Nós-TTS: a Web User Interface for Galician Text-to-Speech

Carmen Magariños[1], Alp Öktem[2], Antonio Moscoso Sánchez[1], Marta Vázquez Abuín[1], Noelia García Díaz[1], Adina Ioana Vladu[1], Elisa Fernández Rei[1], María Baqueiro Vidal[3]
[1]Instituto da Lingua Galega (ILG), Universidade de Santiago de Compostela, Spain
[2]Col·lectivaT, Spain
[3]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain

Abstract

Speech synthesis, also known as text-to-speech (TTS), aims to generate human-like speech from text. The recent emergence of end-to-end deep learning TTS models has led to impressive natural-sounding results. Nevertheless, expanding these models to multiple languages and speakers poses challenges, especially for low- or limited-resource languages. In this context, we introduce Nós-TTS, a user-friendly web interface for Galician TTS developed under the Nós project. The proposed interface offers a choice among three distinct voices trained on diverse conditions regarding data quantity, training approach, and input modality. Although in an experimental stage, informal listening tests have shown a satisfactory performance of the models.

1 Introduction

Speech synthesis, also referred to as text-to-speech (TTS), is the automated generation of human-like speech by machines or computers (Dutoit, 1997; Taylor, 2009). More specifically, TTS techniques aim to synthesize intelligible and natural speech from input text. Over the years, different TTS approaches have been proposed (Tabet and Boughazi, 2011), the most prominent being concatenative unit-selection (Black and Campbell, 1995; Hunt and Black, 1996) and statistical parametric synthesis (Black et al., 2007; Zen et al., 2009).

In recent years, deep learning-based TTS systems have emerged as a powerful alternative to traditional synthesis (Ning et al., 2019; Tan et al., 2021). These systems use deep neural networks (DNNs) as the model backbone and have shown the ability to produce high-quality natural-sounding speech. However, these models often rely on massive single-speaker datasets (20-40 hours) for optimal performance. This high data demand poses a significant drawback, particularly for languages with limited resources, such as Galician, since acquiring this data can be costly and time-consuming. While various techniques like transfer learning, multilingual training, or zero-shot learning have been applied to alleviate this problem (Casanova et al., 2022), their effectiveness still depends on factors such as the data quality, quantity, and the specific traits of the target languages.

Within the framework of the Nós project (Vladu et al., 2022; de Dios-Flores et al., 2022), we present Nós-TTS, a user-friendly web interface for Galician text-to-speech conversion. Nós-TTS allows users to input a text in Galician and synthesize the corresponding speech using one of three distinct voices: Celtia, Sabela, or Icía. The underlying voice models are built on Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) architecture (Kim et al., 2021) and were trained using the Coqui-TTS library (Eren and The Coqui TTS Team, 2021) on different Galician TTS datasets.

The following sections briefly describe the proposed system, discuss the evaluation strategy and main findings, and summarize future work.

2 System Description

Nós-TTS[1] is a web user interface for speech synthesis in Galician. As shown in Figure 1, the proposed interface takes an input text in Galician (up to 1000 characters) and generates the corresponding synthesized speech by clicking the "Xerar voz" button. Once available, the audio will automatically start playing, and audio controls will be shown, including play, pause, volume, and position bar. Depending on the browser, the playback speed may also be adjusted. Other available features are download the synthesized audio (down arrow button) and clear the input text ("Borrar texto" button).

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

2.1 Voice models

In its current version, the TTS system integrates three voice models with different characteristics:

- **Celtia.** Female voice model trained from scratch on a subset of the Nos_Celtia-GL corpus (Vázquez Abuín et al., 2023). This corpus, created under the Nós project, comprises a total of 20.000 sentences recorded by a professional voice talent. Specifically, a subset of 13,000 sentences were used to train the model, which corresponds to 15.5 hours of speech. The Celtia model (Magariños, 2023) was trained directly on grapheme inputs and includes a text normalization step based on the front-end of Cotovía (Rodríguez Banga et al., 2012).

- **Sabela.** Female voice model trained from scratch on the Sabela corpus within the CRPIH UVigo-GL-Voices dataset (CRPIH and GTM, 2023). This corpus comprises 10,000 sentences recorded by a professional radio broadcaster, amounting to approximately 14 hours of speech. The Sabela model (Öktem et al., 2023) was trained on phonemes and incorporates the Cotovía front-end for text normalization and grapheme-to-phoneme conversion.

- **Icía.** Female voice model fine-tuned from the previously described Celtia model using the Icía corpus within the CRPIH UVigo-GL-Voices dataset (CRPIH and GTM, 2023). The Icía corpus comprises around 3,000 sentences, equivalent to approximately 4 hours of speech, recorded by an amateur speaker. Icía (Moscoso et al., 2023) is a phoneme-based model which integrates the front-end of Cotovía for both text normalization and phonetic transcription.

All the models are openly available in Hugging Face[2].

2.2 Models' Architecture

The incorporated voice models are based on VITS (Kim et al., 2021), a fully end-to-end TTS model that leverages cutting-edge deep-learning techniques like adversarial learning (Goodfellow et al., 2014), normalizing flows (Rezende and Mohamed, 2015), variational auto-encoders (Kingma and Welling, 2014) and transformers (Vaswani et al., 2017) to achieve results comparable to ground truth. Its architecture combines the Glow-TTS encoder (Kim et al., 2020) and HiFi-GAN vocoder (Kong et al., 2020) within the same training pipeline. By jointly learning the acoustic model and the vocoder, VITS overcomes some issues of the two-stage models. It also incorporates a stochastic duration predictor that allows synthesizing speech with different rhythms from the same input text.

2.3 Text Processing Module

The proposed TTS system includes a text processing module based on the Cotovía front-end. Depending on the selected voice, as described in Section 2.1, the text processing module performs one or both of the following functions: (1) text normalization; (2) phonetic transcription with stress marks.

Figure 1: View of the Nós-TTS user interface.

3 Evaluation and Discussion

Traditionally, the quality of TTS systems is assessed through perceptual listening tests with human subjects. These tests commonly employ perceptual metrics, such as Mean Opinion Score (MOS) (Ling et al., 2021), to rate speech characteristics including overall quality, naturalness, or similarity to the target voice.

This form of subjective measures are the gold standard for the speech synthesis task, yet it proves to be time-consuming and demanding in terms of test preparation and listener recruitment. Consequently, models are typically initially assessed through informal listening tests, with more extensive formal evaluations reserved for final models.

While the models currently integrated into the Nós-TTS interface are experimental, they have demonstrated competent performance in informal

[2] https://huggingface.co/proxectonos

listening tests, showcasing high naturalness and quality. For each voice, the models exhibiting superior performance in these informal evaluations will undergo subsequent formal evaluations involving a statistically significant number of listeners.

Nevertheless, these informal evaluations reveal insightful findings regarding the performance of the three voice models. Notably, the Celtia model stands out as the undisputed leader in terms of overall quality. Its superior results in audio quality, choice of pronunciation, precision of phonemes and naturalness of prosody, position it as the most robust and satisfactory option. This outstanding performance is directly attributed to the quality of the corpus used in its training, which was meticulously designed and developed to ensure balanced and representative textual content, voice talent with good vocal characteristics, and high-quality recordings.

Second in the ranking, the Icía model is positioned as a solid alternative, despite the limited amount of data and the speaker being non-professional. In this case, the applied fine-tuning techniques have mitigated the data limitations, resulting in a synthetic voice with noticeably more natural prosody compared to the Sabela model.

On the other hand, the Sabela model faces significant challenges, primarily related to the lack of naturalness in prosody. This limitation is evident both in the original recordings used for training and in the generated synthetic voice. The main reason for this lack of naturalness seems to be the particular style and rigid prosody associated with typical news readings on radio and television used during the recordings. This finding underscores the importance of considering not only the quantity but also the quality and diversity of training data to achieve optimal results in speech synthesis.

Another important consideration when comparing the different models pertains to the input modality for training, namely graphemes versus phonemes. A model trained on phonemes is expected to converge more rapidly, and using phonemes as input is also anticipated to aid in disambiguating the pronunciation of specific graphemes. An example of this is the grapheme <x>, which in Galician can be pronounced as [ks] or [ʃ] depending on the word. In this particular case, we have observed that the Celtia model, trained on graphemes, mispronounces this grapheme in some words (e.g., <x> as [ʃ] instead of [ks] in *boxeo* and *axila*), whereas the Icía and Sabela models, trained on phonemes, correctly differentiate between the two pronunciations. We aim to address this minor drawback of the Celtia model by training a new model based on phonemes. This final comparison reveals the importance of having a proficient text processing module to achieve precise phonetic transcriptions.

4 Future Work

The proposed system is in a continuous improvement stage, with ongoing efforts to perfect the quality of the models and expand the voice catalog. Future work will involve testing new architectures for the existing voices and training new models with additional speakers, including male voices. As mentioned in Section 3, formal evaluations will be conducted on models achieving the best-perceived performance in informal listening tests. We also plan to improve the text processing module by implementing changes in the Cotovía front-end.

Acknowledgements

This research was funded by "The Nós project: Galician in the society and economy of Artificial Intelligence", resulting from the agreement 2021-CP080 between the Xunta de Galicia and the University of Santiago de Compostela, and thanks to the Investigo program, within the National Recovery, Transformation and Resilience Plan, within the framework of the European Recovery Fund (NextGenerationEU).

References

Alan W. Black and Nick 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.

Alan W Black, Heiga Zen, and Keiichi 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.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR.

CRPIH and GTM. 2023. CRPIH_UVigo-GL-Voices: Galician TTS dataset.

Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, Jose Ramom Pichel, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín, Manuel González González, Senén Barro, and Xose Luís 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).

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

Gölge Eren and The Coqui TTS Team. 2021. Coqui TTS.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Andrew J. Hunt and Alan 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.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh 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.

Jaehyeon Kim, Jungil Kong, and Juhee 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.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung 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.

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

Carmen Magariños. 2023. Nos_TTS-gl-celtia-vits-graphemes.

Antonio Moscoso, Carmen Magariños, and Alberto Bugarín-Diz. 2023. Nos_TTS-gl-icia-vits-phonemes.

Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. 2019. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19).

Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France. PMLR.

Eduardo Rodríguez Banga, Carmen García-Mateo, Francisco Méndez-Pazó, Manuel González-González, and Carmen 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.

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

Xu Tan, Tao Qin, Frank K. Soong, and Tie-Yan Liu. 2021. A Survey on Neural Speech Synthesis. *ArXiv*, abs/2106.15561.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Adina Ioana Vladu, Iria de Dios-Flores, Carmen Magariños, John E. Ortega, José Ramom Pichel, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín, Manuel González González, Senén Barro, and Xosé Luis 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.

Marta Vázquez Abuín, Noelia García Díaz, Adina Ioana Vladu, Carmen Magariños, Adrián Vidal Miguéns, and Elisa Fernández Rei. 2023. Nos_Celtia-GL: Galician TTS corpus.

Heiga Zen, Keiichi Tokuda, and Alan W. Black. 2009. Review: Statistical Parametric Speech Synthesis. *Speech Commun.*, 51(11):1039–1064.

Alp Öktem, Carmen Magariños, and Antonio Moscoso. 2023. Nos_TTS-gl-sabela-vits-phonemes.