IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Voices from the South: Study and Synthesis of Andalusian Accents with Artificial Intelligence

Jose A. Gonzalez-Lopez[1], *Antonio M. Castilla Rubia*[1], *Alfredo Herrero de Haro*[2], *Angel M. Gomez*[1], *Antonio M. Peinado*[1]

[1]Dept. of Signal Theory, Telematics and Communcations, University of Granada, Spain
[2]Dept. of Spanish Language, University of Granada, Spain
{joseangl, antoniomcr98, ahh, amgg, amp}@ugr.es

Abstract

The Andalusian accent is a key element of the regions cultural heritage and a topic of significant interest for linguists, especially regarding language variation and linguistic typology. This paper introduces "Voices from the South", a project which aims to develop advanced text-to-speech (TTS) voices in a variety of Andalusian accents. The tool is intended to enhance the reading of texts in local accents and assist speech-impaired individuals by preserving their personal Andalusian accent. Additionally, this project can serve as a foundation for similar developments in other varieties of Spanish, with all tools and data made freely available online at the end of the project. The linguistic characteristics for each accent to be synthesised will come from the findings of "Atlas Lingüístico Interactivo de los Acentos de Andalucía", an ongoing project which aims to map accent variation across 500 data points in Andalusia and which is due to be completed in December 2026.

Index Terms: Text-to-speech, Andalusian accents, phonetics, dialectology, speech synthesis, deep learning, embeddings

1. Introduction

This paper presents an ongoing project called "Voices from the South: Study and Synthesis of Andalusian Accents with Artificial Intelligence" (VoSurAI), which focuses on a central aspect of Andalusia's society and culture: the Andalusian accent. Andalusia is a prime laboratory for phonetic research since there is great diversity of accents in a relatively small area. Alvar et al. provided a foundational study in [1], detailing accent variations across 230 towns in Andalusia, which remains the most comprehensive description of Andalusian accent to date. Some of the linguistic phenomena that occur in Andalusia follow universal patterns of language phonetic evolution, such as the omission of consonants at the end of syllables or the weakening of consonants (e.g. /s/ [h]). However, other phenomena, such as vowel lowering [2], do not follow the expected patterns of language evolution. Furthermore, the same phenomenon, such as the omission of consonants in coda, surfaces differently in different parts of Andalusia, such as the aforementioned vowel lowering in Eastern Andalusia or the lenition of /s/ in [h] in Western Andalusia [3].

In a parallel project, "Atlas Lingüístico Interactivo de los Acentos de Andalucía" [1], we are working to expand upon the findings of Alvar et al. [1], with the goal of creating an interactive linguistic atlas that will document accent variations across 500 data points throughout Andalusia. While that project focuses on providing a comprehensive linguistic analysis of Andalusian accents, the current project, VoSurAI, will leverage the data from this linguistic atlas to develop natural-sounding artificial voices representing various Andalusian accents, thereby building on the linguistic insights of the atlas.

VoSurAI will leverage recent advances in generative artificial intelligence (AI) to develop resources and tools that facilitate communication and learning in Spanish. In particular, our project aims to develop state-of-the-art, artificial voices for text-to-speech (TTS) software. We will develop a web-based tool capable of generating synthetic speech that accurately reflects the various accents of Andalusia. To the best of our knowledge, there is currently no such tool available. Offering this tool freely online will bring numerous benefits across various sectors. In entertainment, synthetic voices can provide more authentic experiences in animations, video games, and dubbed films. Media outlets could use them to foster stronger connections with local audiences, and GPS systems could offer directions in regional accents. Additionally, individuals with speech impairments using Augmentative and Alternative Communication (AAC) devices [4] to communicate would benefit from the use of personalized TTS voices that reflect their identity [5, 6].

Despite all potential benefits, various challenges have precluded the widespread development of language tools for the various Andalusian accents. One key issue is the phonetic complexity of these accents, which differs significantly from conservative accents (e.g., those of north-central Peninsular Spanish, where consonants are maintained syllable-finally). For example, while the consonant /s/ is preserved in all contexts in conservative accents, its omission in Andalusian accents triggers a range of phonetic processes that make voice synthesis harder. In Eastern Andalusia, the loss of /s/ at the end of syllables leads to vowel opening, which may extend across the word until the prosodic accent, as well as consonant lengthening [7]. In Western Andalusia, /s/ omission causes consonant lengthening accompanied by aspiration before and after adjacent consonants. Another obstacle has been the lack of accessible tools and resources for designing language and speech technologies specific to Andalusian accents. Only recently, with the rise of Python and ecosystem of deep learning libraries, have these tools become more readily available to the public.

In summary, our proposal VoSurAI delves deeper into the study of Andalusian accents and leverages cutting-edge generative AI techniques to create synthetic voices that encapsulate the regions linguistic richness. This project contributes to the fields of linguistics and speech synthesis and it addresses a societal need by providing more inclusive and representative voice synthesis tools.

The remainder of this paper is organized as follows. In Section 2, we outline the primary objectives of this project. Section 3 details the methodology employed to achieve these objectives, including the experimental design, participant selection, task

[1]https://www.acentosandaluces.com/

10.21437/IberSPEECH.2024-59

descriptions, and the signal processing and machine learning techniques to be implemented. Section 4 provides a summary of the current progress. Finally, Section 5 presents the key conclusions drawn from this work.

2. Project objectives

The specific objectives of the VoSurAI project are as follows:

1. **Extend the linguistic analysis of Andalusian accents** beyond the features typically examined in traditional linguistic studies. In addition to common aspects, such as phonetic and phonological traits, this project will explore prosodic features, fine-grained spectral details, accent embeddings, and other relevant characteristics. These elements are essential for creating natural-sounding artificial voices for TTS systems.
2. **Develop a comprehensive speech corpus** consisting of audio recordings representing a wide variety of Andalusian accents. This corpus will be made freely available to the scientific community, fostering research in phonetics and language technology specific to Andalusia.
3. **Design and implement a TTS synthesiser** capable of generating voices in or with Andalusian accents. The development of this tool will rely on the speech corpus created in Objective 2.
4. **Create an interactive web platform** where users can interact with the synthesiser developed in Objective 3. The platform will allow users to input sentences and choose their preferred Andalusian regional accent for the generated audio output.
5. **Leverage the results of this project to secure external funding** for developing a synthesiser capable of handling a broader range of Spanish accents.

3. Method

The proposed methodology for this project is structured into six key stages. The first stage involves collecting high-quality voice recordings from participants across various regions of Andalusia. These recordings will include a list of words and texts to capture all Andalusian Spanish allophones in diverse linguistic contexts, with each participant contributing between 1 to 2 hours of audio data. In the second stage, the collected audio will undergo filtering and labelling to enhance clarity and prepare the data for analysis, with noise reduction techniques applied and faulty recordings discarded. The third stage focuses on developing a TTS system for neutral Spanish, using state-of-the-art phoneme-to-spectrogram and neural vocoder models. This model will then be adapted to the distinct Andalusian accents in the fourth stage, incorporating both phonetic and prosodic features to ensure the synthesised voices reflect natural regional variations. To this, we will investigate several approaches, such as fine tuning the neutral Spanish models or adaptation by means of accent embeddings. Finally, we will develop a web-based interface to allow the general public to access and interact with the synthesised voices, providing a user-friendly platform to select and generate speech in different Andalusian accents.

Further details on each of these stages are provided in the following sections.

3.1. Participants

Participants will be recruited from each province of Andalusia. Special care will be taken to ensure that participants have a "pure" accent from their province. Additionally, people who previously participated in our ongoing project "Atlas Lingüístico Interactivo de los Acentos de Andalucía" will be invited to participate in the VoSurAI project. These participants were originally recruited through a crowd-sourced campaign that invited people from various parts of Andalusia to complete a web-based speech survey with 140 questions designed to capture different linguistic phenomena and characterise accent variation across the region. As of the writing of this paper, 1,045 participants from over 500 towns in Andalusia have contributed to the "Atlas Lingüístico Interactivo de los Acentos de Andalucía" project.

3.2. Speech tasks

We have designed a comprehensive set of speech production tasks aimed at capturing a wide range of linguistic phenomena to characterise accent variation across Andalusia. These tasks include the production of isolated words, phonetically balanced sentences, and spontaneous speech, as detailed below.

Isolated words: A 262-word questionnaire will be adapted from our previous project, "Atlas Lingüístico Interactivo de los Acentos de Andalucía", to elicit Andalusian consonants and vowels across various phonetic contexts. This includes 143 isolated words carefully selected to encompass all consonants and vowels in Spanish, presented in diverse contexts (e.g., /t/ in word-initial and word-medial positions). Additionally, the questionnaire addresses all phonetic phenomena identified in the literature regarding Andalusian accents. For instance, the word "casas" will be used to analyse word-final /s/ deletion, the lowering of /a/ preceding the deleted /-s/, vowel harmony in the first /a/ of "casas", and the phenomena of "ceceo", "seseo" or "heheo" in intervocalic /-s/. Participants will also read a modified 119-word short text, "The North Wind and the Sun," which has been adjusted to incorporate more examples of Andalusian phonetic phenomena (e.g. "abrigo" is changed to "chaqueta" to investigate the pronunciation of /tʃ/).

Sentences: Participants will also produce sentences extracted from the Sharvard Corpus [8], a phonemically balanced Spanish sentence resource consisting of 700 sentences recorded by both male and female native peninsular Spanish speakers. The corpus includes 70 lists of 10 sentences each, balanced by phoneme content. Originally designed for speech perception testing, the Sharvard Corpus has also been extensively utilised in TTS research in the Spanish language.

Spontaneous Speech: To further explore natural speech patterns, we will include two open-ended questions prompting participants to discuss their hometowns and summarise a movie or book.

3.3. Filtering

During this stage, the previously collected speech material will undergo a thorough process of filtering and labelling. Initially, recordings will be subjected to noise reduction to improve their quality. Also, low-quality or faulty audios will be discarded. Following the filtering process, the refined recordings will be semi-automatically transcribed using the Montreal Forced Aligner [9]. The generated transcriptions will be manually checked for accuracy and saved in Praat format for easy access by researchers. Additionally, each recording will be

accompanied by relevant metadata, such as participant identifiers, age, geographical information, accent, and recording conditions. The final organized corpus will be made available to the scientific community, including recordings, transcriptions, and metadata, to promote further research in the field.

3.4. Development of TTS voices for Andalusian accents

The first step towards our goal of creating diverse synthetic voices for Andalusian accents is to develop a state-of-the-art TTS synthesizer for a neutral Spanish accent. This phase is crucial because training a fully functional TTS model typically requires dozens of hours of data, which is impractical to obtain for each specific Andalusian accent. To overcome this challenge, we will first train a TTS model using the speech databases currently available to our research group. Subsequently, we will utilise the recordings collected from various regions of Andalusia to generate a set of TTS voice models, each corresponding to a distinct Andalusian accent.

For the development of the neutral Spanish TTS synthesizer, we will explore several state-of-the-art models, including FastSpeech 2 [10], VITS [11] or Matcha-TTS [12]. The training of these models will involve a variety of open-source Spanish corpora, such as Spanish LibriSpeech [13] and LibriVox Spanish [14], both of which offer a diverse range of phonetically annotated voice recordings.

Building on the foundational TTS system developed for neutral Spanish, we will explore several approaches to adapt this model for the various Andalusian accents:

- **Fine tuning**: We will refine the neutral Spanish model using recordings obtained from participants in our study to enhance its responsiveness to specific accents.
- **Accent embeddings**: We aim to utilise embeddings to model accent variations effectively. Each embedding will serve as a compact numerical representation that captures distinct phonetic and prosodic features of a spoken accent. By conditioning the TTS model on these accent embeddings during synthesis, we can enable the model to adjust its speech generation to align with the characteristics of each Andalusian accent. This technique has shown promise in previous research for modelling speech characteristics, including speaker identity and accent [15, 16].
- **Accent clustering**: To optimize the adaptation process for the diverse Andalusian accents, we will explore the use of clustering techniques, such as K-means or variational autoencoders. By grouping similar accents based on their phonetic and acoustic traits, we can create embeddings for each cluster. This strategy simplifies adaptation and enhances the TTS models robustness by reducing the risk of overfitting.
- Prosody transfer: Finally, we will also investigate prosody transfer, which involves guiding the neural network to replicate the natural rhythm, stress, and intonation of Andalusian speech by extracting prosodic features from the voice corpus [17]. This approach aims to ensure that the synthesized speech authentically reflects the essence of the accent, rather than merely imitating it.

Throughout this stage, constant evaluations will be performed, leveraging both objective metrics and subjective listening tests.

4. Work in progress

Despite the VoSurAI project officially commencing in January 2024, significant progress has already been made in this short timeframe. We are currently processing and filtering the crowdsourced audio recordings and personal interviews from our concurrent project, "Atlas Lingüístico Interactivo de los Acentos de Andalucía". These recordings will be employed in the development of the accented TTS voices. We have recently entered the third stage of the project, which focuses on developing the neutral Spanish TTS model. At this moment, we are evaluating various models and TTS frameworks for creating this system, including Coqui TTS[2], IMS-Toucan[3], and Matcha-TTS[4]. We anticipate having some speech samples ready for IberSPEECH.

5. Conclusions

The VoSurAI project represents a significant step forward in the study and synthesis of Andalusian accents through advanced TTS technology. By bulding upon the rich linguistic data gathered from the ongoing "Atlas Lingüístico Interactivo de los Acentos de Andalucía" project, VoSurAI not only aims to preserve the unique cultural heritage of the Andalusian accent but also seeks to create a practical tool that benefits a diverse range of users. This paper has outlined the projects objectives, methodology, and current progress, emphasising the novel approaches to TTS synthesis we intend to employ, such as accent embeddings, fine-tuning, and prosody transfer. As we move forward, the integration of advanced deep learning models and the continuous evaluation of synthesised speech will ensure that we meet our goal of creating natural and expressive TTS voices. Ultimately, the tools and data generated through this project will be made freely available, fostering further research in linguistic technology and potentially inspiring similar initiatives in other Spanish-speaking regions. With a projected completion date in December 2026, VoSurAI will aim at making a lasting impact on both the fields of linguistics and speech synthesis.

6. Acknowledgements

This work is part of the R&D&I project C-HUM-223-UGR23, co-financed by the Consejería de Universidad, Investigación e Innovación and by the European Union under the FEDER Andalusia 2021-2027 Program.

7. References

[1] M. Alvar, A. Llorente, and G. Salvador, *Atlas lingüístico y etnográfico de Andalucía. Tomo VI. Fonética y fonología, morfología, sintaxis*. Universidad de Granada/Consejo Superior de Investigaciones Científicas, 1973. [Online]. Available: https://www.cervantesvirtual.com/nd/ark:/59851/bmc1212717

[2] A. Herrero de Haro and J. Hajek, "Eastern andalusian spanish," *Journal of the International Phonetic Association*, vol. 52, no. 1, pp. 135–156, 2022.

[3] A. Herrero de Haro, "The vowel system of eastern andalusian spanish speakers with articulation disorders," *Lingua*, vol. 249, p. 102958, 2021.

[4] D. R. Beukelman, P. Mirenda *et al.*, *Augmentative and alternative communication. 5th ed.* Paul H. Brookes Baltimore, 2020.

[2] https://github.com/coqui-ai/TTS

[3] https://github.com/DigitalPhonetics/IMS-Toucan

[4] https://github.com/shivammehta25/Matcha-TTS

[5] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[6] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[7] A. Herrero de Haro, "Descripción acústica del andaluz oriental: estado de la cuestión," in *Sistematicidad y variación en la fonología del español, 2022, ISBN 9788492658824, págs. 213-250*. Editorial Axac, 2022, pp. 213–250.

[8] V. Aubanel, M. L. G. Lecumberri, and M. Cooke, "The sharvard corpus: A phonemically-balanced spanish sentence resource for audiology," *International Journal of Audiology*, vol. 53, no. 9, pp. 633–638, 2014.

[9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. Interspeech*, vol. 2017, 2017, pp. 498–502.

[10] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[11] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

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

[13] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," *ArXiv*, vol. abs/2012.03411, 2020.

[14] Mena, Carlos Daniel Hernández, "LibriVox Spanish," 2020. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2020S01

[15] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Proc. Advances in neural information processing systems*, vol. 31, 2018.

[16] L. Gutscher, M. Pucher, and V. Garcia, "Neural speech synthesis for austrian dialects with standard german grapheme-to-phoneme conversion and dialect embeddings," in *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, 2023, pp. 68–72.

[17] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Interspeech 2018*, 2018, pp. 3067–3071.