TunArTTS: Tunisian Arabic Text-To-Speech Corpus

Imen Laouirine, Rami Kammoun, Fethi Bougares
ELYADATA, ALGOBRAIN
imen.laouirine@elyadata.com
{rami.kammoun,fethi.bougares}@algobrain.ai

Abstract

Being labeled as a low-resource language, the Tunisian dialect has no existing prior TTS research. In this paper, we present a speech corpus for Tunisian Arabic Text-to-Speech (TunArTTS) to initiate the development of end-to-end TTS systems for the Tunisian dialect. Our Speech corpus is extracted from an online English and Tunisian Arabic dictionary. We were able to extract a mono-speaker speech corpus of +3 hours of a male speaker sampled at 44100 kHz. The corpus is processed and manually diacritized. Furthermore, we develop various TTS systems based on two approaches: training from scratch and transfer learning. Both Tacotron2 and FastSpeech2 were used and evaluated using subjective and objective metrics. The experimental results show that our best results are obtained with the transfer learning from a pre-trained model on the English LJSpeech dataset. This model obtained a mean opinion score (MOS) of **3.88**. TunArTTS will be publicly available for research purposes along with the baseline TTS system demo.

Keywords: Tunisian Dialect, Text-To-Speech, Low-resource, Transfer Learning, TunArTTS

1. Introduction

As the nomenclature implies, TTS is the ability of generating synthesized discourse from written text. Historically, TTS technology relied temporarily on classical methods that proved expensive in terms of data storage and often resulted in robotic-sounding output known as concatenative speech (Tan et al., 2021). However, the landscape of innovation began to shift in the early 2000s with the emergence of parametric models (Story, 2019). These models paved the way for a new era of TTS technology. In the late 2010s, neural network-based models burst onto the scene, revolutionizing the field in terms of naturalness, intelligibility, and prosody (Wang et al., 2017; Chen et al., 2020). Owing to advancements such as the attention mechanism (Bahdanau et al., 2014), models could now learn faster and more efficiently. Notably, newer state-of-the-art models introduced the ability to manipulate speech pitch and energy, further enhancing prediction capabilities (Ren et al., 2022; Tan et al., 2022).

There are mainly two possible approaches to build a TTS system: Cascade and fully End-to-End (Tan et al., 2021). The first approach consists in systems that are composed of two sequential modules: an acoustic module followed by a vocoder. The acoustic model aims to derive acoustic features (mel-spectrograms) from phonemes or characters. Those mel-spectrograms will be fed to a vocoder which will be responsible for waveform generation. The second approach is fully end-to-end, which is directly creating speech waveforms from the input sequence of characters or phonemes. One key advantage of this approach is the avoidance of cascade error propagation (Tan et al., 2021).

The research on TTS technology evolved naturally but not for all languages alike, a high-resource language like English saw a paradigm shift due to the amount of financial and physical resources allocated for research as visible in various products (e.g. Siri or Alexa) (Cambre et al., 2020). Consequently, such interest spiked the amount of research studies as well as the number of available English TTS models.
Contrary to the situation observed for English, TTS datasets and systems are less available for other languages including Modern Standard Arabic (MSA). The situation is even more pronounced for Arabic dialects where TTS systems are almost non-existent. The main contribution of this work can be summarized as follows:

- Releasing +3 hours of high-quality manually diacritized Tunisian Arabic speech[1].
- Comparing TTS results obtained using different training strategies.
- Releasing a baseline model for the Tunisian dialect TTS.

This paper is organized as follows: Section 2 presents the latest benchmarked datasets both in English and MSA that inspired our approach. In section 3, we detail the sourcing, collecting, and cleaning steps when assembling our dataset. Section 4 discusses the training approaches and the experiments we have tried in order to build our TTS system. The fifth section describes our results. Finally, we finish our paper by providing

[1] https://github.com/elyadata/TunArTTS

LREC-COLING 2024, pages 16879–16889

some discussions and concluding remarks in sections 6 and 7, respectively.

2. Related Work

Recently, TTS technology saw a natural development driven by big tech companies. Notably, a recent publication by Meta (Pratap et al., 2023) merits attention, as it reports their work on 1107 languages in TTS. However, none of the Arabic dialects, including the Tunisian dialect, appears to be included within their corresponding published inventory[2].

Prior to work done by Meta, there were multiple studies to train TTS systems for various languages. Despite these efforts, most of them cover only a small number of languages. Without a doubt, English is the most resourced language where multiple datasets are collected and annotated. For instance, LJSpeech (Ito and Johnson, 2017), considered to be one of the standard-bearer for TTS-oriented datasets in English and a source of inspiration in being a teacher model to low resource languages in multiple experiments (Baali et al., 2023; Fahmy et al., 2020; Kim et al., 2022; Jamal et al., 2022). It has been sourced from LibriVox and has a length of around 24 hours of a single-speaker female voice.

Despite being labeled as a low-resource language for so long, several research studies have been conducted for MSA, starting with Arabic Speech Corpus (ASC) by Nawar Al-Halabi (Halabi, 2016), the first Arabic dataset conceived principally for TTS. According to Halabi (2016), the source for ASC is the website Aljazeera Learn[3] since it contains fully diacritized text that helps in phonetization. In Halabi (2016), Nawar started from the buckwalter transliteration (Asgari-Bidhendi et al., 2012) in order to create his own phonetic representation (Halabi and Wald, 2016) suitable for speech synthesis.

With a total duration of almost four hours, ASC was for a long time the only available TTS dataset for the Arabic Language. Recently, two new datasets were introduced, namely Classical Arabic Text-to-Speech (ClArTTS) and QasrTTS. ClArTTS was introduced by Kulkarni et al. (2023) as the longest dataset assembled specifically for TTS. It consists of 12 hours of classical Arabic recorded by a single male speaker for LibriVox audiobooks.

QasrTTS (Baali et al., 2023) was part of the biggest dataset conceived for speech processing purposes (Mubarak et al., 2021). QASR is a large scale annotated speech corpus crawled from Aljazeera news channel and prepared by Qatar Computing Research Institute (QCRI). The corpus is suitable for training multiple NLP systems that do not include TTS. Lately, QASR TTS was created as a TTS corpus for a single speaker after applying multiple cleaning criteria. However, due to its limited duration, only one hour, it was not feasible to build a high-quality TTS system entirely from scratch. As a solution, Baali et al. (2023) decided to use pre-trained models on LJSpeech as a starting point and then fine-tune them with the QasrTTS dataset.
Our work is partially motivated by Baali et al. (2023) who successfully used transfer learning and showed an intelligibility score of 4.4 out of 5 and a naturalness score of 4.2 out of 5 using a limited annotated dataset of 1 hour and a half of a mono-speaker spoken speech.

3. Dataset Creation

Looking at today's literature, the absence of research endeavors focused on TTS systems for the Tunisian dialect is conspicuous. This was mainly due to the lack of TTS training data for the Tunisian dialect. In the following section, we will briefly describe the peculiarities of the Tunisian dialect and detail how we created a labeled dataset which includes speech audio paired with corresponding text in the Tunisian dialect.

3.1. Tunisian Dialect

The Tunisian dialect is heavily influenced by the Arabic language, it is a variety of dialectal Arabic, used mainly for daily spoken communication in Tunisia. Historically, the North African region has seen a diverse set of communities taking turns in living on its lands. Consequently, Tunisian is composed, mainly, of MSA, Tamazight, then on a third degree, of English, French and Turkish with some foreign integrated words adding to this versatility such as Italian and Maltese (Masmoudi et al., 2018).

The aforementioned multiculturalism contributed to the complexity on the phonological, morphological, syntactic, and lexical level. 1) Phonologically with the introduction of foreign vowels such as /P/, /V/ and /G/ (Masmoudi et al., 2018) or with the pronunciation of some consonants like /j/ which is incorporated in some cases as /z/, e.g. the word جزار *jaz∼aAr* 'butcher' is pronounced as /jazza:r/ and /zazza:r/ (Zribi et al., 2014). 2) Morphologically, the Tunisian dialect introduced new clitics and negated others, such as replacing interroga-

[2] https://dl.fbaipublicfiles.com/mms/misc/\language_coverage_mms.html

[3] https://learning.aljazeera.net/en

tion clitics of أ *Âa* or particle هل *hal*, with شي *ŝiy*, (Zribi et al., 2014). 3) Syntactically, Tunisian still follows MSA but without its particularities (Mejri et al., 2009) especially, the emission and merge of some MSA pronouns, from twelve to seven. A concrete example of this is the vanishing of the feminine plural of both second and third persons in some areas, and the duality of the second person which was unified for the masculine You. 4) Lexically, several words are borrowed from Tamazight such as 'فَكْرُونْ' pronounced /fakrOn/ (turtle in English) and from Italian such as 'كوجينة' pronounced /Kuwjiynah/ ('Kucina' in Italian which means 'Kitchen' in English).

Just like all the Arabic dialects, Tunisian does not have an official codified writing system. To remedy this problem, Habash et al. (2019) presented a newly developed research work called Conventionalized Orthography for Dialectal Arabic (CODA) that aimed at proposing orthographic conventions for various Arabic dialects. CODA is under constant research and development. However, it struggles in defining one unified orthographic representation for Tunisian seeing the amount of complexity the latter dialect is derived from Zribi et al. (2014). Due to the unavailability of automatic methods for rendering the dataset CODA-compliant, the amount of manual annotation work needed postponed the idea for future work.

3.2. TTS Dataset Specificities

As discussed above, TTS exaggerates certain constraints to assembling a dataset that guarantees high performance of the TTS system. Below, the list of points to be taken into consideration (Masri and Za'ter, 2022; Bakhturina et al., 2022; Puchtler et al., 2021):

- **Sample Rate**: In TTS, audio recordings need to have a sampling rate equal or higher than 22050 Hz.
- **Alignment**: Every phoneme in the audio has to be aligned with its transcription.
- **Duration**: The duration of the audio recordings needs to be between 2 and 10 seconds.
- **Spelling**: Transcripts need to be grammatically correct and coherent.
- **Normalization**: Symbols and abbreviations need to be written alphabetically.
- **Diacratization**: In Arabic-written dialects, it is necessary to use diacritics in writing in order to minimize ambiguities.
- **Reading Speed**: Consistency in tempo should be upheld throughout the recording process to ensure a high level of uniformity.
- **Tone**: The recordings should feel warm to the listener and neither angry nor confrontational.
- **Pronunciation**: In Arabic, it is crucial to pronounce velarized consonants properly.
- **Background Noise**: Unlike ASR, background noise hurts a TTS system's performance deeply.
- **Silence**: Silence hurts phoneme alignment when training, it needs to be eliminated.
- **Channels**: Audio recordings need to be mono-channeled waveforms.
- **Coverage**: Audio recordings should cover all language phonemes.

3.3. Data Sourcing

As established, the Tunisian dialect is a low resource language. It was challenging to find a publicly available speech with the corresponding transcription. A possible approach is to collect a dataset from Tunisian YouTube videos, manually transcribe, annotate and train using this dataset. However, no clear and consistent videos of one single speaker without any background noise or any other sort of TTS dataset deficiencies were found. Not forget to mention that it was also difficult to extract mono-speaker audios from the videos. After a substantial research and inspection of existing audio resources in the Tunisian dialect, an online English and Tunisian Arabic dictionary[4] that contains Tunisian words and sentences along with their corresponding audio recordings was identified. These data are published by *Derja.Ninja* under the licence CC- BY-SA 4.0. It was, therefore, decided to use them to build the first TTS dataset for the Tunisian dialect.

3.4. Data Collection

The data was collected from the Derja Ninja website. A Python script was developed to harvest the audio files and the corresponding transcriptions. More specifically, *Derija.Ninja* was built around the Tunisian dialect terms, that is, for each term there is an audio of the term and another audio of that term in a context (i.e. sentence). Overall, a multi-speaker dataset with more than 44 hours of audio was collected with four different sample rates: 8KHz, 16KHz, 44KHz, and 48KHz. For the reasons cited above (section 3.2), only audio files with 44KHZ and 48KHz were kept which represent a subset of 15.5 hours of multi-speaker speech.

[4] https://derja.ninja/

Starting from this subset, the speaker with the most abundant spoken data with 4 hours with a sample rate of 44KHz was identified and chosen.

3.5. Audio Cleaning

Once the Tunisian dialect mono-speaker dataset was identified, the audio cleaning phase started. In fact, the original audio files were retrieved in mp3 format. As a first step, they were converted to a mono channel wav format. The recording was designed in a way that the speaker spells a unique ID at the beginning of each audio, followed by the term repeated three times, then, the sentence repeated two times. Consequently, the audio part of the unique ID was trimmed out using the available information related to the start time of each term. The silence was removed with a defined threshold of 55dB.

3.6. Text Cleaning

After cleaning the audio recordings, it turned out that some of them didn't respect the template mentioned in 3.5 (term-term-term sentence-sentence). Given this, while listening to all the audio segments, the following four different forms were discovered:

1. term-term sentence-sentence.
2. term sentence-sentence.
3. sentence-sentence.
4. term-term-term.

In order to obtain a paired speech audio with the correct corresponding text, all the transcriptions in the corpus were manually and thoroughly corrected once they were identified to be following one of the forms mentioned above. Numbers were converted to words and special characters and segments containing ڤ , ڨ , پ were deleted to avoid unexpected issues while converting them to phonemes.
During this text cleaning stage, some misalignment between the audio files and their texts were singled out on the website. That is for some audio files the corresponding text is not the content uttered by the speaker. Therefore, the latter were identified and manually corrected. Table 1, shows some examples with the text before and after correction.

3.7. Diacratization

Diacritics are generally omitted in Standard Arabic and its dialects. In fact, native speakers can recognize the correct pronunciation for a text without diacritics. However, the presence of the diacritic marks is essential for the implementation of a TTS system. Like most online dialectal content, the collected data in the context of this work does not contain diacritics. Diacritizing a dialectal written text is a very challenging task. The diacritized text should match the exact speakers' pronunciation of words even if they are not grammatically correct or borrowed from other languages. Also, a particular attention was paid to the difference between MSA and the Tunisian dialect in order to keep the dialectal peculiarity. For example, the word *she* in the Tunisian dialect : هِيَّ /hiyya/ is written in MSA without gemination (the shada in Arabic : ّ). After listening to the audio recordings, it was decided to write it with gemination. For other cases, it was decided to follow the cognate spelling instead. For example, the term *"from thieves"* is pronounced as مِسُّرَّاقْ /messerrek/ and it was kept written as مِنْ السُّرَّاقْ /men elsorrek/ as its MSA form. Table 2 shows an example showcasing the difference between text with and without diacritics from the corpus.

Original:	ايجا كركر معايا الفرش
CODA:	ija karker m3aya lfarsh.
EN:	Come and pull with me the mattress.
Corrected :	ديما يكركر في الخدمة ياخي طردوه
CODA:	dima ykarker fl lkhedma yekhi t.ardouh.
EN:	He is always slagging at work so he got fired.

Table 1: Example of the identified errors in the dataset. **Original** is before correction while **Corrected** represent the exact text uttered. **CODA** is the CODA transliteration.

Segment (1) :	تعود على الريتم اذاكا
CODA of (1):	t3wd 3la 2eritm 2dheka
Diacritized (2):	تْعَوِّدْ عَلَى الرِّيتِمْ أَذَاكَا
CODA of (2):	t3awwed 3ala 2erritem 2adheka
En:	He got used to that rhythm.

Table 2: Example of a Tunisian dialect text with its transliteration before and after diacritization along with the English translation.

The first sample in Table 2 is a fully dialectal sentence, whereas, the second one is an example where there is a loanword from French: The word الريتم (rhythm in english) is borrowed from the French word *rythme* with an adaptation of its pronunciation (Oueslati, 2021).

3.8. Grapheme to Phoneme Conversion

With the goal of capturing the detailed differentiation between Arabic phonemes, Halabi (2016) proposed to transform the diacritized Arabic alphabets to its phonemicized representation. The

same steps[5] were applied which are consisting of, first, converting the Arabic alphabets to the Buckwalter transliteration, then, applying the system that takes into consideration the position of each consonant with its corresponding clitic and some predefined words that represent a grammatical exception.

Following Halabi (2016), the "*sil*" symbol was also added at the end and beginning of each transcript. It has the purpose of better aligning each phoneme with its corresponding representation and marks the silence with "*sil*". Moreover, given that our audio recordings are characterized with breaks between terms and sentences, it was sought better to add "*sil*" in between them to avoid hurting the model's performance with the inexorable silence.

3.9. Corpus Statistics

Table 3 provides some statistics about our collected dataset named TunArTTS.

Count	TunArTTS
Total Segments	1493
Total Words	20925
Total Phonemes	115966
Total Characters	113221
Total Duration	3 hours and 32 secs
Mean Clip Duration	7.24 secs
Min Clip Duration	3.11 secs
Max Clip Duration	16.3 secs
Mean Words per Clip	14.015
Distinct Words	4491
Distinct Phonemes	76

Table 3: TunArTTS Corpus statistics.

Since TunArTTS was not recorded specifically for the purpose of speech synthesis, it may not include all possible phonetic combinations of the Tunisian dialect. As shown in Table 3, TunArTTS includes 76 different phonemes distributed over around 21k words. The scripts needed to reproduce the dataset preparation described in this work are provided[6]. Table 4 reports the dataset distribution over train, dev and test set.

4. TTS Systems

As can be seen in Figure 1, two different approaches were tried on to train the Tunisian TTS systems: (1) a pipeline consisting of a trained from scratch acoustic model cascaded with a vocoder and (2) a pipeline consisting of a fine-tuned acoustic model from a pre-trained one cascaded with a vocoder. Regarding the first approach, the acoustic model was trained using Tacotron2 (Shen et al., 2018) which is a seq2seq model composed of an encoder, local sensitive attention layer and a decoder. The second approach follows the transfer learning fashion, where pre-trained acoustic models were fine-tuned using the TunArTTS dataset. The vocoder stage is common for both approaches.

Split	Duration	# seg	Mean Words per Clip
Train	2h:47 mins	1384	13.94
Dev	8 minutes	65	14.27
Test	6 minutes	44	15.5

Table 4: Dataset splits of the TunArTTS Corpus

4.1. Vocoder Training

For the vocoder, as an initial step, three different architectures were used and compared, namely, MelGAN (Kumar et al., 2019), HIFIGAN (Kong et al., 2020), and Parallel WaveGAN (Yamamoto et al., 2020). To assess the perceived speech quality, 15 random samples generated by each vocoder had been chosen.
Parallel WaveGAN was selected since it gave the clearest samples. Henceforth, Parallel WaveGAN was adopted in the rest of the experiments. It was trained on TunArTTS dataset following the LJSpeech recipe and the ESPnet compatible Pytorch Parallel WaveGAN implementation[7].

4.2. Acoustic Model Training

The acoustic model was trained from scratch following the ESPnet's LJSpeech recipe[8]. The training step is preceded by the Grapheme to Phoneme (G2P) conversion step as detailed in section 3.8. The configuration followed multiple experimentations by varying the binary cross entropy weights (5, 15 and 20) and the reduction factor (1, 3 and 5) which controls the number of output frames at each time step. Also, phonetic and character sequences representations were attempted, that is with or without using the G2P conversion stage. Table 5 presents the hyper-parameters that led to the best results.

4.3. Transfer Learning

Transfer learning is the alternative to training from scratch presented in the previous section. It has

[5] https://github.com/nawarhalabi/Arabic-Phonetiser

[6] https://github.com/elyadata/TunArTTS/

[7] https://github.com/kan-bayashi/ParallelWaveGAN

[8] https://github.com/espnet/espnet/tree/master\/egs2/ljspeech/tts1

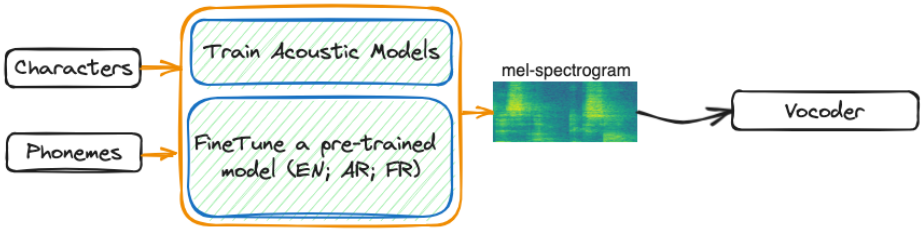


Figure 1: Training approaches of TTS systems. Each approach is a pipeline of (1) an acoustic model that generates mel-spectrograms from a given character or phoneme inputs, followed by a (2) vocoder. The acoustic model can be either trained from scratch or fine-tuned from an existing model.

Hyper-parameter	Value
Reduction factor	5
bce_pos_weight	20
Learning Rate	0.0005
Number of epochs	250
Optimizer	adam
Training unit	Phonemes
Sample rate	44100Hz

Table 5: Hyper-parameters for the acoustic model training experiment.

been proved to enhance various speech processing systems, including Speech-To-Text (Hori et al., 2017) and Text-To-Speech (Xu et al., 2020). The goal is to start from an existing pre-trained model in another language, or, if unavailable, train a TTS model instead using a larger dataset to achieve better performance. In this work, several pre-trained models on various languages were experimented on, namely: English, French, and Arabic. As presented in Table 6, a pre-trained model on the English LJSpeech dataset was sourced and used. In contrast, they were non-existent for French and Arabic, hence, a model had to be trained using the datasets presented in Table 6.

Lang	Corpus	Sample Rate	Size
English	LJspeech	22050Hz	24 h
French	SIWIS	24000Hz	10 h
Arabic	ClArTTS	40100Hz	12 h
	ASC	48000Hz	04 h

Table 6: Datasets used for pre-training TTS models for Transfer Learning scenario.

(1) Pre-trained English TTS: Transfer learning on English has proved its efficiency and been used in several previous works (Fahmy et al., 2020; Durrani and Arshad, 2021). For instance, ESPnet offers a range of pre-trained models in English that can be used in a transfer learning approach (Baali et al., 2023). In our experiments, TransformerTTS[9] (Li et al., 2019) and Tacotron2[10] pre-trained models were used. Both models have a reduction factor of 3, the highest to be found. Consequently, the student models need to have the same reduction factor. These models were fine-tuned following the procedure presented in Baali et al. (2023). During experimentation, both character and phonetic-based representations of the Tunisian text were tried on. After settling the reduction factor to 3, various learning rates were tested and the best results were obtained with Tacotron2 using a learning rate of 0.001. Table 7 shows the best set of hyper-parameters obtained starting from Tacotron2 pre-trained on English. Compared to LJSpeech, TunArTTS has a higher sample rate (44100 Vs. 22050 Hz), we experimented with and without down-sampling it to match LJSpeech's sample rate.

Hyperparameter	Value
Reduction factor	3
bce_pos_weight	20
Learning Rate	0.001
Number of epochs	120
Optimizer	adam
Type of training	Phonemes
Sample rate	22050Hz

Table 7: Hyper-parameters for the transfer learning on English experiment.

The obtained model was used to guide a Fastspeech2 (Ren et al., 2022) based Conformer (Gulati et al., 2020) model. Fastspeech2 is a non autoregressive acoustic model known to fix robotic sounds and misalignment which leads to a better synthesis quality and faster inference. During this experiment, values in Table 7 were kept except the learning rate and the number of epochs which have been set to 1 and 1000 respectively.
In view of the findings set out above, the rest of the work was carried forward using a phonetic-level

[9] https://zenodo.org/record/4643685
[10] https://zenodo.org/record/4643683

Tacotron2 model.

(2) Pre-trained Arabic TTS: Tunisian is one of the Arabic dialects. Thus, it makes sense to experiment transfer learning starting from a pre-trained model on Arabic. Since there was no compatible pre-trained models for Arabic, a similar configuration to Table 5 was used to train two acoustic models from scratch using ASC and ClArTTS datasets. These pre-trained models were intended to be used to apply transfer learning on TunArTTS. The acoustic model trained on ASC corpus was shown to be significantly better than ClArTTS model. Therefore, we proceeded by fine-tuning the ASC-based model and using similar hyper-parameters chosen for the transfer learning presented in Table 7 except the reduction factor, which was set to 5 and the sample rate which was set to 44100Hz.

(3) Pre-trained French TTS: Code-switching (CS) between French and Arabic is common in daily speech of Tunisian speakers. That being so, the next step is to experiment Transfer learning from a French pre-trained model.
As it was the case with the Arabic, a Tacotron2-based acoustic model was trained on the SIWIS[11] (Honnet et al., 2017) dataset using the aforementioned hyperparameters in Table 5. Once the French TTS model was trained, it was fine-tuned on TunArTTS. During this transfer learning experiment, the same set of hyper-parameters used for transfer learning from Arabic pre-trained model was kept. The exception to this is reducing the sample rate to 24000 Hz in order to match that of the SIWIS dataset.

5. Results and Analysis

The developed models were evaluated and tested using the aforementioned test set. Two methods of evaluation were mainly used: (1) Objective evaluation with Mel Cepstral Distortion (MCD), Character Error Rate (CER), and Word Error Rate (WER) and (2) Subjective evaluation, with the Mean Opinion Score (MOS) being the most used metric.

5.1. Objective Evaluation

MCD is inspired by Cepstral Distance (CD) (Kubichek, 1993). It measures the spectral distance between synthesized and ground truth speech by comparing the Mel-frequency cepstral coefficients of the two signals. WER was originally designed to measure the accuracy of Automatic speech recognition (ASR) systems. It calculates the rate of incorrect word predictions, or character prediction for CER, compared to a reference text. The lower the WER, the higher the accuracy of an automatic recognition system. In this work, the WER and CER were used to evaluate the intelligibility of the synthesized speech. That is, after generating the waveforms using a TTS system, it is decoded with a Tunisian dialect ASR system[12]. The WER and CER are then calculated between the ASR output and the reference (undiacritized input text used as input to the TTS).
Table 8 shows the results of the objective evaluation. With a WER of 76%, a CER of 40.6%, and an MCD of 10.43, the transfer learning (TL) experiment on French (TL from French row) is barely intelligible, this is reflected on the quality of synthesized speech where the system is barely pronouncing phonemes. For the transfer learning on Arabic experiment (TL from Arabic row), the experiment has an MCD score of 6.86, a WER rate of 43.6% and a CER of 15.6% where the system could pronounce slightly better some non-velar phonemes. Yet, both results are deemed robotic.

Experiment	MCD ↓	WER(%)	CER(%)
Ground Truth	–	34.4	9.21
From Scratch	5.95	40	11.04
TL from English	5.53	35	9.41
TL from Arabic	6.86	43.6	15.6
TL from French	10.43	76	40.6

Table 8: Comparison of objective evaluation results on the TunArTTS test set.

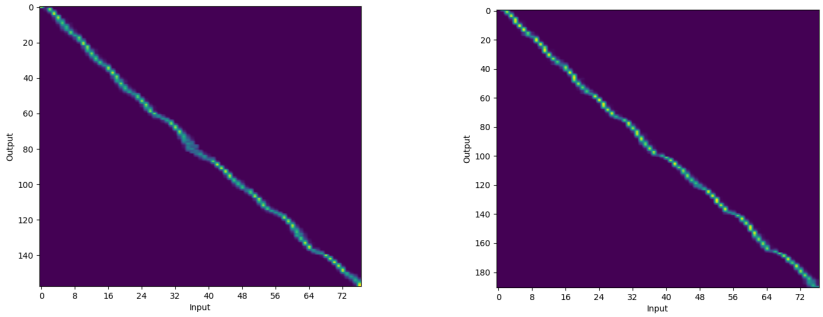
Table 8 states that the experiment of transfer learning on English (TL from English row), performed better than training from scratch by 0.42 points on the MCD metric, 5% and less than 2% on the WER and CER metrics respectively. In addition, transfer learning from English model shows a very close WER, and CER, compared to the Ground Truth.

5.2. Subjective Evaluation

MOS, used for subjective evaluation, is based on the simple concept of averaging the judgements from N assigned listeners. The listeners evaluate the quality, including naturalness and intelligibility, of synthesized speech using a rating scale ranging from 1 to 5 where 1 means "Bad" and 5 means "Excellent" (Kim et al., 2022). In this work, 20 native speakers (7 males and 13 females) were brought to assess 20 audio samples chosen randomly from the test set. Due to the expensive process of MOS evaluation, only the two best systems according to the objective evaluation were evaluated, namely the trained from scratch and transfer learning on English models. In addition to the outputs from these two models, listeners had access to the ground truth as reference. The results are

[11] https://datashare.ed.ac.uk/handle/10283/2353

[12] https://huggingface.co/spaces/SalahZa/Tunisian-Speech-Recognition



(a) TTS system trained from scratch

(b) TTS system trained by TL from English

Figure 2: Comparing attention plots of two TTS systems of the text: Phoneme: t f A qq I1 d t f A qq I1 d t f A qq I1 d t f A qq I1 d < a l x A dd aa m a f ii0 < a l m a E m i1 l t f A qq I1 d < a l x A dd aa m a f ii0 < a l m a E m i1 l

detailed in Table 9. As can be seen, the transfer learning on English outperformed the trained from scratch experiment by 1.02 MOS points. The latter wide disparity is not reflected in the objective evaluation due to the perceptive nature of the MOS evaluation.

Experiment	MOS
Ground Truth	4.78
From Scratch	2.86
TL on English	3.88

Table 9: Comparison of subjective evaluation results on the TunArTTS test set.

Overall, the best Tunisian TTS model achieved a MOS of 3.88. Such performance is comparable to Standard Arabic system trained using the 12 hours of speech of ClArTTS (Kulkarni et al., 2023) sampled at 40100 kHz.

5.3. Attention Plots Analysis

In addition to the objective and subjective evaluations, the attentions plots were visually inspected. Figure 2 shows the attention plots of a given sentence, when training a TTS model from scratch (Figure 2a) compared to the transfer learning fashion (Figure 2b). As the attention plots show, the attention distribution is sharper with more connected alignment paths for the transfer learning experiment. This confirms results presented in Table 9 where the TL on English TTS model gives better MOS score (3.88) compared to training TTS from scratch (2.86).

6. Discussion

The dataset created as part of this work was extracted from a website created as an English and Tunisian Arabic (Derja) dictionary. This website contains 17k entries with example sentences and their audio pronunciations. Even though the dataset was not designed for the purpose of speech synthesis, we were able to use it to train a TTS system that gives an intelligible synthesised speech. Although initial results of TTS systems trained using TunArTTS are encouraging, some annotation improvements are possible by targeting for example the French words which are harder to annotate in Arabic letters while preserving the nature of their tone such as /é/ and /è/.

7. Conclusion

We presented TunArTTS, the first annotated dataset for TTS in the Tunisian dialect. We presented how we collected, cleaned and annotated this dataset. TunArTTS dataset was used to train various TTS systems based on an end-to-end framework that combines a Tactoron2 acoustic model and Parallel WaveGAN as a vocoder. Trained systems were evaluated using a subjective metric, MOS and objective ones, MCD, WER and CER. The best system was achieved by fine-tuning on the TunArTTS of a pre-trained model on the English LJSpeech dataset. Overall, a MOS score of 3.88 was reached. In future work, we plan to extract and annotate audio files of other speakers from the same source of data. This will be done towards building a multi-speaker Tunisian dialect TTS system. We also intend to manually revise the text to be CODA-compliant. We believe that this will generate more consistent phonetic representation and improve the TTS quality.

8. Ethical Considerations

Ethical considerations were a constant matter in our research. For the results generated by our system, we have no intentions in impersonating or cloning someone else's voice or identity. Our TTS system was trained using a voice of a native speaker from Tunis. Therefore, the system is unable to produce sentences in other sub-dialects'

intonations even if provided with the adequate vocabulary. We also acknowledge that the only data available was of a male voice. This will be targeted in future works where we intend to guarantee genders equilibrium.

9. Bibliographical References

Majid Asgari-Bidhendi, Behrouz Minaei, and Hosein Jozi. 2012. Extracting person names from ancient islamic arabic texts.

Massa Baali, Tomoki Hayashi, Hamdy Mubarak, Soumi Maiti, Shinji Watanabe, Wassim El-Hajj, and Ahmed Ali. 2023. Unsupervised data selection for tts: Using arabic broadcast news as a case study.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. 2022. A toolbox for construction and analysis of speech datasets.

Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Miloš Cerňak and Milan Rusko. 2005. An evaluation of a synthetic speech using the pesq measure.

Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, Tao Qin, and Tie-Yan Liu. 2020. Multispeech: Multi-speaker text to speech with transformer.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.

Sara Durrani and Muhammad Arshad. 2021. Transfer learning based speech affect recognition in urdu.

Fady Fahmy, Mahmoud Khalil, and Hazem Abbas. 2020. A transfer learning end-to-end arabictext-to-speech (tts) deep architecture.

Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition.

Nizar Habash, Fadhl Al-Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Alshargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2019. Unified guidelines and resources for arabic dialect orthography.

Nawar Halabi. 2016. *Modern standard Arabic phonetics for speech synthesis*. Ph.D. thesis, University of Southampton.

Nawar Halabi and Mike Wald. 2016. Phonetic inventory for an arabic speech corpus.

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. 2017. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm.

Sahar Jamal, Sadaf Abdul Rauf, and Quratulain Majid. 2022. Exploring transfer learning for Urdu speech synthesis. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 70–74, Marseille, France. European Language Resources Association.

Minchan Kim, Myeonghun Jeong, Byoung Jin Choi, Sunghwan Ahn, Joun Yeop Lee, and Nam Soo Kim. 2022. Transfer learning framework for low-resource text-to-speech using a large-scale unlabeled speech corpus. In *Interspeech 2022*. ISCA.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmonem Mohammad Shatnawi, and Hanan Aldarmaki. 2023. Clartts: An open-source classical arabic text-to-speech corpus.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and Ming Zhou. 2019. Neural speech synthesis with transformer network.

Abir Masmoudi, Fethi Bougares, Mariem Ellouze, Yannick Estève, and Lamia Belguith. 2018. Automatic speech recognition system for tunisian dialect. *Language Resources and Evaluation*, 52.

Hala Al Masri and Muhy Eddin Za'ter. 2022. Arabic text-to-speech (tts) data preparation.

Salah Mejri, Mosbah Said, and Inès Sfar. 2009. Pluringuisme et diglossie en tunisie. *Synergies Tunisie*, 1:53–74.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. Qasr: Qcri aljazeera speech resource – a large scale annotated arabic speech corpus.

Jamila Oueslati. 2021. French loans in tunisian arabic from phonetic and phonological perspective. *Rocznik Orientalistyczny/Yearbook of Oriental Studies*, T. LXXIV(No 1):95–113.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

Pascal Puchtler, Johannes Wirth, and René Peinl. 2021. Hui-audio-corpus-german: A high quality tts dataset.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2022. Fastspeech 2: Fast and high-quality end-to-end text to speech.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.

Brad Story. 2019. *History of speech synthesis*, pages 9–33.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu. 2022. Naturalspeech: End-to-end text to speech synthesis with human-level quality.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

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.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP*, pages 6199–6203. IEEE.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. A conventional orthography for Tunisian Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2355–2361, Reykjavik, Iceland. European Language Resources Association (ELRA).

10. Language Resource References

Honnet, Pierre-Edouard and Lazaridis, Alexandros and Garner, Philip N. and Yamagishi, Junichi. 2017. *The SIWIS French Speech Synthesis Database – Design and recording of a high*

quality French database for speech synthesis. Idiap-RR-03-2017.

Keith Ito and Linda Johnson. 2017. *The LJ Speech Dataset*.