

Masry: A Text-to-Speech System for the Egyptian Arabic

Ahmed Hammad Azab¹^a, Ahmed B. Zaky^{2,4}^b, Tetsuji Ogawa³^c and Walid Gomaa^{1,5}^d

¹Computer Science and Engineering, Egypt-Japan University of Science and Technology, Alexandria, Egypt

²Computer Science and Information Technology Programs (CSIT),
Egypt Japan University of Science and Technology, Egypt

³Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan

⁴Shoubra Faculty of Engineering, Benha University, Benha, Egypt

⁵Faculty of Engineering, Alexandria University, Alexandria, Egypt

Keywords: Natural Language Processing, Text-To-Speech, Egyptian Arabic.

Abstract: This paper presents the improvement and evaluation of **Masry**, an end-to-end system planned to synthesize Egyptian Arabic speech. The proposed approach leverages the capable Tacotron speech synthesis models, counting Tacotron1 and Tacotron2, and integrated with progressed vocoders – Griffin-Lim for Tacotron1 and HiFi-GAN for Tacotron2. By synthesizing waveforms from mel-spectrograms, Masry offers a comprehensive solution for generating natural and expressive Egyptian Arabic speech. To train and validate our system, we construct a dataset including a male speaker describing standard composing pieces and news content in Egyptian Arabic. The sampling rate of recorded data is 44100 Hz, guaranteeing constancy and richness within the synthesized speech output. The execution of our framework was fastidiously assessed through different measurements, with a specific center on the Mean Opinion Score (MOS). The experimental results demonstrated the prevalence of Tacotron2 over Tacotron1, yielding a MOS of 4.48 compared to 3.64. This emphasizes the system's capacity to capture and duplicate the nuances of Egyptian Arabic speech more effectively. Besides, The assessment extended to include fundamental measurements such as word and character error rates (WER and CER). These metrics give a quantitative appraisal of the precision and exactness of the synthesized speech.


1 INTRODUCTION


Text-to-Speech (TTS) technology has become a critical field of research and development. It aims at transforming written text into spoken words, enabling applications like voice assistants, audiobooks, accessibility tools, and language learning platforms with the potential to revolutionize human-computer interactions. (Young et al., 2018)


Arabic stands as one of the most extensively spoken languages, serving as the mother tongue for over 200 million individuals (Versteegh, 2014), and the largest Semitic language. It is composed of two main dialects: Standard Arabic and Dialectal Arabic. While Modern Standard Arabic (MSA) is the formal linguistic standard, Dialectal Arabic represents


the daily spoken variation, exhibiting significant differences within and across countries (Habash, 2022).

However, the literature has been dominated by TTS systems for English, resulting in a gap in developing TTS systems for less commonly spoken low-resource languages and dialects, including Egyptian Arabic (Fahmy et al., 2020). As a widely spoken dialect with unique regional variations and informal characteristics (Abdel-Massih, 2011), Egyptian Arabic requires dedicated attention to its linguistic nuances and cultural relevance. Hence, developing an efficient and accurate TTS system for Egyptian Arabic is paramount to enhancing accessibility and communication for its users. The current work aims to address this imperative by presenting a novel approach for building a high-quality TTS system specifically tailored to the unique characteristics of Egyptian Arabic, drawing insights from existing related work in both Arabic and English TTS systems (Habash, 2022).

^a <https://orcid.org/0009-0007-2461-1040>

^b <https://orcid.org/0000-0002-3107-5043>

^c <https://orcid.org/0000-0002-7316-2073>

^d <https://orcid.org/0000-0002-8518-8908>

Moreover, diacritic and gemination signs, representing short vowels and consonant doubling, respectively, play a crucial role in correctly pronouncing Arabic and its various dialects. However, these signs are often omitted in written texts as most Arab readers are accustomed to inferring them from the context (Habash, 2022). So, this absence poses significant difficulties for TTS systems aiming to accurately represent the diverse pronunciations in Arabic.

The lack of Egyptian datasets poses a significant challenge for training Text-to-Speech (TTS) models. As TTS technology strives for natural and accurate speech synthesis, it heavily relies on large and diverse datasets in various languages, including Egyptian Arabic. Unfortunately, the scarcity of high-quality, annotated data in this specific dialect hinders the development of accurate TTS systems that can effectively mimic the unique characteristics of Egyptian Arabic speech (Baali et al., 2023). Without sufficient training data, TTS models may struggle to capture the nuances of pronunciation, intonation, and linguistic variations specific to the Egyptian dialect, leading to less authentic and less intelligible speech synthesis. Addressing this issue requires collaborative efforts to collect and curate more Egyptian datasets, fostering the advancement of TTS technology to cater to a broader linguistic landscape. Our contributions are as follows:

1. We present a dataset for a male (named, Ashraf) narrating general writing and news in Egyptian Arabic.
2. We present “**Masry**” an end-to-end text-to-speech system for Egyptian Arabic.

The paper is structured as follows: In Section 1, we provide an introduction to the study’s objectives and scope. Section 2 is dedicated to the review of related literature and prior works in the field. Moving forward, Section 3 elaborates on the system architecture, delving into each phase within this framework. Expanding upon the experimental setup, Section 4 elucidates the conducted experiments involving the models. In Section 5, we present the outcomes of these experiments and engage in a comprehensive discussion thereof. Finally, Section 6 encapsulates the paper with a conclusion summarizing our findings and outlining potential avenues for future research.

2 RELATED WORK

Arabic text-to-speech synthesis is one of AI’s NLP challenges. Many attempts have been tried to make systems that can overcome the problem of artificial

voice to create a more human natural voice. Previous works have been done in this area, but most models are made for the English language. Some models have been applied to Arabic without focusing on specific dialects. (Habash, 2022).

2.1 English TTS Models

In English, many TTS models have been developed. The most famous one is Tacotron. Tacotron is an end-to-end generative text-to-speech model that can directly synthesize speech from the text with a simple waveform synthesis module. Its highest achieved MOS score (Mean Opinion Score) is 3.82 (Wang et al., 2017).

The authors in (Ren et al., 2019) proposed a novel method (FastSpeech) using a feed-forward network based on Transformer that generates a mel-spectrogram in parallel for TTS. They extract attention alignments from an encoder-decoder-based teacher model for phoneme duration prediction. A length regulator uses it to expand the source phoneme sequence to match the length of the target mel-spectrogram sequence for producing parallel mel-spectrograms. The conducted experiments on the LJSpeech dataset (Ito and Johnson, 2017) reveal that their parallel model achieves a comparable level of speech quality to autoregressive models.

2.2 Arabic TTS Models

Some TTS models have been developed for the Arabic language. Ossam et al. (Abdel-Hamid et al., 2006) employed an HMM-based (hidden Markov model) approach to enhance synthesized Arabic speech. Their methodology used a statistical model to generate Arabic speech parameters like spectrum, fundamental frequency (F0), and duration of phonemes. They also incorporated a multi-band excitation model and utilized samples from the spectral envelope as spectral parameters.

Zangar et al. (Imene et al., 2018) focused on utilizing Deep Neural Networks (DNN) for duration modeling in Arabic speech synthesis. They compared HMM-based and deep neural network DNN-based duration modeling of various architectures to minimize root mean square prediction error (RMSE). The study concluded that using their DNN for modeling duration outperformed both the HMM-based modeling from the HTS toolkit and the DNN-based modeling from the MERLIN toolkit.

In (Fahmy et al., 2020), the authors proposed a transfer learning end-to-end Modern Standard Arabic (MSA) TTS deep architecture. Their work presents

how they generate high-quality, natural, and human-like Arabic speech that uses an end-to-end neural deep network architecture. The approach is built upon a limited corpus of text and audio pairs, encompassing a relatively small compilation of recorded audio, amounting to 2.41 hours. Notably, it demonstrates the successful utilization of English character embeddings, even when employing diacritic Arabic characters as input. The study further expounds on the pre-processing techniques applied to these audio samples, elucidating the strategies used to optimize outcomes.

The authors in (Abdelali et al., 2022) proposed an end-to-end TTS system for Arabic. They called it the NatiQ system. Their speech synthesizer uses an encoder-decoder architecture with attention. They used the Tacotron model (Tacotron1 and Tacotron2) and transformer model to generate mel-spectrograms from characters. They used WaveRNN vocoder with Tacotron1, WaveGlow vocoder with Tacotron2, and ESPnet transformer with the parallel wavegan to synthesize waveforms from the spectrograms. Two voices, male and female, are used. The authors achieved 4.21 (MOS) for the female voice and 4.40 for the male voice.

3 SYSTEM ARCHITECTURE

The system, **Masry**, is structured into three key elements, depicted in Fig.1. The initial stage involves data preprocessing to refine input data. Subsequently, a text-to-mel-spectrogram model generates mel-spectrum output. Finally, the mel-spectrum is processed by a Vocoder to generate corresponding audio. Additionally, we elaborate on the dataset collected, used for training purposes.

3.1 Dataset

In this section, we will discuss the dataset we collected and its characteristics.

EGYARA-23 Dataset

We collected a dataset and called it EGYARA-23. The EGYARA-23 dataset comprises recordings of a male speaker named Ashraf, who narrates general conversations and news in Egyptian Arabic. The dataset spans 20.5 hours and consists of 105,329 words and 32,716 segments. On average, each segment has a duration of 8 seconds, and the recordings maintain a quality of 44.1 KHz.

We intended to create a comprehensive dataset encompassing a wide range of Egyptian Arabic words, considering that this dialect contains numerous words not present in Modern Standard Arabic (MSA). To en-

sure authenticity, we transcribed the dataset in Egyptian Arabic as used in everyday language, reflecting the actual dialect. Consequently, each segment is accompanied by its respective transcript. Table 1 displays a selection of MSA words and their equivalents in Egyptian Arabic.

Table 1: Difference between words in MSA and Egyptian Arabic.

Egyptian Arabic	MSA	English
استتي	انتظر	Wait
بروح	اذهب	Go
عشان	أجل	Because
عاوز	اريد	Want
معلش	للأسف	Sorry
هات	اعطني	Give me

3.2 Preprocessing

In this section, we will delve into the preprocessing steps, which include diacritization, segmentation, and phonetization.

3.2.1 Diacritization

Egyptian Arabic diacritization encompasses two types of vowels: long vowels, explicitly indicated in the text, and short vowels (diacritics), which are often omitted in modern writing (Habash, 2022), relying on readers' contextual understanding (Abdelali et al., 2022). Accurately restoring these diacritics is paramount for human comprehension and machine-based pronunciation of Arabic words. To address this, we employed Camel Tools (Obeid et al., 2020) as a diacritization tool for Egyptian Arabic, attaining an accuracy exceeding 90%. However, to ensure meticulousness, an expert in Arabic linguistics reviewed the automatically diacritized data, particularly in intricate cases such as named entities and foreign words, which can present challenges even to native speakers. Diacritization holds significant importance in accurately pronouncing words, thus rendering it a crucial step in the preprocessing phase. In Table 2, you can see some Egyptian Arabic words before and after the diacritization of Egyptian Arabic.

3.2.2 Segmentation

Given the constraints of neural architectures in processing lengthy audio samples (Shen et al., 2018), We developed a semi-automated procedure for data collection, where the dataset is partitioned into ap-

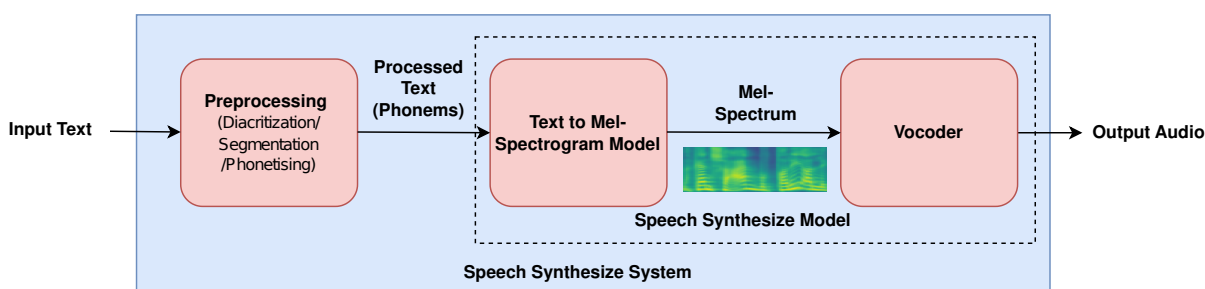


Figure 1: Masry Architecture.

Table 2: Before and after diacritization of samples in Egyptian Arabic.

Before	After	English
استتي	إِسْتِي	Wait
بروح	بَرُوح	Go
عشان	عَشَان	Because
عاوز	عَاوِز	Want
معلش	مَعْلِش	Sorry
هات	هَات	Give me

proximately 8-second frames. This segmentation process involves meticulous attention to maintaining sentence coherence while preserving the overall contextual flow and prosody. Typically, extended pauses between segments are utilized as reliable indicators for segmentation, with exceptions being made when extended pauses are followed by relevant context or supplementary content that remains part of the sentence (Abdelali et al., 2022). This underscores the significance of an intricate approach to segmentation to uphold the precision of audio data representation and its coherent continuity. Employing our semi-automated process, we address this challenge during data collection, resulting in a dataset where each sentence is inherently segmented with established start and end times. Consequently, the audio trimming process is automated to align with each sentence’s corresponding segments. We ensure the entirety of each sentence during data collection, ensuring that no pertinent context or supplementary content is left unresolved within the sentence.

3.2.3 Phonetization

Phonetization converts textual input into its corresponding phonetic representation, mapping each grapheme to its phoneme. In TTS systems, phonetization is crucial to accurately synthesize speech by ensuring correct pronunciation, intonation, and rhythm. For languages like Arabic, with intricate vowel patterns and diacritics, phonetization is particularly im-

portant to capture pronunciation nuances. It enables TTS systems to generate natural and contextually appropriate speech output, creating a more expressive and high-quality auditory experience (El-Imam, 2004). In this phase, our focus shifts towards converting the transcribed text into a phoneme representation, which has already undergone diacritization and segmentation. This transformation occurs through a two-step procedure as you can see in 2. Initially, we transcribe the text into Buckwalter transliteration format. Subsequently, the output from the transliteration step is subjected to further processing to obtain phoneme characters. To accomplish this task effectively, we leverage the Arabic phonetiser tool developed by Nawar Halabi (Halabi, 2016). These sequential steps are essential in preparing the data for subsequent model input, optimizing our research approach in Egyptian Arabic language processing. you can see before and after Phonetization in 3.

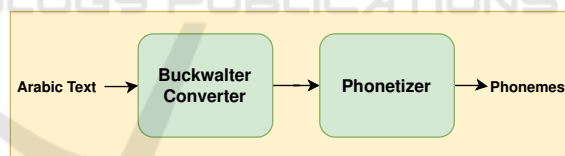


Figure 2: Phonetization Process.

Table 3: Difference between Arabic Text, Buckwalter Transliteration and Phonemes Characters.

Phonemes	Buckwalter	Arabic Text
E a \$ aa n	Ea\$aAno	عَشَان
E aa w i l z	EaAwizo	عَاوِز
m a E l i l \$	maEoli\$o	مَعْلِش
h aa t	haAto	هَات
< i0 s t a nn ii0	Aisotan~iy	إِسْتِي

3.3 Speech Synthesis Model

This section presents an overview of the synthesizer model’s architecture. The synthesizer comprises an encoder-decoder model and a vocoder which is pivotal in generating the desired waveforms. The encoder-decoder module is responsible for converting the preprocessed text into a mel-spectrum representation. Conversely, the vocoder transforms the mel-spectrum representation into the corresponding waveform. To comprehensively explore the effectiveness of the proposed approach, we experimented with two distinct models and vocoders.

Tacotron1, (Wang et al., 2017), adopts an RNN sequence-to-sequence architecture comprising three main components in Fig.3: an encoder, an attention-based decoder, and a post-processing module. The encoder takes text input in the form of characters and transforms it into a mel-spectrogram representation. Subsequently, the post-processing module utilizes this mel-spectrogram to generate the corresponding waveform. The encoder in Tacotron1 utilizes a CBHG-based approach, which involves a bank of 1-D convolutional filters, followed by highway networks and a bidirectional gated recurrent unit (GRU).

Tacotron1 employs the Griffin-Lim algorithm on top of the generated mel-spectrograms to complete the speech synthesis process. The Griffin-Lim algorithm is a famous vocoder used in speech synthesis, including Tacotron1. It converts a mel-spectrogram back into a time-domain waveform. It works iteratively, refining a random waveform estimate to match the target mel-spectrogram. It retains phase information for better waveform reconstruction by alternating between time and frequency domains. This computationally efficient approach strikes a balance between quality and efficiency, making it widely used in real-time speech synthesis.

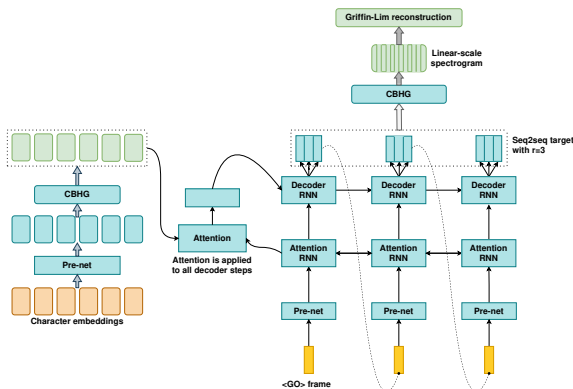


Figure 3: Tacotron 1 architecture (Wang et al., 2017).

Tacotron2 (Shen et al., 2018) is an advanced text-to-speech (TTS) model. In Fig. 4 consisting of a text encoder and a spectrogram generator. The text encoder processes the input text, capturing its linguistic features and context, and produces a fixed-size representation. The spectrogram generator, equipped with attention mechanisms and typically based on recurrent neural networks (RNNs), takes the fixed-size text representation as input and generates mel spectrograms, which represent the spectral content of audio over time. During training, Tacotron2 learns to align the input text with corresponding mel spectrograms using attention mechanisms, enabling it to generate accurate and expressive mel spectrograms. On the other hand, HiFi-GAN (Kong et al., 2020), the high-fidelity generative adversarial network vocoder, is designed to convert mel spectrograms into high-quality audio waveforms. By utilizing a GAN architecture with a generator and a discriminator, HiFi-GAN is trained to synthesize realistic and natural-sounding speech from mel spectrograms. The generator produces the audio waveforms while the discriminator tries to distinguish between real and generated audio, leading to adversarial training that enhances the quality of the generated speech. With the combination of Tacotron2 and HiFi-GAN, the TTS system can generate human-like speech by first generating mel spectrograms that retain essential characteristics of the audio and then utilizing the high-fidelity vocoder to transform these spectrograms into realistic and high-quality speech waveforms.

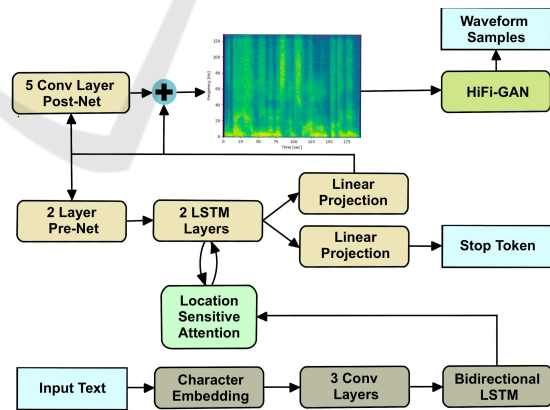


Figure 4: Tacotron 2 architecture (Shen et al., 2018) with modification HiFi-GAN vocoder.

4 EXPERIMENTS

To evaluate the performance of each model, we performed computational experiments with automatic and manual (subjective) evaluations.

4.1 Training Phase

In the training procedure, we adopt a two-step approach. Firstly, we train the feature prediction network independently to predict certain features from phonemes input. Subsequently, we proceed to train a HiFi-GAN separately, utilizing the outputs generated by the feature prediction network as its input. This two-step process allows us to effectively leverage the predictions made by the first network and enhance the overall performance of the HiFi-GAN. The training process involved setting hyperparameters, such as a batch size of 32, and conducting 1000 epochs. Initially, we preprocessed the input text by converting it into phonemes. Subsequently, the training procedure commenced, during which we took checkpoints at every 1000 steps. These checkpoints were used to test and monitor the training progress to ensure effectiveness and efficiency. The computing environment utilized an Intel Xeon 6230R @ 2.1 GHz (52 CPUs) with NVIDIA Quadro RTX5000.

4.2 Testing Phase

To be able to test the models we build a testing dataset called EGYARA-TEST that shares the same characteristics as the EGYARA-23. However, EGYARA-TEST encompasses a duration of 1 hour and comprises 1003 segments. This new dataset was explicitly designed to thoroughly test the performance of the models on a more extensive and diverse set of audio samples. By using this dataset, we aimed to evaluate how well the models perform and assess their robustness and generalization capabilities across a broader range of speech segments.

For model evaluation, we employ two types of assessment: automatic evaluation and manual (subjective) evaluation.

4.2.1 Automatic Evaluation

In this study, we employed an Egyptian Arabic Automatic Speech Recognition (ASR) system developed by (Alyafeai, 2022) to decode the audio files generated by our Text-to-Speech (TTS) models. To compare the generated transcripts with the input sentences used for TTS output, we adapted the reference original text to match the unvoiced output from the ASR system for a fair comparison. Standard evaluation metrics, including Word Error Rate (WER) and Character Error Rate (CER), were used to assess the TTS model's performance. Word Error Rate (WER) is a widely used metric for automatic speech recognition performance. It tackles challenges arising from variable sequence lengths by employing Levenshtein

distance at the word level. WER facilitates system comparisons and improvements assessment, though it lacks specificity on error types. Addressing this, dynamic string alignment aligns recognized and referenced word sequences. The power law theory examines perplexity's correlation with WER, shedding light on language model complexity's impact on error rates (Morris et al., 2004). Word Error Rate (WER) can subsequently be calculated as:

$$WER = (S + D + I) / N = (S + D + I) / (S + D + C) \quad (1)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, N is the number of words in the reference ($N=S+D+C$)

Character Error Rate (CER) is a prevalent performance metric for automatic speech recognition systems. Similar to Word Error Rate (WER), CER operates at the character level rather than the word level. For detailed insights (Morris et al., 2004). The computation of the Character Error Rate involves:

$$CER = (S + D + I) / N = (S + D + I) / (S + D + C) \quad (2)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct characters, N is the number of characters in the reference ($N=S+D+C$).

4.2.2 Manual Evaluation

In Manual Evaluation, we used Mean Opinion Score (MOS) (Guski, 1997), it is a subjective evaluation metric used to measure the perceived quality of generated speech or audio. Human participants rate the samples on a scale; the average score indicates the system's quality. MOS evaluations help identify strengths and weaknesses in TTS and ASR systems and guide improvements for better user satisfaction. It is one aspect of a comprehensive evaluation approach that considers other metrics and user feedback. We conducted an anonymous survey to evaluate the output of our model. 108 participants were asked to rate the audio samples on a scale from 1 to 5, where higher scores indicated better quality. The survey included 10 audio samples, and we calculated the Mean Opinion Score (MOS) based on the participants' ratings. This approach ensured the confidentiality of the participant's responses and provided valuable insights into the perceived quality of the model's output.

5 RESULTS

The evaluation metrics of Character Error Rate (CER), Word Error Rate (WER) based on 1 and 2,

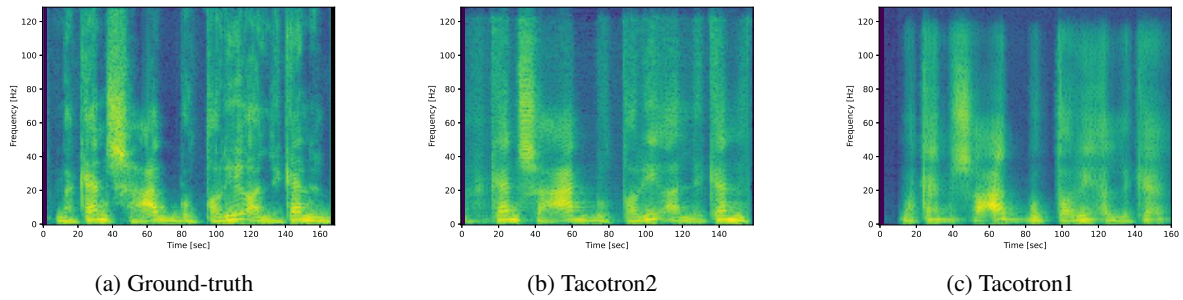


Figure 5: Comparison of predicted Mel-spectrograms for input text (مكنش عاجبه الطريقة اللي كان) (He did not like the way that was).

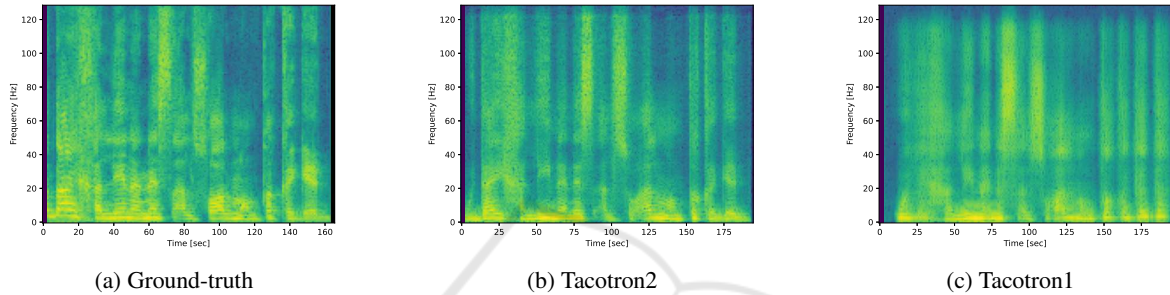


Figure 6: Comparison of predicted Mel-spectrograms for input text (وده يخلينا نسأل سؤال منطقي جدا) (This leaves us to ask a very logical question).

and Mean Opinion Score (MOS) are given in Table 4 and Table 5, Our best-performing speech synthesizer is Tacotron2 with HiFi-GAN for Egyptian Arabic.

The combination of the mel-spectrogram prediction model of Tacotron2 and HiFi-GAN has demonstrated superior performance in low CER and WER, indicating accurate and precise speech generation. Additionally, the high MOS scores obtained from subjective evaluations reflect the perceived quality and naturalness of the synthesized speech. The results from these metrics collectively support the conclusion that Tacotron2 with HiFi-GAN is the most effective and preferred choice among the tested speech synthesis models for Egyptian Arabic. Furthermore, comparing this model's (MOS) scores against those of Modern Standard Arabic TTS systems provides additional validation of its superiority. This affirmation is underscored not only by the observed (Word Error Rate) (WER) and (Character Error Rate) (CER) evaluations but also by the proximity and surpassing of our achieved MOS scores in relation to other systems in the MSA context. Noteworthy examples include the NatiQ system (Abdelali et al., 2022), where our model excels, and the Transfer Learning End-to-End Arabic Text-To-Speech (TTS) Deep Architecture (Fahmy et al., 2020), which further accentuates the efficacy of our approach. Additionally, in terms of (WER) and (CER), our model surpasses the per-

formance of the NatiQ system (Abdelali et al., 2022).

In Fig.5 and Fig.6, a comparison of Mel-spectrograms for three different audios are presented with two different sentences: a) the ground-truth, b) Tacotron2, and c) Tacotron1. Upon observation, we can notice that there is a higher degree of similarity between the ground-truth and Tacotron2 Mel-spectrograms compared to Tacotron1. This finding suggests that Tacotron2 is better at generating Mel-spectrograms that closely resemble the ground-truth, indicating higher accuracy and fidelity in the speech synthesis process. On the other hand, Tacotron1's Mel-spectrograms show more noticeable differences from the ground-truth, suggesting that it may not capture specific acoustic characteristics as effectively as Tacotron2.

Table 4: CER and WER evaluation results.

	CER	WER
NatiQ TTS System for (MSA) (Abdelali et al., 2022)	8.01	24.87
Tacotron1 (Griffin-Lim) for Egyptian Arabic	20.4	66.6
Tacotron2 (HiFi-GAN) for Egyptian Arabic (proposed)	7.3	22.3

Table 5: MOS evaluation results (maximum score is 5).

	MOS
TL TTS System for (MSA) (Fahmy et al., 2020)	4.21
NatiQ TTS System for (MSA) (Abdelali et al., 2022)	4.40
Tacotron1 (Griffin-Lim) for Egyptian Arabic	3.64
Tacotron2 (HiFi-GAN) for Egyptian Arabic (proposed)	4.48

6 CONCLUSION

This paper presents the **Masry** end-to-end text-to-speech system tailored for Egyptian Arabic, combining Tacotron2 with the HiFi-GAN Vocoder. Additionally, a novel dataset and its transcriptions in Egyptian Arabic were introduced. The system's performance was assessed through automatic evaluation metrics, namely Character Error Rate (CER) and Word Error Rate (WER), resulting in scores of 7.3 and 22.3, respectively. Furthermore, a manual evaluation using the Mean Opinion Score (MOS) yielded a score of 4.48. Our findings indicate that the system's performance is in close proximity to that of English and Modern Arabic Standard systems. Future work entails incorporating additional features, such as emotions and multispeaker capabilities, to enhance the system's capabilities further.

REFERENCES

- Abdel-Hamid, O., Abdou, S. M., and Rashwan, M. (2006). Improving arabic hmm based speech synthesis quality. In *Ninth International Conference on Spoken Language Processing*. n/a.
- Abdel-Massih, E. T. (2011). *An Introduction to Egyptian Arabic*. MPublishing, University of Michigan Library.
- Abdelali, A., Durrani, N., Demiroglu, C., Dalvi, F., Mubarak, H., and Darwish, K. (2022). Natiq: An end-to-end text-to-speech system for arabic. *arXiv preprint arXiv:2206.07373*.
- Alyafeai, Z. (2022). Klaam asr. <https://github.com/ARBML/klaam>.
- Baali, M., Hayashi, T., Mubarak, H., Maiti, S., Watanabe, S., El-Hajj, W., and Ali, A. (2023). Unsupervised data selection for TTS: using arabic broadcast news as a case study. *CoRR*, abs/2301.09099.
- El-Imam, Y. A. (2004). Phonetization of arabic: rules and algorithms. *Computer Speech & Language*, 18(4):339–373.
- Fahmy, F. K., Khalil, M. I., and Abbas, H. M. (2020). A transfer learning end-to-end arabic text-to-speech (tts) deep architecture. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 266–277. Springer.
- Guski, R. (1997). Psychological methods for evaluating sound quality and assessing acoustic information. *Acta Acustica united with Acustica*, 83:765–774.
- Habash, N. Y. (2022). *Introduction to Arabic natural language processing*. Springer Nature.
- Halabi, N. (2016). *Modern standard Arabic phonetics for speech synthesis*. PhD thesis, University of Southampton.
- Imene, Z., Mnasri, Z., Vincent, C., Denis, J., Amal, H., et al. (2018). Duration modeling using dnn for arabic speech synthesis. In *Proceedings of 9th International Conference on Speech Prosody*, pages 597–601.
- Ito, K. and Johnson, L. (2017). The lj speech dataset. <https://keithito.com/LJ-Speech-Dataset/>.
- Kong, J., Kim, J., and Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.
- Morris, A., Maier, V., and Green, P. (2004). From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea*.
- Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMEL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.
- Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., 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.
- Versteegh, K. (2014). *Arabic language*. Edinburgh University Press.
- Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R. A. J., and Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In *Interspeech*.
- Young, M., Courtad, C. A., Douglas, K., and Chung, Y.-C. (2018). The effects of text-to-speech on reading outcomes for secondary students with learning disabilities. *Journal of Special Education Technology*, 34:016264341878604.