An Annotation System for Controllable Speech Synthesis in Wolof

Anonymous ACL submission

Abstract

Recent advances in deep learning have enabled the creation of expressive and controllable speech synthesis models. However, the creation of such models requires the collection and annotation of large amounts of data, which limits their applicability to low-resource languages. In this paper, we propose an automatic annotation pipeline to bypass the tedious process of annotating parameters such as prosody or emotion in a text-to-speech dataset. Our system rebalances the distribution of speech features in the dataset and then uses a large language model with Gemma 2 to predict relevant annotations in the form of textual descriptions, with zero minutes of expert annotation. As most of the features extracted are language agnostic, we obtain a generic annotation procedure that we evaluate by finetuning a controllable text-to-speech model on a low-resource language, Wolof. The results show that our model acquires a greater ability to control prosody, with a gain in pitch correlation of +0.09 and a speaker similarity of 0.54. The chosen architecture also performed well on Wolof, with a perceptual quality of 3.34 and a word error rate of 0.45.

1 Introduction

Text-to-speech (TTS) represents one of the most significant advances in human-computer interaction, enabling diverse applications in human communication (Adler et al., 2006), from accessibility tools for visually impaired individuals to virtual assistants and audiobook production.

Recently, with the increasing industrial demand, TTS technologies have evolved beyond the ability to synthesize human-like speech to enable controllable speech generation. This includes fine-grained control over various attributes of synthesized speech such as emotion, prosody, timbre, and duration (Xie et al., 2024). These controllable systems typically employ hierarchical architectures that separate linguistic content from prosodic features, allowing independent manipulation of pitch, rhythm, and emotional tone while maintaining linguistic integrity (Lee et al., 2023).

Controllable TTS for low-resource languages represents a significant frontier in speech synthesis, offering communities the ability to customize speaking styles despite limited training data. Recent advances like Byambadorj et al. (2021) leverage cross-lingual knowledge transfer, where expressive representations from high-resource languages can be adapted to low-resource contexts with minimal target language samples. Other approaches use data augmentation based on voice conversion or synthetic data generation to increase the quality of the system in the target language (Huybrechts et al., 2021). Despite notable advances, the quality of these systems remains well below that of resource-intensive languages due to the lack of quality data. In addition, the control of many characteristics, such as prosody, remains imperfect due to the specific speech characteristics of each underrepresented language.

One of these low-resource languages is Wolof, an African language belonging to the Niger-Congo language family [1]. It is spoken primarly in Senegal and the Gambia, by approximately 80/100 of the inhabitants of these two countries (Omar, 1987). According to Rialland and Robert, the intonational system of Wolof has several interesting typological features, including the absence of any intonational marking of focus. Since Wolof doesn't use intonation to mark focus, the TTS system would need to integrate closely with morphosyntactic information, making it particularly challenging in a low-resource setting.

This perfectly highlights an example of a situation where relevant annotations are needed to adapt to the characteristics of a low-resource language.

[1] https://www.mustgo.com/worldlanguages/wolof/

In this study, we propose an innovative annotation strategy to accurately annotate speech data while adapting to the specificities of the Wolof language. Then, we use the resulting dataset to train and evaluate a controllable TTS model in Wolof. Our test benchmark provides a reference against which many other Wolof text-to-speech models can be evaluated using objective metrics.

2 Related Work

2.1 Controllable TTS

Controllable text-to-speech (TTS) systems require large-scale datasets that exhibit extensive diversity and include fine-grained annotations. Different types of annotation have been used to create these datasets, including annotations based on detailed textual descriptions of speech attributes (Guo et al., 2023; Jin et al., 2024; Ji et al., 2024a). Description-based datasets enable models to interpret nuanced, free-form textual prompts and generate speech that aligns with complex, context-dependent specifications (Xie et al., 2024). This is why we chose a description-based dataset to train our controllable TTS model.

Datasets based on textual descriptions allow precise control, but they are rare and costly to construct. In addition, limited diversity of styles or prosodic variations in datasets can restrict the model's ability to generalize across unseen attributes. Although there are some large-scale datasets, such as LibriTTS (Zen et al., 2019) and TextrolSpeech (Ji et al., 2024a), their diversity is still not enough for fully controllable TTS. One approach is to augment the datasets following certain attributes, while other studies attempt to automatically annotate the features present in speech samples (Lyth and King, 2024). In this study, we combine the two solutions to adapt a traditional text-to-speech dataset to the needs of a controllable model.

In terms of the models used, many recent models have relied on textual descriptions to control speech generation. Most of them use architectures based on large language models (LLMs) to support text descriptions in the form of prompts (Lyth and King, 2024; Ji et al., 2024b). However, determining the appropriate level of granularity for control and devising methods to achieve precise control at a specific granularity or to enable multiscale and fine-grained control remains a significant challenge. Here we analyse the results obtained with an LLM-based TTS model on a low-resource language, as well as future research directions.

2.2 Low-Resource TTS

A lot of research is being carried out into the application of text-to-speech to low-resource languages. Traditional techniques include self-supervised learning (Chung et al., 2019), cross-lingual transfer (Tu et al., 2019; Xu et al., 2020), and back-transformation (Tjandra et al., 2017; Ren et al., 2019). Recent models combine the power of codec models with multilingual capabilities (Zhang et al., 2023; Zhou et al., 2024). These models encode text from multiple languages into a shared latent representation space, enabling cross-lingual transfer learning and unified processing.

Despite these advances, African languages remain particularly marginalised. However, there have been some studies on the creation of datasets and text-to-speech models for African languages (Meyer et al., 2022; Ogayo et al., 2022). With regard to the control of certain parameters such as African accents, Ogun et al. have explored such multi-accent models, but in English. As for Wolof, which is the subject of this study, some text-to-speech systems exist in the form of proofs of concept (Gauthier et al., 2024), alongside more recent open source initiatives [2]. Our system uses the latest architectures based on LLMs on a Wolof dataset annotated with various speech attributes.

3 Methodology

3.1 Data Preprocessing

The dataset used in this work is a text-to-speech (TTS) dataset that contains recordings from two native Wolof actors, a male and female voice [3]. The audio samples were cut with an average duration of 3.78 seconds. A large number of operations were carried out to process this dataset. A comparison of the details of the dataset before and after preprocessing can be found in table 1. First of all, corrupted files were detected and deleted. The transcriptions were then cleaned up by removing foreign language characters, unpronounced emojis and special characters, while the numbers were converted to letters to match the reading in the audios. All these transformations to the text are designed

[2] https://huggingface.co/galsenai/xTTS-v2-wolof

[3] https://zenodo.org/records/4498861

Dataset	Number of samples	Speakers	Duration	Mean Duration	PESQ	SI-SDR
Original	40042	2	42h	3.78s	2.41	13.62
Processed	39555	2	41h 30min	3.78s	3.52	22.84

Table 1: Comparison of the details of the dataset before and after preprocessing.

to ensure better text-audio alignment for a more robust text-to-speech model (He et al., 2019).

In addition, further operations have been carried out on the audio samples. Most of the audio was noisy or poor quality, so we used an enhancer [4] to improve the quality. Table 1 shows two relevant metrics for assessing the quality of enhancement. These are the perceptual audio quality (PESQ) (Recommendation, 2001) and the scale-invariant signal-to-distortion ratio (SI-SDR) (Le Roux et al., 2019), which measures the rate of sound distortion relative to the useful signal. Significant gains are observed compared with the original data.

3.2 Annotation System

The aim of our annotation system represented at figure 1 is to automatically generate detailed textual descriptions that match the speaking style of each audio sample. These descriptions are then added to the text-audio pairs to form a controllable text-to-speech dataset. To generate the descriptions, we rely on Data-Speech (Lacombe et al., 2024), an automatic annotation method inspired by the work of Lyth and King. The system calculates a set of metrics from an audio sample, such as speaking rate, pitch, signal-to-noise ratio (SNR) and speech clarity (C50). These metrics are then used by a large language model to predict the following five attributes:

- **Speaker:** the speaker's identifier, which can be his or her name or gender;
- **Pace:** the speech rate, which can be slow, moderate speed, fast, slow pace or fast pace;
- **Tone:** the expressiveness of the speaker which can be monotone or expressive and animated;
- **Noise:** the level of unwanted sound which can be clear, noisy, good recording or poor recording;
- **Reverberation:** the level of persistence of a sound due to echo which can be close-sounding, distant-sounding, roomy sounding, moderate reverberation or confined.

The language model used in our case to make these predictions is Gemma 2 (Team et al., 2024). It predicts a natural language text description including each of these five attributes to accurately to describe the style of the audio sample.

Most of the metrics used to determine these attributes are language-independent, so we can use them directly on our data. However, the speaking rate is calculated according to the number of phonemes per second of audio, which makes it a language-dependent metric. Our contribution was therefore to adapt Data-Speech to Wolof for this metric, by creating a grapheme-to-phoneme model. Our phonetic transcription model is a multilingual version of ByT5 (Xue et al., 2022) which we fine-tuned on phonetic data in Wolof. The output of the model is remarkably accurate, which we discuss in the section on experiments.

One of the constraints with controllable speech synthesis is that there needs to be diversity of prosodic features in the dataset. To measure the diversity of these characteristics, we use a given metric for each characteristic. For pitch, we calculate the standard deviation of the pitch countour for each sample; for speaking rate, the speed in phonemes per second; and for noise and reverberation, the signal-to-noise ratio (SNR). An analysis of the distribution of these three metrics on a sample of 1,000 examples from the dataset shows that they are not balanced in relation to expressivity, speaking rate and noise level, as shown in the first part of figure 2 in the case of the speaking rate.

To rebalance these metrics in the dataset and thus have the diversity needed to learn how to control these parameters, we use signal processing techniques such as time stretching for speaking rate and pitch shifting for pitch. In the case of noise and reverberation, we do not modify the distribution to avoid negatively impacting data quality. Even in the case of pitch and speaking rate, manipulating their values can be tricky because of the risk of adding distortion to the audios produced. So to ensure that we get a quality rendering, we perform

[4] https://github.com/resemble-ai/resemble-enhance

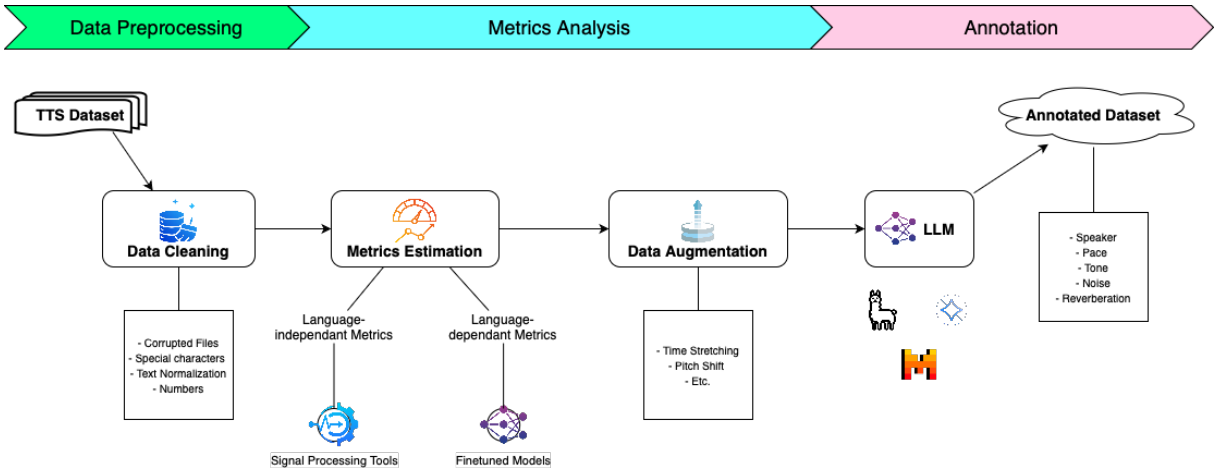

Figure 1: Schema of the annotation system.

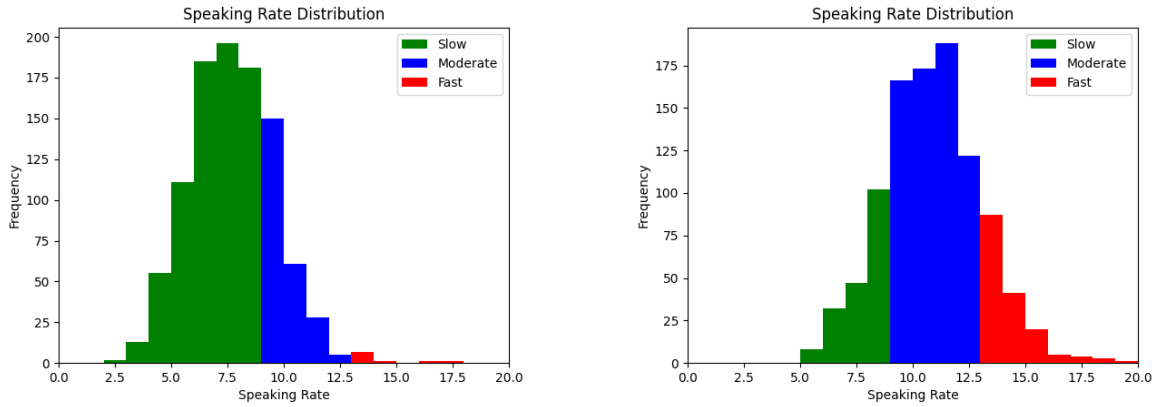

Figure 2: Comparison of the distribution of the speaking rate in the original and augmented dataset.

these operations using CLPCNet (Morrison et al., 2021), which applies machine learning techniques to signal processing to edit the speech parameters of audio samples not seen during its training. The results obtained, which can be seen in the second part of figure 2, show that we managed to significantly rebalance the pitch and speaking rate values across our dataset.

In short, we propose a methodology for annotating text-to-speech datasets in three stages: prepare the data appropriately, using an enhancement model if necessary, analyse the distribution of key speech metrics and rebalance their values across the data, and finally generate text descriptions that match those metrics using a large language model. Such a system can produce controllable synthetic datasets without expert annotation, and can be applied to a wide range of languages, as most of the metrics extracted are language agnostic.

3.3 TTS Model

Among the models based on large language models (LLMs), we have chosen Parler-TTS, which is a lightweight speech synthesis model that can generate high-quality natural speech in the style of a given speaker (gender, pitch, speaking style, etc.). It is a reproduction of the work carried by Lyth and King. This choice is justified by the model's ability to handle textual descriptions to control audio output, as well as its architecture that uses a latent representation allowing high audio fidelity. Its multilingual version we use is pre-trained on a cleaned CML-TTS dataset (Oliveira et al., 2023), which supports eight European languages.

Here we test the ability to transfer European languages to an African language by finetuning the multilingual version of Parler-TTS on our Wolof dataset. As the audio descriptions are in English, this makes it all the easier for the model to represent them semantically, for controlled generation based on those descriptions.

4 Experiments

4.1 Experimental Settings

To demonstrate the efficacy of the proposed methodology, we conducted experiments to build two different models: a grapheme-to-phoneme model and a controllable TTS using text descriptions. Both models are trained on Wolof data that we have cleaned up and improved. We describe the experimental conditions for each model below.

The **grapheme-to-phoneme model** is trained on a corpus of 1000 lines that we have collected and processed. The phonetic data is written in the International Phonetic Alphabet (IPA) (Ladefoged, 1990), which ensures uniformity with other languages. The dataset is then divided into training, validation and test sets in an 80/10/10 ratio. We then finetuned a model based on ByT5 (Xue et al., 2022) on this data with a learning rate of 3e-4 and a batch size of 16 for 30 epochs on a T4 GPU. The model is then evaluated on the test set with a batch size of 32.

The **controllable TTS model** is a multilingual version of Parler-TTS (Lyth and King, 2024) with 880M parameters that we finetuned into a multi-speaker dataset that we processed and improved. Details of the processing on the data can be found in section 3.1. The model is trained on four V100 GPUs over 100 epochs. The learning rate is set to 1e-4 while the duration of audio samples is limited to between 1 and 20 seconds during training.

4.2 Evaluation Benchmark

In order to evaluate our text-to-speech model, and provide a basis for comparison with all existing text-to-speech models in Wolof, we have set up an evaluation benchmark. It consists of a test dataset of 100 audios collected in such a way as to have a diversity of speakers (male and female), prosody and noise level. We then implemented a set of scripts to calculate objective metrics suitable for evaluating controllable text-to-speech models in Wolof.

PESQ: This is an estimate of the perceptual quality of audio, which we use here to get an idea of the naturalness of the audio produced. As there is no model for estimating perceptual quality in Wolof, we used this method inspired by telecommunications (Recommendation, 2001).

WER: This is a metric that compares the error rate between a reference text and a transcribed text. We use it here to evaluate the intelligibility of our model, which improves as the WER increases. To adapt this metric to Wolof, we use a speech recognition model trained on Wolof [5] to transcribe the audio samples generated by our model.

Speaker Similarity: It measures how closely a synthesized voice matches the characteristics of a target speaker. To do so, we compare acoustic features using audio embeddings. To calculate the embeddings of the audios, we used the speaker encoder of a version of XTTS v2 (Casanova et al., 2024) finetuned to Wolof. This allows us to better separate the linguistic characteristics specific to Wolof from the vocal characteristics of the speakers.

Pitch Correlation: It quantifies how well fundamental frequency patterns align between a synthesized speech sample and reference speech. Strong correlation suggests the synthesized speech preserves natural intonation patterns, which is crucial for expressiveness. In our case, we use the probabilistic YIN algorithm (pYIN) (Mauch and Dixon, 2014), which is an improved version of the YIN algorithm for pitch detection that reduces octave errors and handles noisy conditions better.

SNR: This is a measure of the level of desired speech signal compared to background noise. In speech synthesis, higher SNR values indicate cleaner audio with fewer artifacts. In the controllable generation, it also measures the model's ability to control the isolation and clarity of the generated signal. The model we use here to separate audio from noise and calculate the SNR is Sepformer (Subakan et al., 2021), a Transformer-based neural network for speech separation.

Altogether, these different metrics provide an objective and appropriate assessment of the naturalness, intelligibility, clarity and controllability of audio samples generated by a TTS in Wolof. Thanks to the diversity of prosody and speaker parameters present in our benchmark, we provide a reference tool for evaluating the quality, expressiveness and controllability of speech synthesis models in Wolof.

4.3 Results

The grapheme-to-phoneme model was evaluated using two metrics to assess its ability to correctly transcribe a Wolof text into the phonetic alphabet. These are the word error rate (WER), which calculates the error rate per word between the original

[5] https://huggingface.co/CAYTU/whosper-large-v2

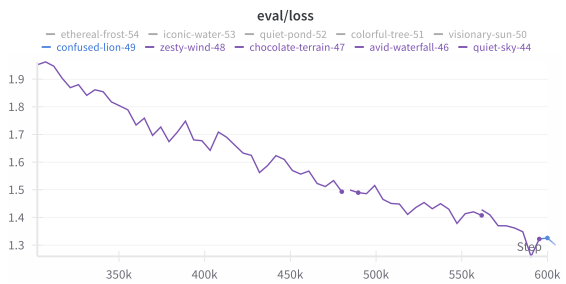

Figure 3: Validation loss descent curve

phonetic transcription and that of the model described by the formula [x], and the character error rate (CER), which measures the error rate per character between these transcriptions described by the formula [y].

The grapheme-to-phoneme model has a word error rate of 0.13 and a character error rate of 0.02. These metrics demonstrate the effectiveness of the model in transcribing phonetics in Wolof and in being used in systems such as our automatic annotator. The low character error shows that even when the model is wrong on a word, it is generally an error on few characters.

For our speech synthesis model, its finetuning on the Wolof dataset shows a stable and rapid loss descent curve, as can be seen in figure 3. The rapid decrease in the validation loss shows that the model learns quickly from the knowledge acquired about the languages used during pre-training. Each different colour in the figure represents the resumption of training from the last checkpoint, which does not affect the course of the training. The model was evaluated on the proposed benchmark using the following five metrics: perceptual quality (PESQ), word error rate (WER), speaker similarity, pitch correlation and signal-to-noise ratio (SNR). Each of these metrics has been detailed in the previous section.

All the metrics at benchmark level are perfectly adapted to Wolof data thanks to the models we used to calculate them. Only perceptual quality, which is supposed to assess naturalness, is not specifically adapted to Wolof data. We therefore reinforce this metric with a subjective metric, the mean opinion score (MOS), which is an average evaluation of the quality of the audios by a group of human experts. This is the last piece of data to be added for an exhaustive evaluation.

In addition to our model, we evaluated Galsen AI's open source model based on XTTS v2 (Casanova et al., 2024) on the benchmark. This allows us to compare our model with the existing one and to justify the choice or not of the Parler-TTS architecture according to its performance compared with XTTS v2. We have also evaluated the performance of the multilingual Parler-TTS model without finetuning in order to measure the performance gains provided by our annotated dataset. The results of these different models according to the different metrics are shown in table 2.

Analysis of the results of our model shows that it produces very natural audios with satisfactory perceptual quality, as well as good intelligibility with a WER of 0.45, which means that more than half of the words present in the speech are perfectly transcribed. The errors present are due to a few cases of mispronunciation of words, but overall the speech remains very comprehensible for the Wolof speaker, as demonstrated by the MOS of [-].

Comparing its performance with the multilingual version of Parler-TTS (Lyth and King, 2024) without finetuning shows a clear improvement on all metrics. This shows the effectiveness of our cross-lingual transfer learning approach, and just how necessary it is to adapt existing models to obtain satisfactory results on a low-resource language such as Wolof. In terms of the objective quality of the generated audio, there is a considerable improvement of +0.5, while the audios are much clearer with less noise. The increase in pitch correlation shows that the model learned to adapt better to the tonal variations specific to the Wolof language. This is an important point, as it proves that the prosodic features learned from European languages need to be adapted to better match the speech of an African language like Wolof.

As for the comparison with the Galsen AI XTTS v2 open source model, it shows that our model outperforms its counterpart on most metrics, particularly intelligibility, with a significant gap of 0.73. Another important point to note is that our model produces a voice more similar to the original voice, which is very important for a controllable model to be able to control the speaker. All this proves the advantages of the Parler-TTS architecture over XTTS v2 in this study, since both models were trained on the same data. The metrics on which the Galsen AI's model outperforms ours are mainly explained by the fact that our model controls the output voice quality parameters, which it deliberately lowers in certain audios where the description requires noise to be added.

In short, the subjective and objective results of our evaluation show the superiority of our speech synthesis model over existing models and support the relevance of our approach. The metrics obtained show that the model does indeed manage to control the speaker with a high degree of similarity, to control rhythm and tone with a higher pitch correlation, and to control noise and reverberation by lowering the audio quality where necessary.

5 Conclusion

In this work, we have built an annotation pipeline that constructs controllable text-to-speech datasets without the need for expert annotation. We applied it to a low-resource language dataset to create the first controllable text-to-speech system in Wolof. This is a major step forward for this underrepresented African language, opening the door to numerous applications. The results obtained with our model show that this annotation process is a robust methodology that can be replicated for other low-resource languages. It also shows the importance of having reliable benchmarks for evaluating text-to-speech models for low-resource languages. With more powerful models, such advances will enable speech technology to become more widely available and benefit all populations.

6 Limitations and Future Work

Although this is the first controllable text-to-speech model in Wolof, it has a number of limitations. The quality of the voice generated remains average, with a perceptual quality (PESQ) of [x], which needs to be improved. Similarly, the number of parameters controlled remains relatively modest, although satisfactory for a start. Parameters such as dialect will need to be taken into account in future studies, as they are of significant importance in the social context of African languages.

Other limitations are inherent in the architecture of Parler-TTS itself, which we have used as a base model. For our model, as for other versions of Parler-TTS, we have noticed difficulties in correctly pronouncing words that are unknown or rarely encountered in the training dataset. Parler-TTS also has difficulty handling long utterances. One solution is to use the model in streaming mode so that Parler-TTS processes the text by chunk and thus generates coherent speech despite the length. We have also noticed that Parler-TTS, once finetuned, tends to forget the characteristics of the speakers it has been pre-trained on. This limits its potential applications in speech cloning for a very large number of speakers. Improvements to the Parler-TTS architecture or the exploration of other architectures for controllable speech synthesis are therefore avenues worth exploring.

For the data, the main limitations are the quality of the data and the lack of diversity of prosodic features. Quality has been improved using an enhancement system, while diversity has been increased using signal processing techniques. However, these modifications fall short of studio recording quality and do not allow natural emotions and variations in tone to be rendered with sufficient fidelity. Research into the collection of high-quality, expressive text-to-speech data would therefore be salutary and complementary to work such as this.

7 Ethical Considerations

Ethical aspects are an important part of this work, given the potential applications of such a system. On the one hand, this text-to-speech system can make a significant contribution to language preservation and the digital inclusion of marginalised communities, thereby reducing the technological divide. However, this technology raises critical questions about the use of such a model for the purpose of cloning other people's voices without their consent. This may open the door to new methods of voice spoofing in languages where people are not used to encountering this kind of problem. It is therefore important to impose strict conditions of use on these models, which is what we intend to do in its future deployment.

Furthermore, the potential mismatch between imported technologies and local sociolinguistic norms could lead to such a system being rejected or judged negatively by its users. It is therefore important to put the people who speak these low-resource languages at the heart of voice data collection. It is also important that we have the outputs of this model validated by Wolof speakers, beyond the simple estimation of vocal quality. Any study of speech systems for low-resource languages should give high priority to these considerations. This is how these speech synthesis models will become truly useful tools, adapted to the context in which they are to be deployed.

Metrics	MOS	PESQ	WER	Speaker Sim.	Pitch Corr.	SNR
Galsen AI XTTS v2	3.06	**3.90**	1.18	0.40	0.07	**49.31**
Parler Mini Multilingual	1.55	2.84	0.99	0.13	0.05	14.17
Ours (Finetuned Parler)	**3.73**	3.34	**0.45**	**0.54**	**0.14**	26.67

Table 2: Table of objective and subjective metrics across models.

References

Ronald Brian Adler, George R Rodman, and Carrie Cropley. 2006. *Understanding human communication*, volume 10. Oxford University Press Oxford.

Zolzaya Byambadorj, Ryota Nishimura, Altangerel Ayush, Kengo Ohta, and Norihide Kitaoka. 2021. Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):42.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and 1 others. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan. 2019. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6940–6944. IEEE.

Elodie Gauthier, Papa-Séga Wade, Thierry Moudenc, Patrice Collen, Emilie De Neef, Oumar Ba, Ndeye Khoyane Cama, Cheikh Ahmadou Bamba Kebe, Ndeye Aissatou Gningue, and Thomas Mendo'O Aristide. 2024. Preuve de concept d'un bot vocal dialoguant en wolof. *arXiv preprint arXiv:2404.02009*.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Mutian He, Yan Deng, and Lei He. 2019. Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural tts. *arXiv preprint arXiv:1906.00672*.

Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, and Jaime Lorenzo-Trueba. 2021. Low-resource expressive text-to-speech using data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6593–6597. IEEE.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024a. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

Shengpeng Ji, Jialong Zuo, Wen Wang, Minghui Fang, Siqi Zheng, Qian Chen, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, and 1 others. 2024b. Controlspeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec. *arXiv preprint arXiv:2406.01205*.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 2024. Speechcraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1255–1264.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. 2024. Data-speech. https://github.com/ylacombe/dataspeech.

Peter Ladefoged. 1990. The revised international phonetic alphabet. *Language*, 66(3):550–552.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. 2019. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE.

Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. 2023. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *arXiv preprint arXiv:2311.12454*.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Matthias Mauch and Simon Dixon. 2014. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 659–663. IEEE.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, and 1 others. 2022. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. *arXiv preprint arXiv:2207.03546*.

Max Morrison, Zeyu Jin, Nicholas J Bryan, Juan-Pablo Caceres, and Bryan Pardo. 2021. Neural pitch-shifting and time-stretching with controllable lpcnet. *arXiv preprint arXiv:2110.02360*.

Perez Ogayo, Graham Neubig, and Alan W Black. 2022. Building african voices. *arXiv preprint arXiv:2207.00688*.

Sewade Ogun, Abraham T Owodunni, Tobi Olatunji, Eniola Alese, Babatunde Oladimeji, Tejumade Afonja, Kayode Olaleye, Naome A Etori, and Tosin Adewumi. 2024. 1000 african voices: Advancing inclusive multi-speaker multi-accent speech synthesis. *arXiv preprint arXiv:2406.11727*.

Frederico S. Oliveira, Edresson Casanova, Arnaldo Cândido Júnior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. Cml-tts a multilingual dataset for speech synthesis in low-resource languages. *Preprint*, arXiv:2306.10097.

Ka Omar. 1987. Wolof phonology and morphology: A non-linear approach. *Urbana, IL: University of Illiniois Ph. D. dissertation*.

ITU-T Recommendation. 2001. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*.

Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Almost unsupervised text to speech and automatic speech recognition. In *International conference on machine learning*, pages 5410–5419. PMLR.

Annie Rialland and Stéphane Robert. 2001. The intonational system of wolof.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. 2021. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2017. Listening while speaking: Speech chain by deep learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 301–308. IEEE.

Tao Tu, Yuan-Jui Chen, Cheng-chieh Yeh, and Hung-Yi Lee. 2019. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*.

Tianxin Xie, Yan Rong, Pengfei Zhang, and Li Liu. 2024. Towards controllable speech synthesis in the era of large language models: A survey. *arXiv preprint arXiv:2412.06602*.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. 2024. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 554–563.