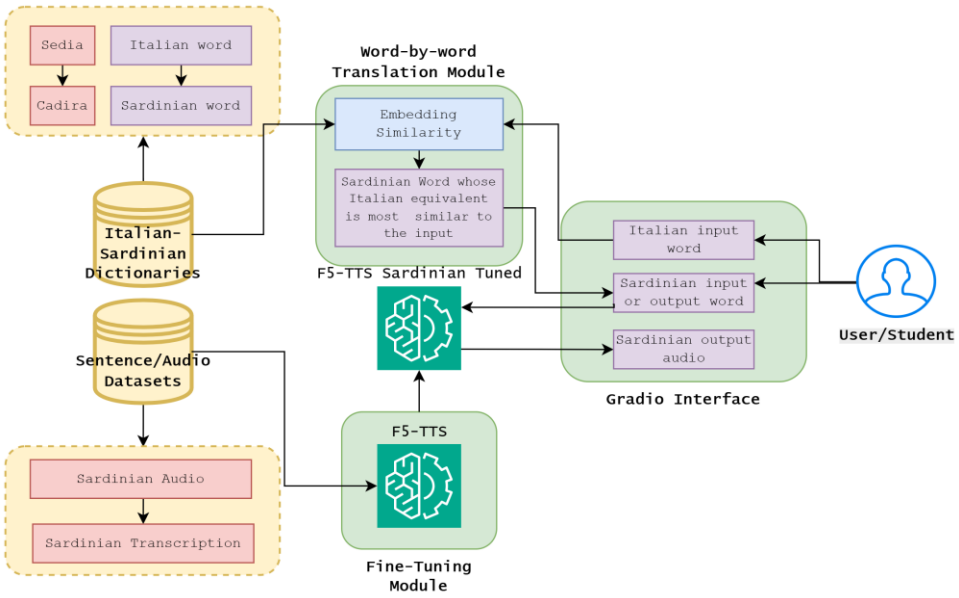

Figure 1. End-to-end pipeline: fine-tuning, translation, and graphical interface

2.1 Framework Overview

The proposed system is implemented as a web-based educational environment designed to provide accessibility, scalability, and ease of integration in classroom contexts. The prototype was developed using the Gradio framework**1**, which provides an interactive interface through a lightweight, browser-accessible architecture. At its

[1] https://www.gradio.app/

core, the framework integrates a TTS module fine-tuned for the Sardinian language varieties, e.g., Campidanese, Logudorese, and Nuorese, and a word-level translation component. The platform enables users, particularly teachers and students, to enter words in Italian or Sardinian and synthesize the corresponding speech in the selected Sardinian variety. The user interface was designed to be intuitive and pedagogically oriented, providing a minimal number of steps between input and output.

The main workflow includes:

- Text input field: users can type or paste text in Sardinian (or in Italian, when using the translation mode).
- Variety selector: a dropdown menu allows choosing among the three supported Sardinian varieties, enabling comparison between phonological and lexical realizations.
- Speech synthesis output: upon submission, the system generates a natural-sounding speech waveform using the related fine-tuned TTS model.
- Audio playback: an embedded audio player allows listening, pausing, and repeating the generated voice for training or classroom activities.

The entire interaction loop is managed through the Gradio interface, which communicates with backend inference services, ensuring that all TTS operations occur server-side and avoid the need for local computation. Figure 2 reports a screenshot of the current prototype of the user interface.

Figure 2. Screenshot of the graphical interface

2.2 Corpus Design and Preprocessing

The development of the TTS component is based on a comprehensive Sardinian speech corpus, composed of two datasets, i.e., Mannigos and SardinianVox.

In detail, Mannigos consists of approximately 5.5 hours of semi-structured interview videos, recorded between 2008 and 2010 in domestic environments. The audio data, resampled to 24 kHz and encoded in 32-bit floating-point, captures natural conversational speech in four varieties (Nuorese, Logudorese, Campidanese, and Arborense). For the present study, the Arborense variety was excluded due to data scarcity. Each transcription was manually revised with attention to dialectal features, providing valuable phonological depth despite the lack of time-aligned annotations.

SardinianVox comprises approximately 125.5 hours of high-fidelity recordings, resampled to 24 kHz and encoded in 32-bit floating-point. The dataset includes semi-structured interviews conducted in diverse acoustic conditions (indoor and outdoor) and aligned, but partially noisy, transcriptions. Unlike Mannigos, SardinianVox lacks fine-grained labeling for Sardinian varieties, which was subsequently enriched through cross-validation with the Mannigos linguistic annotations.

Both corpora exhibit unbalanced gender distributions: 41.4% female vs. 58.6% male in Mannigos, and 25.8% female vs. 74.2% male in SardinianVox. After preprocessing, (silence trimming, noise filtering, alignment correction, and normalization of orthographic variants), the following structured datasets were obtained:

- Mannigos: 1,522 <audio, transcription> pairs (779 Nuorese, 440 Logudorese, 293 Campidanese)
- SardinianVox: 26,544 pairs (2,665 Nuorese, 7,397 Logudorese, 16,482 Campidanese)

Such a heterogeneous, complementary resource base allowed us to build a robust multilingual–multivariate dataset, providing both acoustic richness and dialectal coverage for training and evaluation.

2.3 Text-to-Speech Model

The TTS module aims to generate natural-sounding, variety-specific speech from Sardinian text inputs. The goal is to synthesize fluent speech while preserving the prosodic and phonological traits of each Sardinian variety.

To this end, a state-of-the-art model was loaded and fine-tuned. In detail, we adopted F5-TTS (Chen et al., 2024), a fully non-autoregressive system based on flow-matching with the Diffusion Transformer (DiT). This architecture offers improved efficiency and expressive zero-shot generalization, outperforming previous autoregressive systems in terms of inference speed and speech naturalness. The model was fine-tuned on the preprocessed Mannigos and SardinianVox datasets, using Mel-spectrogram representations of the audio signal as training targets. During the training phase, the model learned the mapping from textual (phonemic) sequences to spectrogram features, which were later converted to a waveform through a neural vocoder.

Model evaluation combined automatic metrics, in addition to perceptual evaluation by native Sardinian speakers, who assessed intelligibility, naturalness, and fidelity

to each variety's prosodic patterns. Overall, the model demonstrated a good level of intelligibility, expressiveness, and consistency in speech variety.

2.4 Word-by-Word Translation Module

In addition to speech synthesis, the system includes a word-by-word translation module, designed to support the development of lexical and metalinguistic awareness. Such a component performs a direct mapping between Sardinian and Italian vocabularies, using curated lexicons for each variety. The system calculates the embeddings when the user enters an Italian word and selects a specific Sardinian language variety from those available. It uses cosine distance to find the Italian word most similar to the input word and responds with the word in the desired Sardinian variety.

Although conceptually simple, the module represents an essential educational aid: it encourages learners to compare orthographic, phonological, and morphological patterns across varieties, fostering a deeper understanding of linguistic structure and variation. The mapping process can be further extended or refined by integrating morphosyntactic rules and context-aware translation models in future iterations.

3. Results

The experimental phase focused exclusively on the TTS component of the framework, namely the word-level translation module, which was conceived primarily as a proof-of-concept to support linguistic accessibility. Therefore, efforts were concentrated on assessing the speech synthesis quality, intelligibility, and phonetic alignment of the TTS model, the most complex and computationally demanding part.

Specifically, the evaluation concerned the F5-TTS model fine-tuned on the two underlying Sardinian datasets (Mannigos a SardinianVox). The goal is to assess the model's ability to synthesize natural and authentic Sardinian speech while maintaining high perceptual and objective quality.

3.1. Evaluation Strategy

As an assessment, we compared each of our models against the corresponding ground truth, i.e., the set of original text and audio. In particular, we adopt a standard set of 192 utterances, randomly drawn with a fixed seed, from the corresponding test partition of each training corpus. The F5-TTS model was evaluated using multiple objective and perceptual metrics, as well as an internal language-similarity metric. In detail, the adopted metrics are briefly described below:

- DNSMOSPro: a probabilistic, lightweight DNN-based non-intrusive speech quality model (Cumlin, 2024). It estimates the overall Mean Opinion Score on a 1-5 scale (the higher the score, the better the quality). Scores are computed with the official Microsoft implementation.
- STOI (Short-Time Objective Intelligibility) (Taal et al., 2011): measures the linear correlation between short-time spectro-temporal envelopes of clean and processed signals; values span 0 (unintelligible) to 1 (perfect intelligibility).
- PESQ (Perceptual Evaluation of Speech Quality) (Rix et al., 2001): maps perceptual differences onto a MOS-like scale from 1 (poor) to 4.5 (transparent).

- SI-SDR (Scale-Invariant Signal-to-Distortion Ratio): expresses, in decibels, the energy ratio between the target waveform and the residual error after the optimal scaling (Le Roux, 2019); clean studio recordings typically exceed 25 dB.
- SardinianOrigin: An internal language-aware similarity metric that outputs the probability that an utterance is spoken in Sardinian. The score is derived from the softmax of a ResNet-152V2 model trained on 20 hours of speech (10 hours of Sardinian and 10 hours of Italian) from an internally created dataset. The Sardinian portion comprises five hours from the SardinianVox training split and five hours from the Mannigos training split. In comparison, the Italian portion consists of 10 hours from the OpenSLR Italian speech corpus[2]. The scores vary in the 0 -1 interval, where 1 denotes perfect Sardinian match.

3.2. Text-To-Speech Assessment

Table 1 reports the results of the TTS task on the selected datasets.

Table 1 – Summary of F5-TTS Evaluation Results

Dataset and F5-TTS Model	**DNSMOSPro**	**STOI**	**PESQ**	**SI-SDR (dB)**	**SardinianOrigin**
Mannigos (Original Model)	3.80	0.90	2.08	10.70	0.30
Mannigos (Fine-tuned Model)	3.68	0.89	2.14	10.76	0.43
SardinianVox (Original)	3.80	0.90	2.08	9.37	0.59
SardinianVox (Fine-tuned)	3.69	0.92	2.57	13.76	0.72

The fine-tuned models demonstrated consistent improvements in linguistic alignment and naturalness compared to the original checkpoints. On the Mannigos dataset, F5-TTS increased the SardinianOrigin score from 0.30 to 0.43, confirming that even limited adaptation can realign the model toward Sardinian phonetics. However, the relatively small dataset constrained further improvements, leaving some residual English bias in pronunciation. On SardinianVox, fine-tuning yielded substantial gains in intelligibility and overall perceptual quality, with DNSMOSPro = 3.69, STOI = 0.92, and SardinianOrigin = 0.72. These results indicate that dataset quality and recording conditions play a decisive role: higher-fidelity audio enables the model to generalize more effectively and produce speech that is both intelligible and perceptually natural. Notably, signal fidelity (SI-SDR) remained stable after fine-tuning, confirming that linguistic adaptation does not introduce significant distortions.

[2] https://www.openslr.org/94/

4. Discussion

4.1 Technical Performance and Model Adaptation

The experimental results confirm the overall robustness of the F5-TTS model and the efficacy of fine-tuning in aligning pre-trained multilingual models to low-resource languages such as Sardinian. Fine-tuning on domain-specific speech corpora resulted in a consistent improvement of the SardinianOrigin score (from 0.59 to 0.72), alongside high intelligibility (STOI > 0.90) and stable signal fidelity (SI-SDR ≈ 13.8 dB). These findings indicate that dataset quality and recording consistency are decisive for TTS adaptation: more transparent, balanced datasets enable the model to internalize phonetic traits without sacrificing acoustic realism. While F5-TTS preserves a slight English prosody bias due to its pre-training, the observed enhancement in Sardinian authenticity shows that even limited fine-tuning can bridge linguistic gaps.

A residual trade-off remains between perceptual naturalness and language fidelity, suggesting that future versions could benefit from hybrid adaptation strategies - for example, combining multilingual pre-training with phoneme-level Sardinian augmentation or using synthetic-parallel data to improve prosodic balance.

4.2 Educational Impact

Beyond technical performance, the integration of the TTS module within a web-based learning framework contributes to advancing AI literacy at multiple levels. By exposing users to a tangible example of speech synthesis grounded in their own linguistic variety, the platform fosters a deeper understanding of AI capabilities and limits, bridging the gap between abstract technological concepts and everyday linguistic experience. In educational contexts, the system acts as both a didactic aid - helping learners visualize and listen to linguistic variants - and a pedagogical tool for teaching fundamental principles of AI explainability, model bias, and data quality. In this sense, the platform not only provides linguistic functionality but also acts as a living laboratory for AI education, promoting reflective awareness about how technologies shape language perception, access, and preservation.

4.3 Cultural and Linguistic Valorization

The development of a Sardinian-capable TTS system contributes to a broader movement to revitalize minority languages digitally. The system's ability to synthesize natural-sounding speech across multiple Sardinian varieties—combined with the word-level translation module - enhances the visibility and usability of the language in digital spaces, which national or global idioms have historically dominated. By integrating Sardinian into an accessible technological interface, the framework promotes digital inclusion, enabling speakers to see and hear themselves represented within modern AI tools. Furthermore, the lexical mapping mechanism underlying the translation module provides a transparent, extensible basis for future lexicon expansion and semantic enrichment, potentially serving as a foundation for Sardinian knowledge graphs and culturally aware conversational agents.

4.4 Societal and Research Implications

The prototype demonstrates that small-scale, community-centered AI systems can have tangible impacts beyond their immediate technological scope. In practical terms, the Sardinian TTS framework offers an accessible entry point for municipalities, schools, and local institutions, supporting the creation of locally tailored digital services (e.g., audio guides, storytelling apps, or speech-based archives). From a research perspective, the results suggest promising directions for low-resource TTS adaptation pipelines, combining minimal fine-tuning data with transfer learning strategies to achieve effective localization without prohibitive costs.

5. Conclusions

This study presented a web-based educational tool integrating the F5-TTS system for the Sardinian language, demonstrating that neural TTS can generate intelligible, natural speech in a low-resource, minority language context. Fine-tuning on domain-specific corpora improved Sardinian phonetic and prosodic fidelity, confirming the feasibility of AI-driven speech synthesis as a pedagogical resource.

Beyond technical outcomes, the platform promotes language awareness, phonological training, and AI literacy, while contributing to the digital revitalization of Sardinian. The modular, user-friendly design supports potential extensions to other minority languages and participatory educational initiatives.

This work shows that voice technologies, when combined with careful pedagogical and cultural design, can function as transformative tools in language education, bridging technological innovation with linguistic, cultural, and educational sustainability. Future work will focus on school-based testing, longitudinal evaluation, and participatory refinement to consolidate the educational impact.

Acknowledgments

We also acknowledge financial support from the project ”AI4LIMBA”, funded by the Autonomous Region of Sardinia, CUP: F89B25000220002.

We also acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR).

References

Brookings. (2024). "Closing the gap: A call for more inclusive language technologies". Brookings Institution. https://www.brookings.edu/articles/closing-the-gap-a-call-for-more-inclusive-language-technologies/

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., and Chen, X. "F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching". arXiv repository arXiv:2410.06885

Cumlin, F. "DNSMOS Pro: A Reduced-Size DNN for Probabilistic MOS of Speech". In: Interspeech 2024. 2024, pp. 4818–4822.
doi: 10.21437/Interspeech.2024-478.

D'Angelo, F. (2024). "The hegemony of the English language in the digital era". inTRAlinea Vol. 27. https://www.intralinea.org/archive/article/2688

Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2023). "Diversity and language technology: How techno-linguistic bias can cause epistemic injustice". arXiv repository. https://doi.org/10.48550/arXiv.2307.13714.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). "The state and fate of linguistic diversity and inclusion in the NLP world. arXiv repository". https://doi.org/10.48550/arXiv.2004.09095

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J.R., "Sdr–half-baked or well done?" in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 626–630.

Long, D., & Magerko, B. (2020). "What Is AI Literacy? Competencies and Design Considerations". In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

Rivoltella, P.C. (2017). "Media Education. Idea, metodo, ricerca". Brescia : ELS La Scuola.

Rix, A.W., Beerends, J.G., Hollier, M.P., and Hekstra, A.P. "Perceptual evaluation of speech quality (pesq)-a new method for speech
quality assessment of telephone networks and codecs", in 2001 IEEE international conference on acoustics, speech, and signal processing.
Proceedings (Cat. No. 01CH37221), vol. 2. IEEE, 2001, pp. 749–752.

Taal, C.H., Hendriks, R.C., Heusdens, R., and Jensen, j. "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech", in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125-2136, Sept. 2011, doi: 10.1109/TASL.2011.2114881

UNESCO. (2024). "AI literacy and the new digital divide: A global call for action". UNESCO. https://www.unesco.org/ethics-ai/en/articles/ai-literacy-and-new-digital-divide-global-call-action