INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# ClArTTS: An Open-Source Classical Arabic Text-to-Speech Corpus

*Ajinkya Kulkarni*[*], *Atharva Kulkarni*[†]
*Sara Abedalmon'em Mohammad Shatnawi*[*], *Hanan Aldarmaki*[*]

[*]MBZUAI, UAE, [†]Erisha Labs, India

[*]firstname.lastname@mbzuai.ac.ae, [†]atharva7kulkarni@gmail.com

## Abstract

We present a Classical Arabic Text-to-Speech (ClArTTS) corpus to facilitate the development of end-to-end TTS systems for the Arabic language. The speech is extracted from a LibriVox audiobook, which is then processed, segmented, and manually transcribed and annotated. The ClArTTS corpus contains about 12 hours of speech from a single male speaker sampled at 40100 Hz. In this paper, we describe the process of corpus creation, details of corpus statistics, and a comparison with existing resources. Furthermore, we develop two TTS systems based on Grad-TTS and Glow-TTS and illustrate the performance of the resulting systems via subjective and objective evaluations. The ClArTTS corpus is publicly available at www.clartts.com for research purposes, along with the baseline TTS systems and an interactive demo.

**Index Terms**: arabic speech corpus, text-to-speech, corpus creation

## 1. Introduction

Neural text-to-speech (TTS) models are becoming mainstream due to their superior performance in synthesizing intelligible and natural-sounding speech [1, 2, 3]. Compared to older concatenative (e.g. [4]) or HMM-based [5] TTS models, neural models can generate raw waveform directly from text inputs without complex pre-processing and phonetic feature extraction. Neural TTS models commonly have two main components: an acoustic model that generates acoustic features (e.g. mel-spectrograms) directly from the text, and a vocoder to generate a waveform from the acoustic features (see for example [6]). Fully end-to-end TTS models that combine both stages have also been explored [7]. While these neural architectures can be complex, end-to-end training alleviates the need for feature engineering and other design choices that are prone to be sub-optimal. One of the bottlenecks in TTS system design, however, is the availability and quality of the speech corpora used for training. Unlike ASR datasets, where it is desirable to have a variety of speakers and recording conditions to achieve robust performance, it is far more advantageous to have consistent single-speaker corpora for TTS to achieve intelligible and natural-sounding synthesis. Therefore, speech data used for training TTS models need to have more consistent acoustic features that ideally only vary along phonetic and prosodic dimensions.

Such datasets can be conveniently extracted from pre-existing audiobooks; for example, the LJ Speech corpus[1] includes $\sim$ 24 hours of speech extracted from seven audiobooks by the same female speaker. Meanwhile, due to the scarcity of freely available speech corpora of this kind, a larger gap exists in Arabic TTS research and development. Most of the existing freely available Arabic speech corpora are not suitable for TTS training as they contain multi-speaker casual speech with variations in recording conditions and quality, whereas the corpora curated for speech synthesis are generally small in size and not suitable for training state-of-the-art end-to-end models [8, 9]. The existing corpora for Arabic TTS are carefully designed and reduced datasets that are optimized for phonetic coverage while maintaining a relatively small number of units [10, 11]. This choice is partially a remnant of early concatenative models that have a real-time computational cost proportional to the size of the dataset. Another reason for this choice is the relative difficulty of constructing consistent datasets that are suitable for TTS training, especially if they need to be annotated at the phonetic level for traditional TTS systems, so a reduced dataset that maintains phonetic coverage is more manageable to construct. For example, one of the most commonly used public TTS datasets for Arabic is the Arabic Speech Corpus (ASC) [10], which has around 3.4 hours of speech. The ASC was designed to maximize phonetic coverage using a greedy optimization strategy. While such optimization technique is commonly used in most TTS data construction projects, there is some evidence that a random subset of the same size could potentially lead to similar or even more natural-sounding speech synthesis [12]. In addition, for neural TTS models, quantity is more beneficial to the overall quality of the synthesized speech as they are more robust to small variations in input conditions. Moreover, neural TTS models can work directly with text utterances as input without the need for phonetic annotations, which makes the construction of larger datasets more feasible.

In this work, we construct a relatively large single-speaker corpus to enable wider exploration and adaptation of end-to-end TTS approaches for the Arabic language and bridge this gap in data availability, which is a stepping stone towards more inclusive spoken language technologies. To our knowledge, ClArTTS is the largest freely available single-speaker speech corpus in the Arabic language. In particular, the corpus consists of audio recordings by a male speaker of a book written in Classical Arabic sampled at 40100 Hz, which is available in the LibriVox project. To create a corpus for TTS synthesis, we segmented the audio into short utterances, checked for quality and consistency of recording conditions, and manually annotated the audio segments with fully diacritized transcriptions. The final corpus comprises 12 hours and 10 minutes of speech, which is segmented into 10334 utterances. We also built several neural TTS systems using this corpus to demonstrate the quality of the synthesized speech using subjective and objective evaluations. We evaluate the speech synthesis performance for both Classical and Modern Standard Arabic.

[1]https://keithito.com/LJ-Speech-Dataset



10.21437/Interspeech.2023-2224

## 2. Related work

The most commonly used approaches for Arabic speech synthesizers are either based on unit selection or parametric speech synthesis [13, 14, 15]. The Arabic Speech Corpus [10], which is one of the most cited corpora for Arabic speech syntehsis, contains around 3.4 hours of Modern Standard Arabic (MSA) speech recorded at 48KHz.To reduce the size of the corpus, diphone-based greedy optimization strategies were used, and nonsense or dummy utterances were added to cover the gaps of underrepresented phonemes. The Standard Arabic Single Speaker Corpus (SASSC) [16] is another MSA corpus containing 7 hours of speech, professionally recorded for TTS and ASR applications. Zine et al. [17] describe a process of extracting a 4-hour Arabic speech synthesis corpus using a pre-recorded audiobook from the Masmoo3 Audiobooks website, which was then used to construct a concatenative TTS system. Amrouche et al. [18] describe the process of creating a phonetically balanced speech corpus for Arabic, and use the constructed dataset to train a Hidden Markov Model (HMM)-based speech synthesis model. In [19], an end-to-end TTS system based on Tacotron architectures [2] is described. The models were trained using in-house professionally recorded data by two speakers. The resulting systems generate high-quality speech, but the speech datasets are not publicly available for research use.

In terms of the availability of resources for training high-quality TTS systems, Arabic is still considered a low-resource language. The ClArTTS corpus is an attempt to fill the resource gap in Arabic speech research by providing a relatively large single-speaker corpus that can be used to train end-to-end TTS models. The LibriVox [2] project contains free public-domain audiobooks in many languages and has been the basis for many text-to-speech corpora, including the LJ Speech Dataset [3], the M-AILABS multi-lingual corpus [4], and the HI-FI English TTS corpus [20]. The ClArTTS corpus is the first speech corpus extracted from the LibriVox project for the Arabic language.

## 3. Corpus construction

### 3.1. Audio pre-processing

For the creation of a classical Arabic text-to-speech (ClArTTS) corpus, we selected an audiobook recorded by a single speaker from the LibriVox project. The classical book is titled *Kitab Adab al-Dunya w'al-Din* by Abu al-Hasan al-Mawardi (972-1058 AD). The audiobook is recorded by a single speaker and consists of approximately 16 hours of audio without accompanying text. While scanned copies of the book exist, we opted for manual annotation of the audio data to create text transcripts that truly match the audio recording using the Praat annotation tool[5].

The audiobook consists of 20 long audio files, each representing a chapter of the book in MP3 format. We converted this audio to WAV format using *ffmpeg* command-line tool to ensure compatibility with the Praat program. We kept the original sampling rate of 40100 Hz. We ran a rule-based Praat script to mark pauses and speech segments in the long audio files. This script created a TextGrid object for a LongSound object and set boundaries at pauses based on intensity analysis. We validated the marking of pauses and speech segments provided by the Praat tool using energy-based VAD from the Kaldi toolkit[6].

### 3.2. Annotation process

The process of annotating the audiobook involved transcribing audio content into written text, along with additional tags for speech pauses, background noise, inaudible speech segments, and stuttering. The Praat tool was used for the annotation, and the annotators were given TextGrid Praat files that contained the audio recording and a framework for marking speech and pause segments. This helped the annotators efficiently and accurately transcribe the speech segments into written text.

A team of three Arabic annotators was involved in the transcription process to ensure a reliable and accurate final transcript that considered multiple perspectives. To enhance the quality of the transcripts, two rounds of validation were conducted. The first validation was done by the annotators themselves, followed by a check by two other annotators for accuracy and consistency. The text transcripts were marked with Arabic diacritical marks to increase the accuracy of the transcripts for speech analysis and pronunciation.

In addition to the TextGrid Praat files, the annotators were also given a text image of the original book for reference. This made it easier for the annotators to transcribe the speech segments accurately by referring to the original text. Guidelines were provided to the annotators during the annotation process, including instructions for using abbreviations, numbers, special characters, and punctuation according to Arabic language rules. Specific speech segments were marked with tags, including [B] for background noise, [H] for stuttering or hesitation, [*] for unclear speech, and [O] for human noise. The combination of the Praat tool, three annotators, two levels of validation, text transcripts with Arabic diacritization markers, and reference materials assisted in ensuring the accuracy and reliability of the final transcripts.

### 3.3. Final corpus creation

The total amount of original audio is around 16 hours, spanning 20 chapters, so it was recorded in multiple sessions. We observed slight variations in speaking style between the chapters, even though it was neutral (non-emotional) overall. Therefore, we conducted subjective listening tests by listening to random parts of each chapter and removed three chapters that diverge in speaking style compared to the rest.

For segmentation, we split each long-audio file using the textgrid obtained through the Praat tool and the manual annotation process with speech and silence segments. For ensuring high audio quality, we used a signal-to-noise ratio (SNR) to guide the selection process. We estimated the waveform amplitude distribution analysis SNR [21] by taking into account the noise power in silence (non-speech) segments adjacent to the given speech segment. We used a threshold value of 20dB SNR for the first level of speech segment selection. We concentrated adjacent speech segments to create a minimum speech segment duration of 2 seconds. Furthermore, during the concatenation process, we kept only 20% of silence segments between two speech segments if the silence segment duration was exceeding the average silence duration computed across the long audio. We also removed the preamble speech segments, during which the reader briefly talked about the LibriVox project, stated their name and book information, and may have mentioned copyright descriptions or LibriVox project-related content.

[2] https://librivox.org/

[3] https://keithito.com/LJ-Speech-Dataset/

[4] https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/

[5] `https://www.fon.hum.uva.nl/praat/`

[6] `https://kaldi-asr.org`

Table 1: *Corpus statistics comparison between Arabic speech corpus (ASC), Balanced Arabic corpus (BAC) and ClArTTS.*

| Count | BAC | ASC | ClArTTS |
|---|---|---|---|
| Sentences | 202 | 1,913 | 10,334 |
| Words | 1,254 | 17,275 | 82,970 |
| Words/sentence (Avg) | 6 | 9 | 8 |
| Unique words | 975 | 12,144 | 27,870 |
| Phonemes | 6,174 | 135,232 | 518,682 |
| Diphones | 3,614 | 72,797 | 282,487 |
| Unique diphones | - | 682 | 520 |

Table 2: *Percentage of a subset of frequent (Top) and infrequent (bottom) diphones in the ClArTTS corpus vs. a larger text corpus (Tashkeela). In the bottom half, the diphones start with a diacritic, which we display on a dummy character.*

| Diphone | ClArTTS | Tashkeela |
|---|---|---|
| وَ | 3.62% | 3.21% |
| لَـ | 3.09% | 3.00% |
| أَ | 2.89% | 3.53% |
| لا | 2.82% | 1.6% |
| ال | 2.53% | 1.39% |
| عَـ | 2.32% | 2.34% |
| مَـ | 2.24% | 1.95% |
| نا | 2.19% | 1.03% |
| نُص | .00035% | .00065% |
| نُط | .00035% | .00175% |
| نُث | .00035% | .00034% |
| نِط | .00035% | .00163% |
| نُعـ | .00035% | .00009% |
| نَجـ | .00070% | .00011% |
| نَخـ | .00070% | .00082% |
| نُغـ | .00070% | .00154% |

During the segmentation process, we ensured that each segmented speech utterance had a duration of at least 2 seconds and a maximum duration of 10 seconds. We also observed that the Praat pause marking script was unable to tag the last silence segments. Therefore, we manually removed the silence frame in the last audio segments marked by Praat tools. We also removed the speech segments consisting of text transcripts with non-Arabic characters.

We used 3% of the corpus as the test set and 97% as the training set, which results in 10000 utterances in the training set with a total duration of 11 hours and 45 minutes. For the test set, we have a total of 334 utterances, for a total duration of 25 minutes. All text files were saved in UTF-16 encoding and non-Arabic characters were removed. In addition to Arabic transcripts, we have also provided the Buckwalter [22] transliterated transcripts.

## 4. Corpus statistics

Speech corpora that are recorded specifically for the purpose of speech synthesis typically follow a specific procedure to maximize phonetic coverage while minimizing total corpus size [11].

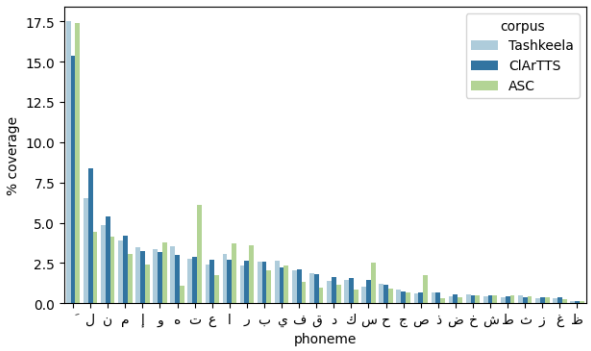


Figure 1: *Percentage coverage of phonemes for ASC, ClarTTS, and the Taksheela text corpus.*

However, since we do not record the corpus and instead use a pre-existing audiobook, we are constrained only by the size of the audiobook. As a result, ClArTTS may not include all possible phonetic combinations, but instead follows the phonetic distribution of the language. In Figure 1, we illustrate a comparison of monophone coverage across three corpora: the Arabic Speech Corpus (ASC), he Tashkeela text corpus [23], and our ClArTTS corpus. The ASC was curated for maximum phonetic coverage, whereas the Tashkeela corpus contains randomly sampled text from various resources, which we assume represents a better representation of the natural phonetic coverage in the Arabic language. We picked this corpus in particular as it contains manually diacritized text, which is essential for our comparison. As shown in Figure 1, ClArTTS phonetic distribution is closer to the natural text distribution than the curated ASC corpus.

We compare our corpus statistics with previously published statistics, namely the Balanced Arabic Corpus described in [11] and the ASC corpus. The statistics are shown in Table 1. ClArTTS is the largest corpus in terms of the number of sentences, words, unique words, phonemes, and diphones, indicating that it is a more extensive and diverse corpus than the other two. ASC has the second-largest number of sentences and words, but the number of unique words is less than half of that in ClArTTS. BAC is the smallest corpus in terms of all the measures listed in the table. The only statistic where we observe a shortage in ClArTTS compared to ASC is the number of unique diphones. In the ASC, dummy utterances were recorded to artificially maximize the total number of diphones, even though these diphones are rare or impossible in the language. Therefore, this shortage in diphone coverage is unlikely to degrade TTS performance for most utterances. In Table 2, show a subset of frequent and infrequent diphones and compare their coverage in ClArTTS and the Tashkeela text corpus, which shows that ClArTTS has good representation of frequent diphones, roughly similar to their natural distribution in the language.

## 5. Baseline TTS systems

To verify the usability of the ClArTTS corpus for speech synthesis, we compare the performance of two baseline text-to-speech (TTS) systems, Grad-TTS [24] and Glow-TTS [25], trained on ClArTTS corpus vs. the Arabic Speech Corpus (ASC). We used the default network parameters as mentioned in the papers [24] and [25] respectively for these TTS systems without using any explicit Arabic grapheme to phoneme mod-

Table 3: *Evaluation metrics computed to measure the performance of baseline end-to-end TTS systems on two Arabic speech synthesis corpora, namely Arabic speech corpus (ASC) and Classical Arabic TTS corpus (ClArTTS).*

| System | Corpus | MOS | PESQ | MCD | Lf0RMSE | BAP | Speaker similarity |
|---|---|---|---|---|---|---|---|
| GroundTruth | ASC | 4.01 ± 0.1 | — | — | — | — | — |
| GroundTruth | ClArTTS | **4.39 ± 0.1** | — | — | — | — | — |
| Grad-TTS | ASC | 3.02 ± 0.2 | 1.48 | 6.38 | 12.25 | 1.14 | 0.51 |
| Glow-TTS | ASC | 3.19 ± 0.2 | 1.41 | 6.27 | 10.03 | 1.12 | 0.56 |
| Grad-TTS | ClArTTS | 3.63 ± 0.2 | **2.25** | 4.94 | 9.03 | 0.85 | 0.71 |
| Glow-TTS | ClArTTS | **3.84 ± 0.1** | 2.23 | **4.83** | **8.04** | **0.93** | **0.78** |

ule on text transcripts. We used the train set and test set as discussed in section 3.3 for training baseline TTS systems on ASC and ClArTTS. We trained the Grad-TTS[7] and Glow-TTS[8] systems individually on both corpora for 1000 epochs.

To synthesize the speech from the predicted Mel spectrograms, we opted for a Hi-Fi GAN-based neural vocoder [26]. The ASC and ClArTTS corpora have speech utterances with different sampling rates, 48000 Hz and 40100 Hz, respectively. Therefore, we trained two Hi-Fi GAN[9] neural vocoders to be compatible with the different sampling rates of both corpora. We used the V1 configuration of the Hi-Fi GAN neural vocoder for training both neural vocoders as detailed in [26]. We applied the short-time Fourier transform (STFT) with an FFT length of 1024, a hop length of 256, and a window size of 1024, and extracted Mel spectrograms using 80 Mel filters.

## 6. Evaluation and results

In Table 3, we present the performance of the two TTS systems using subjective and objective evaluations. We evaluated E2E TTS systems using a Mean Opinion Score (MOS) [27] based listening test. Each listener had to assign a score for synthesized speech utterance on a scale between 1 to 5 considering the intelligibility, naturalness, and quality of speech utterance. The speech utterances for subjective evaluation were selected randomly from test sets of ClArTTS and ASC. A total of 30 Arabic listeners participated in this MOS test and results are displayed in Table 3 with an associated 95% confidence interval. Furthermore, To validate the coherence of subjective listening test with objective evaluation, we opted for Perceptual Evaluation of Speech Quality (PESQ) [28] as an automated assessment of audio quality which takes into account various factors such as Audio sharpness, volume, background noise, lag in audio, clipping and audio interference. PESQ is computed on a scale from -0.5 to 4.5, where 4.5 represents the best similarity.

We used MCD (Mel Cepstral Distortion), an objective evaluation metric that measures the spectral distortion between the synthesized speech and the original speech signal. Lf0 RMSE (Root Mean Square Error of Log F0): an objective evaluation metric that measures the pitch accuracy of synthesized speech. BAP (Band Aperiodicity): an objective evaluation metric that measures the spectral envelope accuracy of synthesized speech. These evaluations are conducted by computing errors between reference speech utterances and synthesized speech utterances aligned using the dynamic time-warping algorithm. We selected a cosine distance-based speaker similarity score [29] to measure the consistency of the speaker's voice quality in synthesized speech. We utilized the pre-trained ECAPA-TDNN[10] based speaker embedding extractor to measure the similarity scores from synthesized speech and reference speech from the original speech synthesis corpus [30].

Table 3 shows that the ground truth samples of both corpora have higher MOS scores than the synthesized speech generated by the two TTS systems. The Glow-TTS system outperforms the Grad-TTS system in terms of MOS and PESQ scores for both corpora. The ClArTTS corpus has higher MOS scores and lower MCD, Lf0 RMSE, and BAP scores than the ASC corpus, indicating the ClArTTS corpus provides better synthesis quality. The speaker similarity scores of the synthesized speech are relatively low for ASC-based TTS systems, compared to the ClArTTS corpus counterpart. Thus, it shows that ClArTTS-based systems are better at retaining the speaker's voice characteristics in synthesized speech. The synthesized speech quality of baseline TTS systems can further be improved by using grapheme to phoneme system as additional pre-processing.

## 7. Conclusion

In this work, we presented the ClArTTS corpus, a single male speaker corpus extracted from a pre-recorded LibriVox audiobook. The ClArTTS corpus was developed to facilitate research in Arabic end-to-end TTS synthesis, which generally requires larger datasets compared to earlier models. The final corpus consists of a total of 12 hours and 10 mins of annotated speech, the largest single-speaker corpus freely available in the Arabic language. We illustrated the comparative advantage via corpus statistics compared to the available smaller corpora that were previously curated for the purpose of speech synthesis. In addition, we trained Glow-TTS and Grad-TTS systems using our ClArTTS corpus and compared the performance against the systems trained on the smaller Arabic Speech Corpus. Using both subjective and objective evaluations, our results indicate an overall better quality of synthesized speech using the ClArTTS corpus. The results also indicate that there is room for improvement, which we leave for future research. The ClArTTS corpus is now freely available for research purposes along with an interactive TTS demo at `www.clartts.com`.

Finally, we emphasize the caveat that the corpus is based on Classical Arabic text, which can be different from Modern Standard Arabic (MSA) in some aspects, such as lexical distribution. While many Arabic datasets liberally mix the two variants (e.g. the Tashkeela corpus contains text from both CA and MSA) as they seem indistinguishable, they do have differences that may prove to be consequential [31]. In our reported results, we included MSA utterances from ASC as well as utterances from the ClArTTS test set to reflect the potential quality in both of these variants. We leave any additional analysis of these differences for future work.

[7] https://github.com/huawei-noah/Speech-Backbones/tree/main/Grad-TTS

[8] https://github.com/jaywalnut310/glow-tts

[9] https://github.com/jik876/hifi-gan

[10] `https://github.com/microsoft/UniSpeech`

# 8. References


[1] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2017.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[4] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE, 1996, pp. 373–376.

[5] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0." *SSW*, vol. 6, pp. 294–299, 2007.

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *proceedings of INTERSPEECH*, 2017.

[7] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5679–5683.

[8] Z. Alyafeai, M. Masoud, M. Ghaleb, and M. S. Al-shaibani, "Masader: Metadata sourcing for arabic text and speech data resources," in *International Conference on Language Resources and Evaluation*, 2021.

[9] A. Ahmed, N. Ali, M. S. Alzubaidi, W. Zaghouani, A. A. Abdalrazaq, and M. J. Househ, "Free and accessible arabic corpora: A scoping review," *Computer Methods and Programs in Biomedicine Update*, 2022.

[10] N. Halabi, "Modern standard arabic phonetics for speech synthesis," Ph.D. dissertation, UNIVERSITY OF SOUTHAMPTON, 2016.

[11] A. Amrouche, A. Abed, K. Ferrat, K. N. Boubakeur, Y. Bentrcia, and L. Falek, "Balanced arabic corpus design for speech synthesis," *International Journal of Speech Technology*, vol. 24, no. 3, pp. 747–759, 2021.

[12] T. Lambert, N. Braunschweiler, and S. Buchholz, "How (not) to select your voice corpus: random selection vs. phonologically balanced." in *SSW*. Citeseer, 2007, pp. 264–269.

[13] R. Abdelmalek and Z. Mnasri, "High quality arabic text-to-speech synthesis using unit selection," *2016 13th International Multi-Conference on Systems, Signals & Devices (SSD)*, pp. 1–5, 2016.

[14] A. A. Shalaby, O. A. Dakkak, and N. Ghneim, "An arabic text to speech based on semi-syllable concatenation," *International Review on Computers and Software*, vol. 11, pp. 1178–1186, 2016.

[15] O. O. Khalifa, M. Z. Obaid, A. W. Naji, and J. I. Daoud, "A rule-based arabic text-to-speech system based on hybrid synthesis technique," vol. 5, pp. 342–354.

[16] I. A. Almosallam, A. Alkhalifa, M. Al-Ghamdi, M. I. Alkanhal, and A. Alkhairy, "Sassc: a standard arabic single speaker corpus," in *Speech Synthesis Workshop*, 2013.

[17] O. Zine and A. Meziane, "Novel approach for quality enhancement of arabic text to speech synthesis," *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 1–6, 2017.

[18] A. Amrouche, A. Abed, K. Ferrat, K. N. Boubakeur, Y. Bentrcia, and L. Falek, "Balanced arabic corpus design for speech synthesis," *International Journal of Speech Technology*, vol. 24, pp. 747 – 759, 2021.

[19] A. Abdelali, N. Durrani, C. Demiroğlu, F. Dalvi, H. Mubarak, and K. Darwish, "Natiq: An end-to-end text-to-speech system for arabic," *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pp. 394–398, 2022.

[20] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-fi multi-speaker english tts dataset," in *Interspeech*, 2021.

[21] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Interspeech*, 2008.

[22] T. Buckwalter, "Buckwalter arabic morphological analyzer version 1.0," in *Linguistic Data Consortium*, 2002.

[23] T. Zerrouki and A. Balla, "Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems," *Data in Brief*, vol. 11, pp. 147 – 151, 2017.

[24] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," *proceedings of International Conference on Machine Learning (ICML)*, 2021.

[25] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative Flow for text-to-speech via monotonic alignment search," *In proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2020.

[26] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2020.

[27] M. D. Polkosky and J. R. Lewis, "Expanding the MOS: development and psychometric evaluation of the MOS-R and MOS-X," *International Journal of Speech Technology*, vol. 6, pp. 161–182, 2003.

[28] A. W. Rix, J. G. Beerends, M. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752 vol.2, 2001.

[29] A. Kulkarni, V. Colotte, and D. Jouvet, "Analysis of expressivity transfer in non-autoregressive end-to-end multispeaker tts systems," in *Interspeech*, 2022.

[30] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech*, 2020.

[31] I. S. Alkhazi and W. J. Teahan, "Classifying and segmenting classical and modern standard arabic using minimum cross-entropy," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 4, 2017.