LLM-to-Speech: A Synthetic Data Pipeline for Training Dialectal Text-to-Speech Models

Ahmed Khaled Khamis
Georgia Institute of Technology
akhamis6@gatech.edu

Hesham Ali
Nile University
he.ali@nu.edu.eg

Abstract

Despite the advances in neural text to speech (TTS), many Arabic dialectal varieties remain marginally addressed, with most resources concentrated on Modern Spoken Arabic (MSA) and Gulf dialects, leaving Egyptian Arabic—the most widely understood Arabic dialect—severely under-resourced. We address this gap by introducing NileTTS: 38 hours of transcribed speech from two speakers across diverse domains including medical, sales, and general conversations. We construct this dataset using a novel synthetic pipeline: large language models (LLM) generate Egyptian Arabic content, which is then converted to natural speech using audio synthesis tools, followed by automatic transcription and speaker diarization with manual quality verification. We fine-tune XTTS v2, a state-of-the-art multilingual TTS model, on our dataset and evaluate against the baseline model trained on other Arabic dialects. Our contributions include: (1) the first publicly available Egyptian Arabic TTS dataset, (2) a reproducible synthetic data generation pipeline for dialectal TTS, and (3) an open-source fine-tuned model. All resources are released to advance Egyptian Arabic speech synthesis research.

1 Introduction

Neural text-to-speech (TTS) has made remarkable progress in recent years, with models like Tacotron (Wang et al., 2017), FastSpeech (Ren et al., 2019), and VITS (Kim et al., 2021) achieving near-human naturalness for high-resource languages. More recently, multilingual TTS systems such as XTTS (Casanova et al., 2024) and VALL-E (Wang et al., 2023) have demonstrated impressive zero-shot voice cloning capabilities across different languages. However, this progress has not been evenly distributed, as low-resource languages and dialectal varieties remain significantly underserved.

Arabic presents a particularly challenging case for TTS research. While Modern Standard Arabic (MSA) has received considerable attention, the spoken reality of the Arab world is far more diverse. Arabic has many regional dialects that differ substantially in phonology, vocabulary, and syntax, often to the point of mutual unintelligibility (Abu Kwaik et al., 2018). Among these, Egyptian Arabic holds a unique position: spoken natively by over 100 million people and widely understood across the Arab world due to Egypt's dominant media presence, it is arguably the most accessible Arabic variety.

Despite its prominence, Egyptian Arabic remains under-resourced for speech synthesis. While prior work has explored Egyptian Arabic TTS (Azab et al., 2023), (Lodagala et al., 2025), existing resources are limited in scale, domain coverage, or public availability. Current Arabic TTS systems mainly target MSA or Gulf dialects, leaving Egyptian Arabic speakers without state of the art tools. As a result, Egyptian Arabic speakers lack access to quality TTS in applications like voice assistants and audiobooks.

In this work, we address this resource gap by introducing NileTTS[1] , a large-scale Egyptian Arabic TTS dataset along with a fine-tuned speech synthesis model. Our dataset comprises 38 hours of transcribed Egyptian Arabic speech from two speakers across three domains: medical, sales and customer service, and general conversation.

A key contribution of our work is the novel synthetic data generation pipeline used to construct the dataset. Rather than relying on costly manual recording, we leverage recent advances in generative AI: large language models (LLMs) generate Egyptian Arabic content across diverse topics, which is then converted to natural-sounding

[1] Code: https://github.com/KickItLikeShika/NileTTS

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 47–54
March 28, 2026. ©2026 Association for Computational Linguistics

speech using neural audio synthesis tools that support Egyptian Arabic. The resulting audio is automatically transcribed using Whisper (Radford et al., 2022) and segmented into utterances, with speaker identities assigned via *ECAPA-TDNN-based* speaker diarization (Desplanques et al., 2020). Manual quality verification ensures transcription accuracy and speaker consistency. This pipeline offers a reproducible and scalable approach for creating TTS datasets for other low-resource dialects.

To demonstrate the utility of our dataset, we fine-tune XTTS v2 (Casanova et al., 2024), a state-of-the-art multilingual TTS model with zero-shot voice cloning capabilities. We evaluate the fine-tuned model against the baseline XTTS v2, which was trained on Arabic data from other dialectal varieties. Our experiments show substantial improvements in intelligibility and speaker similarity. Our contributions are as follows:

- We release **NileTTS**[2], a large-scale Egyptian Arabic TTS dataset comprising 38 hours of transcribed speech across multiple domains.
- We present a **reproducible synthetic data generation pipeline** combining LLM-based content generation, neural audio synthesis, automatic transcription, and speaker diarization.
- We provide an **open-source fine-tuned XTTS model**[3] for Egyptian Arabic, serving as a baseline for future research.

We publicly release all resources to facilitate further research in Egyptian Arabic speech synthesis.

2 Related Work

2.1 Arabic Text-to-Speech

Arabic TTS research has primarily focused on Modern Standard Arabic (MSA), with systems leveraging both traditional concatenative methods and neural approaches (Lodagala et al., 2025). For dialectal Arabic, resources remain scarce. Notable exceptions include work on Gulf Arabic dialects, which benefit from commercial interest in the Gulf region.

For Egyptian Arabic specifically, two prior efforts are most relevant. Azab et al. (2023) introduced EGYARA-23, a 20.5-hour dataset featuring a single male speaker narrating news and general conversations, comprising 32,716 segments. While substantial in size, the dataset is limited to one speaker and two domains. More recently, Lodagala et al. (2025) presented SawtArabi, a multi-dialect Arabic speech corpus that includes approximately one hour of Egyptian Arabic among several other varieties. While valuable for cross-dialectal research, the Egyptian Arabic portion is limited in scale for dedicated TTS training.

Our work complements these efforts by providing a larger, more diverse resource: 38 hours of Egyptian Arabic speech from two speakers (male and female) across three distinct domains. Additionally, we introduce a synthetic data generation pipeline that offers a reproducible approach for future dataset expansion.

2.2 Synthetic Data for Speech

Synthetic data generation has emerged as a promising approach for low-resource speech tasks. Prior work has explored using TTS systems to generate training data for automatic speech recognition (Fazel et al., 2021), and text augmentation via LLMs has shown success in NLP tasks (Ding et al., 2024). Our work extends this paradigm to TTS dataset construction, using LLMs for content generation and neural audio synthesis for speech production—creating a fully synthetic pipeline that requires no manual recording.

2.3 Multilingual TTS and XTTS

Recent advances in multilingual TTS have enabled models to synthesize speech across many languages from a single model. XTTS v2 (Casanova et al., 2024), built on a GPT-style architecture with voice cloning capabilities, supports over 16 languages including Arabic. However, its Arabic training data primarily covers MSA and Gulf dialects. We finetune XTTS v2 on our Egyptian Arabic dataset to adapt it to this under-served variety.

3 Dataset Construction

This section describes the construction of the NileTTS dataset. We present a synthetic data generation pipeline that leverages large language models for content creation, neural audio synthesis for speech generation, and automatic tools for transcription and speaker identification. Figure 1 illustrates the complete pipeline.

[2] Dataset: `https://huggingface.co/datasets/KickItLikeShika/NileTTS-dataset`

[3] Model: `https://huggingface.co/KickItLikeShika/NileTTS-XTTS`

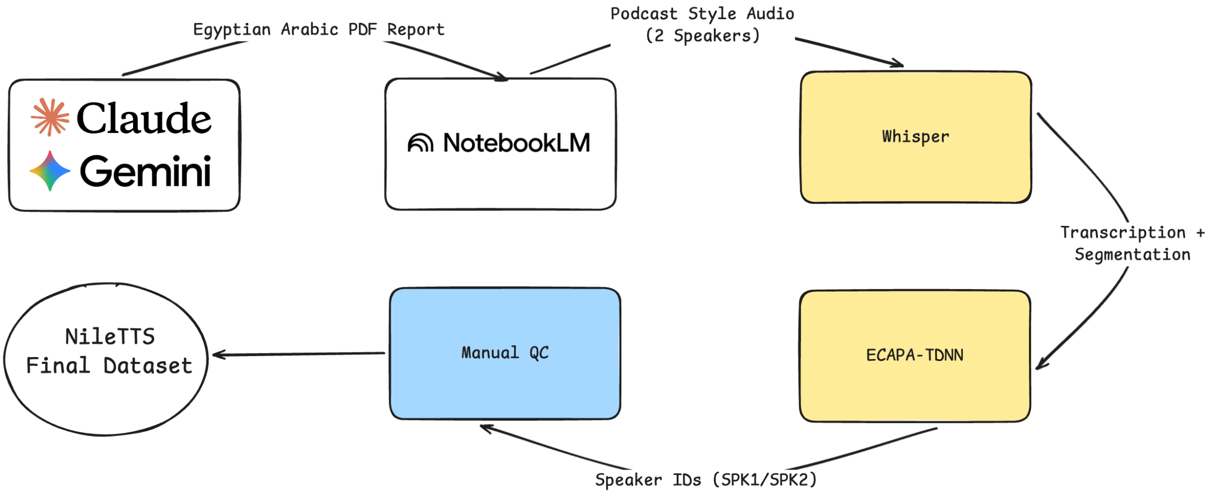

Figure 1: Overview of the NileTTS data generation pipeline. Egyptian Arabic content is generated by LLMs, converted to speech via neural audio synthesis, transcribed and segmented using Whisper, and annotated with speaker identities using ECAPA-TDNN embeddings. Manual quality control ensures accuracy before final dataset compilation.

3.1 Content Generation

The first stage of our pipeline involves generating Egyptian Arabic textual content using large language models. We employ variants ***Gemini*** and ***Claude*** to generate PDF-style reports on diverse topics in authentic Egyptian Arabic dialect. We target three domains to ensure topical diversity:

- **Medical:** Health topics, symptoms, treatments, and medical advice
- **Sales and Customer Service:** Product discussions, negotiation scenarios, customer interactions
- **General Conversations:** Everyday topics, social commentary, cultural discussions

For each generation, we prompt the LLM to write an a report entirely in Egyptian Arabic dialect, explicitly avoiding Modern Standard Arabic. The prompts specify the domain and request natural conversational language that reflects how Egyptians actually speak. This approach yields content that is both topically diverse and linguistically authentic to Egyptian Arabic.

The prompts used for content generation were intentionally **simple**. We did not use complex prompting strategies or explicit normalization rules; instead, we directly instructed the model to generate reports in Egyptian Arabic while avoiding Modern Standard Arabic. Dialect authenticity was assessed **qualitatively** through manual inspection. In practice, both models consistently produced fluent Egyptian Arabic content.

3.2 Audio Synthesis

The generated textual reports are then converted to speech using ***NotebookLM's*** audio generation feature. NotebookLM produces podcast-style audio discussions where two virtual hosts engage in an in-depth natural conversation about the input report. Crucially for our purposes, NotebookLM supports high-quality Egyptian Arabic synthesis with authentic dialect pronunciation. The audio generation produces conversations featuring two distinct speakers: one male and one female voice. Both speakers maintain consistent voice characteristics across all generated audio, which is essential for TTS training data. Each generated audio file is approximately 10-15 minutes in length, covering the content of one PDF report in a conversational format. We selected NotebookLM for several reasons: (1) it produces natural, conversational Egyptian Arabic speech rather than formal MSA; (2) the two-speaker format provides speaker diversity within a consistent voice identity; and (3) the audio quality is suitable for TTS training without significant artifacts.

3.3 Transcription and Segmentation

The generated audio files are processed using OpenAI's Whisper Large model (Radford et al., 2022) for automatic transcription. Whisper provides accurate Arabic transcription with word-level timestamps, which we use for segmentation. We segment the continuous audio into utterance-level chunks. This constraint ensures manageable sequence lengths. Segmentation is performed at natural speech boundaries (pauses between utterances) using the timestamp information from Whisper. Segments shorter than 1 second or containing only silence are discarded. Each segment is saved as an individual WAV file along with its corresponding transcription. This produces the paired (audio, text) format required for TTS training.

3.4 Speaker Diarization

Since the source audio contains two speakers in conversation, we must identify which speaker produced each segment. We employ a speaker diarization approach based on speaker embeddings. We use the ECAPA-TDNN model (Desplanques et al., 2020) from SpeechBrain (Ravanelli et al., 2021) to extract speaker embeddings. ECAPA-TDNN produces a 192-dimensional embedding vector for each audio segment that captures the speaker's voice characteristics independent of linguistic content. Our diarization process works as follows: 1) **Embedding Extraction:** We extract ECAPA-TDNN embeddings for all segments across multiple audio files, 2) **Centroid Computation:** Using K-Means clustering with $k = 2$, we identify two cluster centroids representing the two speakers' average voice characteristics, and 3) **Speaker Assignment:** For each segment, we compute the Cosine Similarity between its embedding and both centroids. The segment is assigned to the speaker whose centroid is closest.

This approach reliably separates the two speakers, as their voice characteristics (male vs. female) are sufficiently distinct in the embedding space. The computed centroids are saved and reused for processing new audio files, ensuring consistent speaker labels across the entire dataset.

3.5 Quality Control

While our pipeline is largely automated, we incorporate manual quality control to ensure dataset quality. Human annotators reviewed a the whole category for Sales and Customer Service, along with a sample of the other 2 sections to verify the following: 1) **Transcription Accuracy:** Checking that the Whisper transcription correctly captures the spoken content, particularly for Egyptian Arabic vocabulary and expressions that may differ from MSA, 2) **Speaker Consistency:** Verifying that the automated speaker labels correctly identify the speaker in each segment, and 3) **Audio Quality:** Ensuring segments are free from artifacts, truncation, or overlapping speech.

Statistic	**Utterances**	**Hours**
Total	9,521	38.1
Training Set	8,571	–
Evaluation Set	950	–
Sales & Customer Service	4,975	21.0
General Conversations	2,979	11.2
Medical	1,567	5.9
SPEAKER_01 (Male)	4,865	–
SPEAKER_02 (Female)	4,656	–
Average Utterance Length	14.4 seconds	

Table 1: NileTTS dataset statistics.

Segments with significant errors are corrected or removed. This quality control step is essential for maintaining dataset integrity, particularly for dialectal content where automatic tools may have higher error rates than for standard language varieties.

3.6 Dataset Statistics

Table 1 summarizes the NileTTS dataset. The final dataset comprises 38.1 hours of transcribed Egyptian Arabic speech, totaling 9,521 utterances. We split the data into training (90%) and evaluation (10%) sets, ensuring both speakers appear in both splits while keeping specific utterances exclusive to one split. We ensure that there is **no report-level overlap** between the training and evaluation sets, and that the evaluation set contains **unseen topics and prompts** not used during training. The dataset covers three domains: Sales and Customer Service is the largest (4,975 utterances, 21.0 hours), followed by General Conversations (2,979 utterances, 11.2 hours) and Medical (1,567 utterances, 5.9 hours). Speaker representation is well-balanced, with SPEAKER_01 (male) contributing 4,865 utterances and SPEAKER_02 (female) contributing 4,656 utterances. The conversational format naturally produces roughly equal speaking time between both voices. The average utterance length of 14.4 seconds provides sufficient context for TTS training while remaining within typical sequence length constraints. The dataset is formatted following the XTTS v2 training data specification: each utterance is stored as a WAV file, paired with its transcription and speaker identifier in a metadata

Hyperparameter	Value
Epochs	30
Batch Size	2
Gradient Accumulation Steps	8
Effective Batch Size	16
Learning Rate	5e-6
Optimizer	AdamW
Weight Decay	1e-2
Max Text Length	400 tokens

Table 2: Finetuning hyperparameters for XTTS v2 on NileTTS.

CSV file. This ensures direct compatibility with the XTTS fine-tuning pipeline and facilitates reproducibility.

4 Model Finetuning

4.1 Base Model: XTTS v2

We finetuned XTTS v2 (Casanova et al., 2024), a state-of-the-art multilingual text-to-speech model developed by Coqui. XTTS v2 employs a GPT-style autoregressive architecture that generates discrete audio tokens, which are then decoded into waveforms. The model supports zero-shot voice cloning, allowing it to synthesize speech in a target voice given only a short reference audio clip. XTTS v2 is pretrained on a large multilingual corpus covering 16 languages, including Arabic. However, the Arabic training data primarily consists of Modern Standard Arabic and Gulf dialects, leaving Egyptian Arabic underrepresented. Our finetuning adapts the model to Egyptian Arabic while preserving its voice cloning capabilities.

4.2 Finetuning Configuration

We finetuned the GPT component of XTTS v2 on the NileTTS training set while keeping the DVAE (audio tokenizer) frozen. We largely adopt the default hyperparameters and training setup provided by the Coqui team's finetuning codebase, with minimal modifications. Table 2 summarizes the key training parameters.

Our primary modifications to the training pipeline involve integrating Weights & Biases for experiment tracking and implementing evaluation metrics—including Word Error Rate, Character Error Rate, and Speaker Similarity—computed periodically during training to monitor convergence and enable checkpoint selection.

5 Experiments and Results

5.1 Evaluation Setup

We evaluate our finetuned NileTTS model against the baseline XTTS v2 model to measure improvements in Egyptian Arabic synthesis quality. The baseline is the pretrained XTTS v2, which includes Arabic but primarily covers Modern Standard Arabic and Gulf dialects.

We use the following evaluation metrics, computed on the held-out evaluation set:

- **Evaluation Loss**: Combined text and mel-spectrogram cross-entropy loss as defined by the XTTS architecture.
- **Word Error Rate (WER)**: We synthesize speech from text, transcribe it using Whisper Large (Radford et al., 2022), and compute WER against the original text. Lower WER indicates higher intelligibility.
- **Character Error Rate (CER)**: A finer-grained intelligibility metric computed at the character level.
- **Speaker Similarity**: Cosine similarity between ECAPA-TDNN (Desplanques et al., 2020) speaker embeddings of synthesized and reference audio. Higher similarity indicates better voice cloning.

5.2 Results

Figure 2 illustrates the progress of evaluation metrics throughout training. All metrics show rapid improvement in early training, with loss decreasing and intelligibility metrics (WER, CER) improving substantially within the first 20,000 steps. Beyond this point, metrics begin to look more horizontal, indicating diminishing returns from continued training.

Checkpoint Selection. Although we initially planned for 30 epochs of training, we observe that after approximately 8 epochs (around 35,000 steps), the evaluation metrics stabilize with minimal further improvement. Training was stopped after 13 epochs (55,719 steps) due to this reason. We select the checkpoint at step 34,289 (epoch 8), which achieves a strong balance across all metrics. To validate this selection, we synthesized 50 randomly sampled utterances from the evaluation set and conducted manual listening evaluation. The synthesized speech demonstrated natural prosody, accurate pronunciation of Egyptian Arabic phonemes,

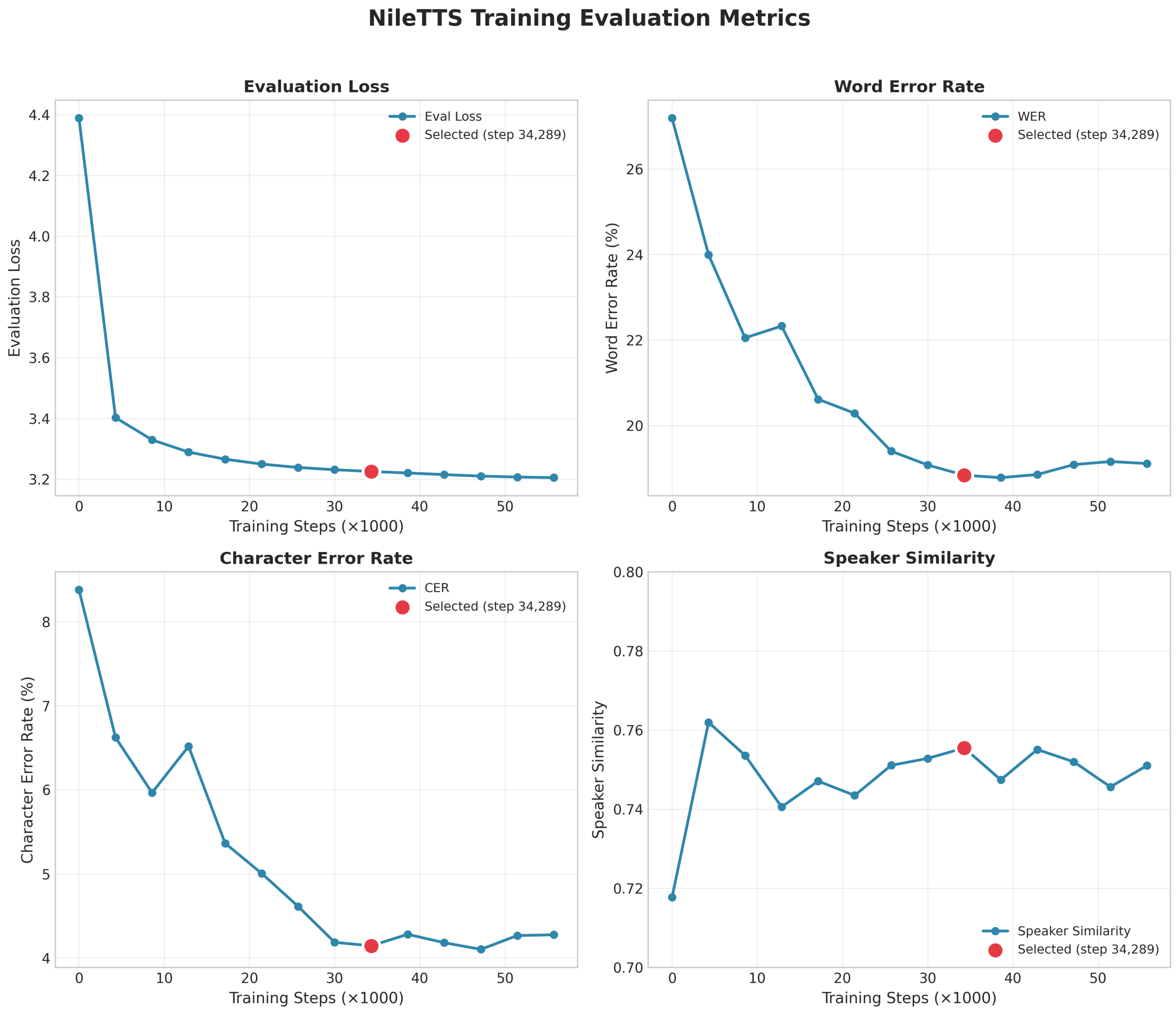

Figure 2: Evaluation metrics throughout training: (a) Evaluation Loss, (b) Word Error Rate, (c) Character Error Rate, (d) Speaker Similarity. The red marker indicates the selected checkpoint at step 34,289 (epoch 8).

Model	WER ↓	CER ↓	Spk Sim ↑
XTTS v2	26.8%	8.1%	0.713
NileTTS	**18.8%**	**4.1%**	**0.755**

Table 3: Comparison of baseline XTTS v2 Baseline and finetuned NileTTS on Egyptian Arabic evaluation set.

and consistent preservation of speaker identity, confirming the checkpoint's suitability for release.

Table 3 presents the final comparison between the baseline XTTS v2 model and our finetuned NileTTS model.

NileTTS achieves a **29.9% relative reduction in Word Error Rate** (from 26.8% to 18.8%) and a **49.4% relative reduction in Character Error Rate** (from 8.1% to 4.1%), indicating significantly improved intelligibility for Egyptian Arabic synthesis. Speaker similarity improves from 0.713 to 0.755 (**+5.9%**), demonstrating better voice cloning.

These results confirm that finetuning on dialect-specific data yields substantial improvements in TTS quality, even when the base model already supports the target language family. We release the NileTTS model weights publicly on Hugging Face to serve as a foundation for future Egyptian Arabic speech synthesis research.

6 Discussion and Limitations

6.1 Dataset Limitations

While NileTTS represents a significant resource for Egyptian Arabic TTS, limitations should be acknowledged. First, the dataset contains only **two speakers** (one male, one female), which limits speaker diversity. TTS models trained on limited speaker data may not generalize well to synthesizing voices with different characteristics. Future work should expand the dataset with additional speakers to improve voice diversity.

Second, our dataset is constructed from **synthetically generated audio** rather than recordings of human speakers. Even though the used audio synthesis tool produces high-quality Egyptian Arabic speech with natural prosody. However, our evaluation results suggest that the synthetic data is sufficient for training effective TTS models, and the pipeline's reproducibility enables future expansion with additional synthetic or natural data.

Third, although we cover three domains (medical, sales, and general conversations), certain **specialized domains** such as news broadcasting, poetry, or technical content are not represented. Expanding domain coverage would improve the model's versatility.

6.2 Evaluation Limitations

Our evaluation relies on **automatic metrics** (WER, CER, Speaker Similarity) rather than formal human evaluation studies such as Mean Opinion Score (MOS) assessments. While automatic metrics correlate with perceived quality, they do not fully capture subjective aspects like naturalness, expressiveness, or listener preference. We mitigate this limitation through manual listening evaluation of synthesized samples, but a comprehensive human evaluation study remains valuable future work.

Additionally, WER and CER are computed using Whisper Large as the transcription model. While Whisper performs well on Egyptian Arabic, transcription errors from the ASR system may introduce noise into these metrics. Moreover, since Whisper is also used to generate the dataset transcripts, this setup may introduce a form of **ASR self-consistency bias**, potentially inflating evaluation scores. Future work should evaluate WER and CER using alternative ASR models and include a small human-verified transcription subset to improve robustness.

Similarly, speaker similarity is computed using cosine similarity between ECAPA-TDNN speaker embeddings. As the same embedding architecture is also used in the pipeline, this may introduce a degree of **model-specific bias** in the speaker similarity scores. Future work will explore the use of additional speaker embedding models for confirmation and incorporate human verification to further validate speaker similarity assessments.

6.3 Synthetic-to-Real Generalization

A common concern with synthetic speech datasets is whether models trained on them generalize to real human speech. Although NileTTS is built from synthetically generated audio, the text content is written in authentic Egyptian Arabic, and the synthesis preserves key dialectal properties such as pronunciation, intonation, and prosodic patterns. These aspects are essential for learning dialect-specific speech characteristics.

In low-resource settings, synthetic speech has been shown to be a practical and effective training signal when natural data is limited. In this sense, NileTTS provides a scalable source of Egyptian Arabic speech data that captures core linguistic properties of the dialect and can complement future datasets based on real human recordings.

6.4 Future Work

Several directions could extend this work:

- **Speaker expansion**: Adding more speakers with diverse voice characteristics, ages, and speaking styles.
- **Other Arabic dialects**: Applying the synthetic data pipeline to other under-resourced Arabic varieties.
- **Human evaluation**: Conducting formal MOS studies to complement automatic metrics.
- **Robust evaluation**: Evaluating WER, CER, and speaker similarity using multiple independent models and human-verified subsets.

7 Conclusion

We presented **NileTTS**, a large-scale Egyptian Arabic text-to-speech dataset and finetuned model. Our dataset comprises 38 hours of transcribed Egyptian Arabic speech from two speakers across medical, sales, and general conversation domains. We introduced a novel **synthetic data generation pipeline** that leverages large language models for content creation, neural audio synthesis for speech generation, and automatic transcription with speaker diarization—offering a reproducible and scalable approach for creating TTS datasets for low-resource dialects.

By finetuning XTTS v2 on NileTTS, we achieved substantial improvements over the baseline Arabic model: **29.9% relative reduction in Word Error Rate**, **49.4% reduction in Character**

Error Rate, and **5.9% improvement in speaker similarity**. These results demonstrate that dialect-specific finetuning significantly enhances TTS quality for underrepresented language varieties.

We publicly release the NileTTS dataset, model weights, and pipeline code to facilitate further research in Egyptian Arabic speech synthesis. We hope this work contributes to closing the resource gap for Arabic dialects and inspires similar efforts for other low-resource language varieties.

References

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A lexical distance study of arabic dialects. *Procedia Computer Science*, 142:2–13.

Ahmed Hammad Azab, Ahmed B. Zaky, Tetsuji Ogawa, and Walid Gomaa. 2023. Masry: A text-to-speech system for the egyptian arabic. *Proceedings of the International Conference on Informatics in Control, Automation and Robotics*, 2:219–226. Publisher Copyright: © 2023 by SCITEPRESS - Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0).; 20th International Conference on Informatics in Control, Automation and Robotics, ICINCO 2023 ; Conference date: 13-11-2023 Through 15-11-2023.

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

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using large language models: Data perspectives, learning paradigms and challenges. *Preprint*, arXiv:2403.02990.

Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. Synthasr: Unlocking synthetic data for speech recognition. *Preprint*, arXiv:2106.07803.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Preprint*, arXiv:2106.06103.

Vasista Lodagala, Lamya Alkanhal, Daniel Izham, Shivam Mehta, Shammur Chowdhury, Aqeelah Makki, Hamdy Hussein, Gustav Henter, and Ahmed Ali. 2025. Sawtarabi: A benchmark corpus for arabic tts. standard, dialectal and code-switching.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. Speechbrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624.

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. *Preprint*, arXiv:1905.09263.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. *Preprint*, arXiv:1703.10135.