Voices of Latin America: A low-resource TTS system for multiple accents

Jefferson Quispe Pinares
Vozy Inc / Miami, Florida, USA
jquispe@vozy.co

Abstract

In this article, we present the implementation of a low-resource Text-to-Speech (TTS) system with accents from various regions in Latin America. We evaluate different sources of databases such as freely accessible corpora, synthetic data, crowdsourcing, and audio recordings made in a professional studio by ourselves. We adopt the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology for the development lifecycle of different models. The performance results of the TTS are represented by metrics such as Word Error Rate (WER) and Comparative Mean Opinion Score (CMOS) with interesting results like CMOS: Spanish:-1.13, Colombian:-0.34, Mexican:-0.23. WER: Spanish: 15.54, Colombian: 15.65, Mexican: 16.1. Additionally, metrics are proposed to evaluate its performance in a conversational agent.

1 Introduction

Text-to-Speech (TTS) is a voice synthesis technology that converts written text into a synthetic or digital voice that can be heard. TTS utilizes natural language processing algorithms and voice models to produce an artificial voice that sounds as natural as possible (Zevallos, 2022). This technology is used in a variety of applications, such as virtual assistants. The quality of the generated voice largely depends on the voice model used, the collected data, and the quality of the synthesis algorithm.

The project focuses on the development of a Spanish voice synthesis system with various accents from different countries that can overcome the challenges presented by this language, such as the scarcity of available resources. To achieve this, an approach is proposed that combines open data from literature, professionally recorded audio, and synthetically generated phrases by a Large Language Model (LLM).

The addressed topic is voice synthesis in Spanish, specifically the challenges it presents and the possible solutions to overcome them. The main contribution of the work lies in the combination of different types of data for the creation of a more robust and versatile Spanish TTS system. This combination allows:

Expanding the quantity and variety of data available for model training. Improving the sound quality of synthesized voice with accents from Latin American countries. Adapting the synthesized voice to different domains or specific contexts in a conversational environment.

2 Background

The development of a Spanish TTS synthesis system is a topic of great relevance for several reasons. Spanish is the second most spoken language in the world, with over 500 million speakers. A TTS system in Spanish could have a significant impact on the lives of millions of people, especially those with visual impairments or dyslexia. The development of a TTS system in Spanish could generate new business opportunities in areas such as education, customer service, and entertainment.

The development of a TTS system in Spanish presents various challenges, such as resource scarcity, dialectal variety, and sound quality. As the quality of recent TTS systems has reached a natural level, several attempts have been made to apply databases of voice synthesized by TTS to voice applications. For example, Laptev et al. (2020) and Jia et al. (2019) improved the performance of automatic voice recognition and translation systems by training models with databases of voice synthesized by Tacotron. In TTS applications, Sharma et al. (2020) demonstrated that data augmentation driven by WaveNet is effective in improving system quality Oord et al. (2018). Note that we primarily investigate the effectiveness of TTS-driven data augmentation on acoustic model performance, which was not considered in Sharma et al. (2020).

On the other hand, Ren et al. (2019) adopted the idea of using the result generated by the autoregressive model to train the non-autoregressive model. Although this and our methods commonly transfer the quality of the autoregressive model to the non-autoregressive model, there are clear differences: our method uses the autoregressive TTS model to increase the size of the training database for data augmentation purposes, while Butryna et al. (2020) present a multidialectal corpus approach for building a text-to-speech voice for a new dialect in a language with existing resources, focusing on various South American dialects of Spanish.

3 Metodology

To develop Latin accents, it was important to inventory many data sources to obtain good results, so it is distributed into 3 types of data based on their origins.

3.1 Free Access Data

Currently, many datasets have been published to create automatic voice recognition and text-to-speech conversion applications for languages and dialects from South and Southeast Asia, Africa, Europe, and South America. In this case, data in Spanish has been collected since it is considered an underrepresented linguistic community. For this purpose, databases such as Ward and Marco (2024), Pratap et al. (2020), Butryna et al. (2020) were ideal for training different accents in Spanish. The quality and diversity of the corpus are ensured with different Latin American accents.

3.2 Synthetic Data Based on LLM

In this case, synthetic data has been generated by creating text phrases using LLM and other TTS models, ensuring they are well-structured and normalized to achieve greater variability than that generated by different APIs available in the state of the art. This allows for a larger quantity across various domains, which are manually validated by a team.

3.3 Generation of Professional Audios

This is the most costly process, as after generating phrases with LLM across different domains, it allows for voice customization depending on the desired accent. Professional voice actors with various Latin American accents are hired. Phrases and sentences are recorded with high-quality audio, and the audios are transcribed or aligned with the corresponding texts. This process is crucial for inclusion within the model, as the quality of the audio ensures the desired outcome for the target accent.

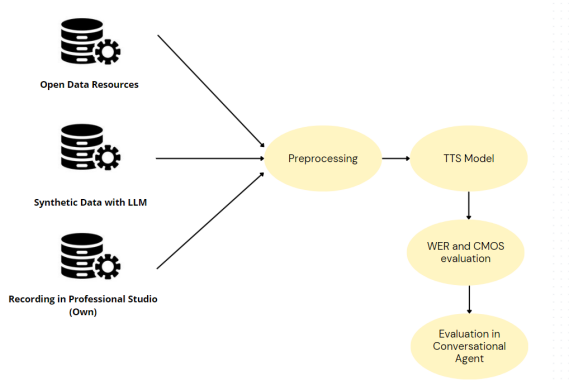

Figure 1: Data collection uses data from different sources to generate more data for a general corpus and refine it from a general to a specific accent in a given domain. Subsequently, the TTS model is selected, the parameters are adjusted and evaluated with subjective and objective metrics selected by us, to finally be evaluated in an environment with a real conversational agent.

4 Results

Evaluating a TTS system is complicated. Listener perception is key, so both subjective metrics (surveys, naturalness) and objective metrics (similarity to real voice) are used. For a comprehensive analysis, both are combined, and their performance is even evaluated in real conversations (questions and answers). In this work, WER and CMOS were initially used to measure the results of the TTS.

4.1 *Word Error Rate* (WER)

Subgeneration in synthesized speech, such as early truncation and word omission, is reflected in a higher WER (Elias et al., 2021). We using Automatic Speech Recognition (ASR) where 0 iWhere the error equivalent to zero is as accurate as possible. Since the ASR system make errors, the aforementioned metrics serve only as an upper limit of the TTS system's actual failures.

In this case, we evaluate forced alignment in synthesized speech against verbalized text and report two metrics measuring overgeneration and undergeneration. As a result, using the VITS(e2e) model yields better results compared to other models. It's worth noting that the Spanish accent has a larger amount of generated data, which is reflected in the results of Table 1.

Author	Model	Spanish Accent	Colombian Accent	Mexican Accent
Shen et al. (2018)	Tacotron2-Wavenet	28.5	19.4	18.5
Łańcucki (2021)	FastPitch-Univnet	22.75	20.45	17.5
Kim et al. (2021)	VITS(e2e)	15.54	15.65	16.1

Table 1: VITS(e2e) has the lowest WER and the best forced alignment among the evaluated models. The Spanish, having more training data, also presents better performance.

Evaluator	Spanish Accent	Colombian Accent	Mexican Accent
1st	-0.85	-0.79	-1.05
2nd	0.38	0.9	1.31
3rd	-1.8	-1.32	-0.83
4th	-0.68	0.12	0.06
5th	-2.31	-0.58	-0.53
6th	-1.5	-0.35	-0.31
Average	-1.13 ± 0.95	-0.34 ± 0.77	-0.23 ± 0.85

Table 2: The Mexican accent has the best perception by the evaluators using CMOS, while the Spanish voice has the worst. A difficulty is observed for the TTS to synthesize audios with long texts in a natural way.

4.2 *Comparative Mean Opinion Score* (CMOS)

While WER is used to measure the quality gap between generated speech and human recordings, it is not entirely reliable for assessing perception quality in TTS. Therefore, we will use subjective evaluation to measure voice quality and reinforce the evaluation.

Previous experiments used Mean Opinion Score (MOS) with 5 points (from 1 to 5) to compare generated speech with recordings. However, MOS is not sensitive enough to differences in voice quality, as the judge simply rates the quality of each phrase from the two systems without paired comparison. Hence, we chose CMOS scoring, evaluated with 7 points (from -3 to 3) as an evaluation metric where -3 indicates that the new synthesized TTS is much worse than the recorded audio, and 3 indicates that the new synthesized TTS is much better than the recorded audio, while 0 represents equivalent quality for both audios (Tan et al., 2022). Each evaluator assesses voice quality by comparing samples of recorded and synthesized audio.

As a result of the process, an arithmetic average has been established, which can be seen in Table 2. The deviation is added to the average to quantify the variation in ratings among evaluators. The Mexican accent scored the highest, followed by the Colombian accent. Finally, the Spanish voice had the worst perception by the evaluators. This is because all the audios are long texts that are difficult for the TTS to synthesize in a natural-sounding manner. This occurred with some sentences from the other voices when the text was too long. Additionally, the Spanish voice is perceived as having a less natural tone, while the other voices are close to the evaluation equivalent to zero, meaning the synthesized audios were similar to the original audios.

5 Conclusions

TTS has made significant progress in recent years, with the emergence of deep learning-based language models producing more natural and coherent voices. In this study, the VITS e2e model has been defined to synthesize voices with greater naturalness and fewer errors. A new metric called CMOS has been introduced according to the state of the art, where we obtain Colombian and Mexican voices close to the value of 0, representing a voice identical to recorded audios. Regarding the Spanish voice, it is noted with a lower rating due to the bias present among recorded audios with the character length they possess, making it sound as unnatural as possible. On the other hand, the WER metric determines that it has better results than previous TTS versions implemented with other models. This implies that the TTS service in production performs better in user conversations, providing a better user experience.

In future work, we will train the voices of the Google corpus (Peruvian, Venezuelan, Puerto Rican, Argentine, Colombian, Chilean accents) and an Ecuadorian accent recorded by us.

6 Acknowledgments

We wish to extend our heartfelt gratitude to Vozy for their invaluable support in the development of this paper. We would particularly like to acknowledge Ricardo Marin, Humberto Pertuz, Alejandro Lopez, and Helmuth Corzo, the co-founders of Vozy, for granting us the opportunity to conduct research within the company. Additionally, we extend our thanks to my AI Research Team for their consistent contributions and the Conversational User Experience (CUX) Team for their invaluable feedback. Their collaboration has been instrumental in the success of this endeavor.

References

Alena Butryna, Shan-Hui Cathy Chu, Isin Demirsahin, Alexander Gutkin, Linne Ha, Fei He, Martin Jansche, Cibu Johny, Anna Katanova, Oddur Kjartansson, et al. 2020. Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: an overview. *arXiv preprint arXiv:2010.06778*.

Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J Weiss, and Yonghui Wu. 2021. Parallel tacotron: Non-autoregressive and controllable tts. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5709–5713. IEEE.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE.

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.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. 2020. You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Manish Sharma, Tom Kenter, and Rob Clark. 2020. Strawnet: Self-training wavenet for tts in low-data regimes. In *INTERSPEECH*, pages 3550–3554.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2022. Naturalspeech: End-to-end text to speech synthesis with human-level quality. *arXiv preprint arXiv:2205.04421*.

Nigel G. Ward and Divette Marco. 2024. A collection of pragmatic-similarity judgments over spoken dialog utterances. In *Linguistic Resources and Evaluation Conference (LREC-COLING)*.

Rodolfo Zevallos. 2022. Text-to-speech data augmentation for low resource speech recognition. *arXiv preprint arXiv:2204.00291*.