Machine Learning for Arabic Text To Speech Synthesis: a Tacotron Approach

A.M. Mutawa

Kuwait University, Box 5969, Safat, 13060, Kuwait

Abstract

A Text-to-Speech (TTS) method converts a language’s standard utterance to speech. In contrast, other systems provide symbolic linguistic representation, such as phoneme transcriptions into speech. Synthesized speech is generated by bringing together units of recorded speech that are saved in a repository. The speech units, which come in diphones or tablets, differ in size depending on the system. Even though this provides the most comprehensive performance range, it can also require more clarity. The storage of whole words or sentences in explicit usage areas is the foundation of high-quality production. There have been pre-trained end-to-end TTS systems applied to other languages. In this work, we study the use of an end-to-end Tacotron when applied Arabic text.

Arabic is morphologically rich and vague, and it has many dialects that vary significantly from one another. There are no official spelling norms in the dialects, and uncorrected standard Arabic includes numerous spelling and grammar errors and hence is a very challenging problem.

Keywords 1

Arabic Text to speech, Tacotron, Machine Learning, end-to-end

1. Introduction

Speech is the most common means of human expression, and Speech Technology is rapidly becoming the most prevalent mode of knowledge delivery today [1-2]. It is more intuitive to communicate with computers through speech rather than pressing buttons, for example. Text to speech serves as a link between humans and machines, so it’s critical to build a robust and dependable framework. Building such structures necessitates keen observation and a thorough understanding of all speech and language technology [3]. This will include a significant description of the human speech development and interpretation process, in addition to the exciting uses that the TTS method itself promises, and has been proven to be the best way to explain cognitive ability.

It is challenging to construct a wide-scale text-to-to-speech infrastructure that leaves the language rules uncertain. Individuals can say the same things but have different degrees of significance. This fact alone highlights that while dialects can be used to differentiate areas, the same names can also constitute different accents, depending on where a person speaks with the local accent. Expansion: So, where two people use the same set of words in two languages, the sound, and inflection of such words differ in such a manner that two regions can use those languages. This takes us to the issue of whether men’s and women’s voices are distinct [4]. Often, we can make use of sound to talk and think about our thoughts. Also, the accessibility has additional requirements for collecting data to allow them to fully manage both of these varieties. Most speech synthesis systems fail to correctly generate speech data while supplying text data on a dialect or accent, leaving them unable to send a text with a certain validity or a particular sound. A few apps use multiple voices to provide users with a known pronunciation and accent sounds as an alternative to common words. Despite the current development

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine

EMAIL: dr.mutawa@ku.edu.kw (A.M. Mutawa)
ORCID: 0000-0002-5707-2692 (A. M. Mutawa)

efforts in NLP to develop systems, the representation of text in speech still persists, suggesting that speech technology needs to focus on more general-purpose applications for larger-scale use. Even as more use was found for multidimensional speaker divergence, no single learning methods could effectively produce language-invariant expression while using several dimensions. much like other kinds of speech synthesizers, it often becomes impossible not to use when faced with the quotidian issues of removing robotic voices

To use a machine learning approach, it is assumed that a few steps in the text-to-to-speech algorithm are taken into account. Datasets are then harvested. The model will understand the knowledge gathered from The testing is completed on the datasets for the TTS method. Then it goes on to the research process. During this step, both the audio files that correspond to the model's outputs and the text files make the model processed. The first step in text function expansion is to collect linguistic and phonetic properties, including details about the current phoneme or expression.

Normalization of the text is the first step of text analysis. Normalization consists of segmenting the text into titles, sentences, paragraphs, and so on. These are again classified into units such as formed words or words. After that, to clean the text, preprocessing is applied to these elements. So all ambiguities concerning terms are eliminated to give a pronounceable structure for phonetization [5]. The preprocessing shows the task as easy in languages like English since its features can convert the capitalized form. But for languages like Arabic, the absence of capitalized words or rules for punctuation makes the preprocessing task a little complicated. Part of Speech tagger (POS) is the second step. It adds grammar like a verb, subject, and object to every cleaned section and then splitting it into lexemes such as suffix, derivative, and prefix.

However, this development on TTS conversion is restricted from changing some languages yet remains an active topic where numerous works are accomplished by developing technology to change TTS models indistinguishable from human capacity [6]. The letter sets of the Arabic language are made out of 28 letters. These 28 characters are grouped based on the articulation points in the human vocal system, as shown in Figure 1 below. The Arabic language is known to have unique personalities that sound from different parts of the vocal system. There are two characters from the nasal, ten characters from the plosive, one character is from the trill, fourteen charactrs from the fricative, and three characters from the approximant as seen in Table 1 [7-8].

Table 1: Arabic character grouping based on the different articulation points.

	Bilabial	Labiodental	dental	Dento-	Post-alveolar	Palatal	velar	Pharyngeal	Glottal
Plosive	ب		ض د ط ت				ق ك		أ ء ئ
Nasal	م			ن					
Trill				ر					
Fricative		ف	ظ ذ ث	س ص ز		ج ش	غ خ	ح ع	ه
Appr ximant	و			ل		ي			

Every one letter in the Arabic alphabet is a consonant, three of them (ي ، و ،ا) may show up as consonants or long vowels as indicated by their context. All remaining letters in the Arabic alphabet can act as consonants or short vowels if one of the three diacritics (َ، ِ، ُ) is placed on the character. Traditional Arabic is usually written without diacritics which adds a new level of complexity to the mixture. Diacritics in the Arabic language are only present in books for primary peoples or Arabic for non-Arab speakers; Arabic text is naked from diacritics in a standard form used in public media or on the internet. Figure 2 depicts the challenges when Arabic text is stripped off diacritics, then all short vowels are removed from the text. This is similar to writing “bed and breakfast” with no vowels as “bd and brkfst” then asking the TTS to guess the missing short vowels as a preprocessing step then to produce the vocal-based on the assumed best fit when compared with a dictionary entry. Additional complexity is when the same character set can have more than one valid entry based on different diacritics placement, such as in Figure 2. The three characters “W,” “L” and “D” followed with different set of vowels such as “WaLaLd” or “WuLiDa” will have two valid but opposite meaning; the first word means “to give birth,” but the second means “was born”. Hence permutation of vowels following characters can generate different meanings for the same word. Hence the preprocessing must be smart enough to consider the full semantic of the sentence based on the context of the word that will add a new layer of preprocessing to the TTS system. This layer can be very useful when implemented in other languages so that TTS systems can tolerate miss-spelled or missing characters from the words. Also will have a more broad meaning of the text; instead of reading Dr. as “D R” I will read it as “Doctor.”

Building a solid synthesis structure for the Arabic language opens the door for others to expand it in various ways. The proposed method would produce an expression that strongly resembles the speaker features used to create the system in the first place, involves feelings, and is adaptable to new unseen text.

In this work, we study the application of using an end-to-end system Tacotron [9-13] on the Arabic text to synthesize speech.

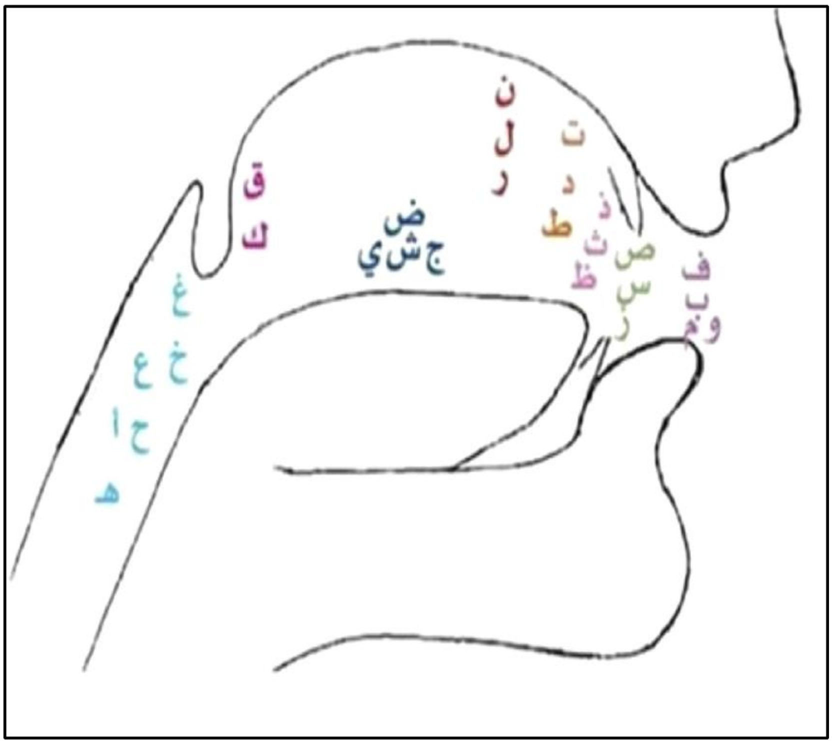
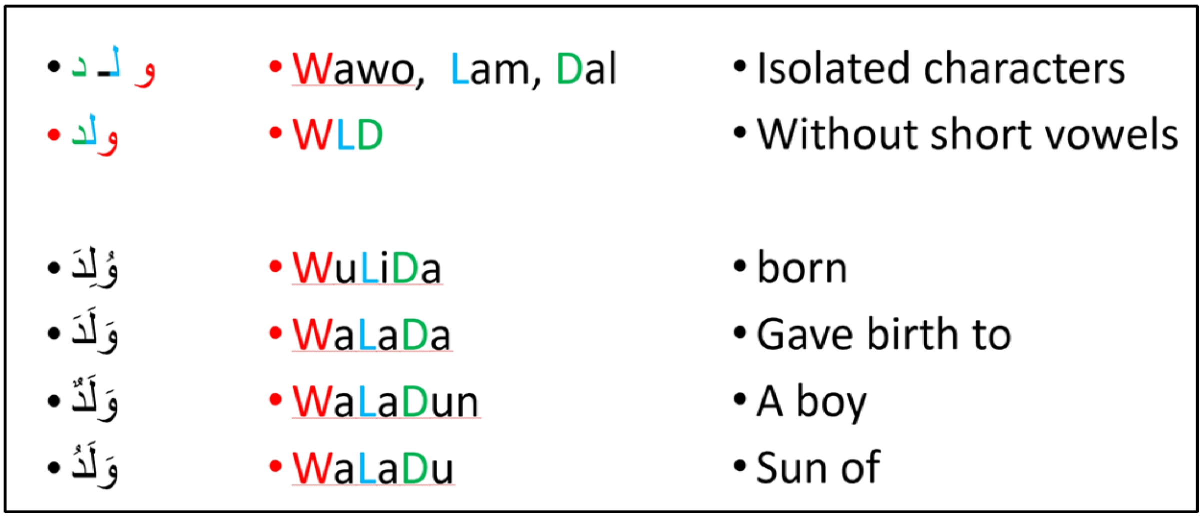


Figure 2: Arabic Characters and it’s articulation points in the human vocal system



و لـ د	Wawo, Lam, Dal	Isolated characters
ولد	WLD	Without short vowels
وُلِدَ	WuLiDa	born
وَلَدَ	WaLaDa	Gave birth to
وَلَدٌ	WaLaDun	A boy
وَلَدُ	WaLaDu	Sun of

Figure 2: Challenges in Arabic written text: Arabic text followed by the English literal equivalent followed by the comment/semantic.

2. Methodology

Beginning with studying the grammar of the Arabic language's graphemes as a focal point in the conversion. The algorithm will be applied to determine phonemes. A data-based method will be proposed to make up for the lack of efficiency. We need to collect phonemes before we set up a benchmark for a Unit Selection scheme. When you're thinking about the process of Unit Synthesis, one of the first things that would probably come to mind is the unit scale. Placement of differentiates the audio expansion elements are widely available. Still, sizes that hold consistent units of extended tone frequencies (e.g., Diphones, tablets, and syllables) are often required.

Often, the boundaries of the audio must be kept low, which minimizes audible discontinuities when concatenation is needed [15]. In certain instances, the system extends when put in the word, which helps one significantly decrease the time spent on a quest on the backend. Additionally, some research has shown that syllables follow these two properties and are thus natural choices as the base unit for aspects such as found in Arabic. Furthermore, it seems as if syllables are much more often preferred over diphones in language communication. They can be broken up into smaller chunks for ease of expression.

In comparison, diacritic languages, such as Arabic, have a vast co-articulative influence on the syllables found in the alphabetic and the canal phonetic regions. Theoretically, this justifies the decision since it ensures greater access across borders. A syllable-dependent unit collection and concatenation process will be developed based on these intuitions. Letter to Sound Laws, founded on heuristics and commonly adopt the C*VC* sequence, where C is a consonant, and V is a vowel, are used to generate syllables [15].

Few syllabograms are permitted, and the total amount of syllabograms that are perceivable is finite. A single-Unlike most languages, in which each syllable begins with a consonant, the Arabic language allows each syllable to be connected to the preceding one immediately with a vowel. In the case of short vowels, V, and in the case of long vowels signifies VV. That is, these components reside in the 2nd place of the spoken syllable.

3. Dataset

There is an unprocessed Speech corpus of Nawar Halabi [11], a professor at the University of Southampton, which can be used to construct templates. The body of the text was recorded in South Levantine. We synthesized a text that combines the human speech of the same elements to increase the overall value of the corpus. This includes translations, digital audio records, raw recordings of the phonemes, and stored data containing demodulated time stamps.

Once it has been expanded, it will be split into three separate datasets: preparation, validation, and testing. Synthesized audio and synthetic voice recordings are not used for voice simulation; they are the only sounds that are never revealed to the model. We have a dataset of 1814 audio files that you can listen to, all of which can be heard in the corpus. The first and second sets of preparation and validation files are stored together. In contrast, the others are only used for test training. This is important to consider since the dataset files can be jumbled due to the effects of the time remapping transformations. This is removed during preprocessing to ensure they are not mixed up with the dataset rearrangement in the next epoch.

The Tacotron model consists of an encoder, an attention-based decoder, and a post-processing net. Tacotron has a CBHG building block, which includes a one-dimensional convolutional filter bank (CB), highway networks (H), and a bi-directional gated recurrent unit (GRU or G) [16-17].

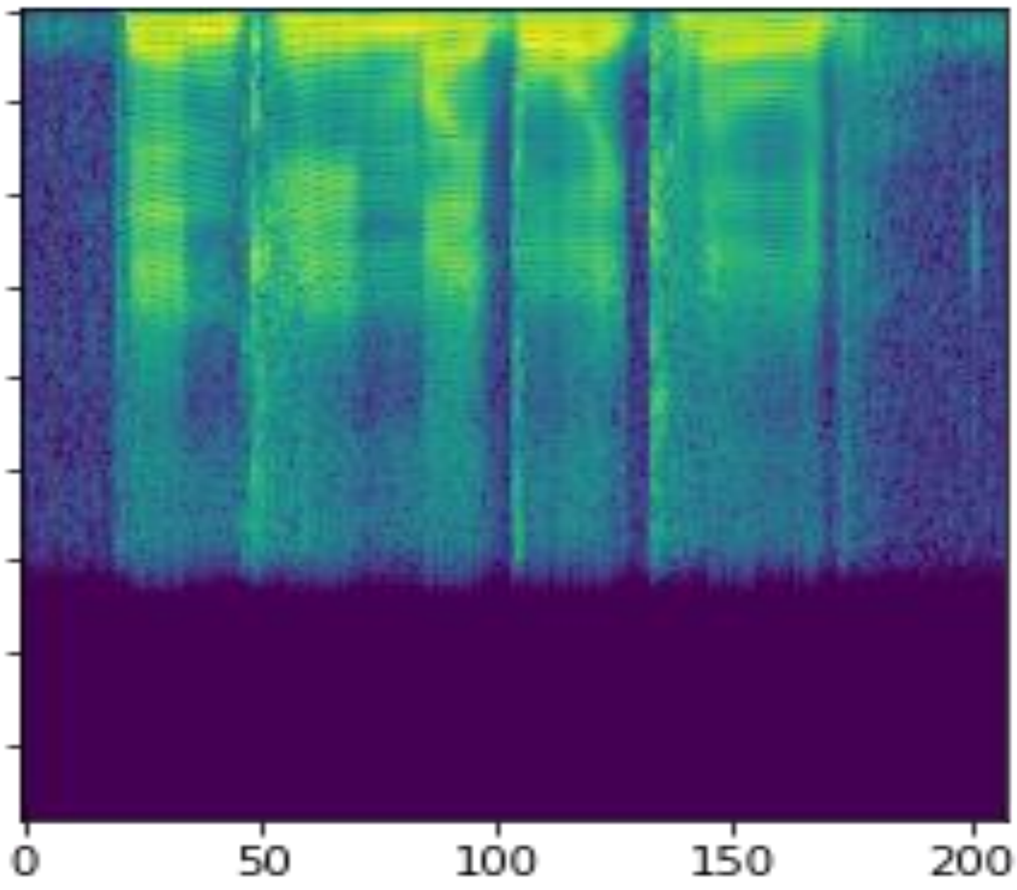


Figure 3: spectrogram output of a poorly trained Tacotron model

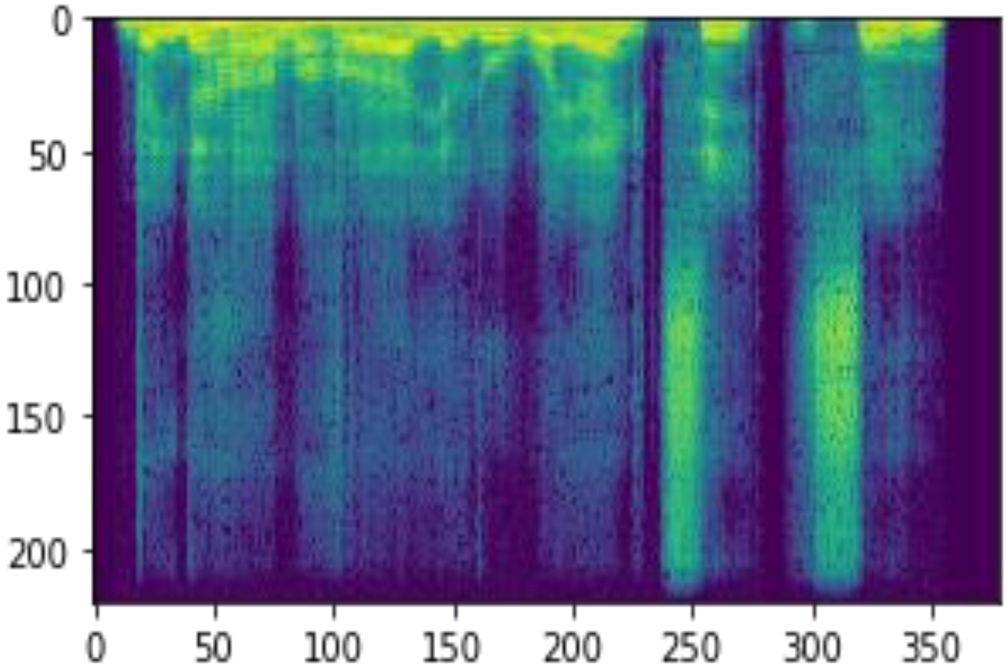


Figure 4: spectrogram output of a well trained Tacotron model

4. Results and Conclusion

This dataset includes 1814 audio files that were used to help Tacotron train the model using spoken utterances. We use visual models trained on different learning or running phases and varying scales to extend our models to show signal dynamics in Spectrogram visualizations [18]. The spectrum (or graphic representation) of frequencies is the sudden fluctuation of a signal at a specified moment. It indicates frequencies on the horizontal and vertical axis of a spectrum graphic representation. The colors equate to the sum of recurrence at any given point, if you catch my drift.

We did a subjective study of the speech to validate the Tacotron models and then measured the extent of the expansion to see how accurate their results were. The first Tacotron model is focused on PyTorch implementation, and the second Tacotron models are of the TensorFlow variety. The second model consists of an Arabic TTS model that has been fine-tuned and pre-trained beforehand. This will be advantageous to do audio spectrogramming during testing and assist in assessing focus quality and assessing for the sound of attention. Regarding expense, most people who deal with WaveNet have contended that it will be an extravagance to use it for training; thus, the general opinion has always been that if it’s being used that you have to amplify sounds, you shouldn’t.

The spectrogram of linked audio with zoomed-in capabilities is seen in the graphs. The performance is graphically visualized in Figs. 2 and 3, which compare four-second spectrograms with the same utterance except with insufficient training in Fig. 2 and better learned in Fig. 3. The spectra in Fig. 3 are considerably more precisely defined, as can be observed.

The Tacotron pre-trained system’s best advantage is that it doesn’t need phoneme segmentation. Without messing with speech, the automated phonemealization process provides extra sound to the device. High-frequency audio can give a distorted sound quality, rendering it unsuitable for use in specific devices.

The success of Tacotron implementation to Arabic text was due to the written Arabic script’s flexibility and general characteristics; nevertheless, the vague spelling of spoken Arabic words adds to the magnificence of the results. To achieve better results, more work in training the model should be done. We can only hope for a short-term expansion of our performance to match the system's potential before further work is done and revised results are achieved.

Building a solid pedagogical and scientific basis to cover the Arabic language grants other applications, such as Spanish, German, Japanese, and, or Sanskrit. The initially proposed framework would generate audio/expand behavioral features similar to those used to create the current system that covers specific feelings and different application features for unknown text. Our primary goal in this paper was to explore new approaches to building an Arabic Speech Synthesis framework using contemporary methodology.

5. Acknowledgements

This research is funded by Kuwait University Research with Grant Number QE04-16.

6. References

[1] I. Isewon, O. J. Oyelade, O. O. Oladipupo, Design and implementation of text to speech conversion for visually impaired people, International Journal of Applied Information Systems 7.2 (2012): 26-30.

[2] M. C. Anil, S. D. Shirbahadurkar, S. S. Shakil, A mapper and combiner based Marathi text to speech synthesis using English TTS Engine, in: 2015 Annual IEEE India Conference (INDICON), IEEE, New Delhi, India, 2015, pp. 1-5, doi: 10.1109/INDICON.2015.7443570.

[3] P. Jayawardhana, A. Aponso, N. Krishnarajah and A. Rathnayake, An Intelligent Approach of Text-To-Speech Synthesizers for English and Sinhala Languages, in: 2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT). IEEE, Kahului, HI, USA, 2019 pp. 229-234. doi: 10.1109/INFOCT.2019.8711051.

[4] M. Macchi, Issues in text-to-speech synthesis, in: Proceedings. IEEE International Joint Symposia on Intelligence and Systems (Cat. No. 98EX174), IEEE, Rockville, MD, USA, 1998, pp. 318-325, doi: 10.1109/IJSIS.1998.685467.

[5] A. Farghaly, K. Shaalan, Arabic natural language processing: Challenges and solutions, ACM Transactions on Asian Language Information Processing (TALIP) 8.4 (2009): 1-22.

[6] D. Newman, The phonetic status of Arabic within the world's languages: the uniqueness of the lughat al-daad, Antwerp papers in linguistics, 100 (2002): 65–75.

[7] I. Sabir and N. Alsaeed, A brief description of consonants in modern standard Arabic, Linguistics and Literature Studies 2.7 (2014): 185-189.

[8] S. Abu-Rabia, The role of vowels in reading Semitic scripts: Data from Arabic and Hebrew, Reading and Writing 14.1-2 (2001): 39-59.

[9] Y. Wang, R. Skerry-Ryan, D. stanton, Y.Wu, R. J.Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, 2017. Tacotron: Towards end-to-end speech synthesis, 2017. URL: https://arxiv.org/abs/1703.10135.

[10] Y. Sherif, Arabic Tacotron Text-To-Speech, 2018. URL: https://github.com/youssefsharief/arabic-tacotrontts.

[11] N. Halabi, M. Wald., Phonetic inventory for an Arabic speech corpus (2016): 734-738.

[12] I. Elias, H. Zen, J. Shen, Yu Zhang, Ye Jia, R. Weiss, and Y. Wu, Parallel Tacotron: Non-Autoregressive and Controllable TTS, 2020. URL: https://arxiv.org/abs/2010.11439.

[13] R. J.Weiss, R. J. Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad, and Diederik P. Kingma, Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis, 2020. URL: https://arxiv.org/abs/2011.03568.

[14] T. Hanane, H. Maamar and A. Hamid, TTS-SA (A text-to-speech system based on standard arabic), in: 2014 Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP). IEEE, Bangkok, Thailand, 2014, pp. 337-341, doi: 10.1109/DICTAP.2014.6821707.

[15] N. Y. Habash, Introduction to Arabic natural language processing, Synthesis Lectures on Human Language Technologies 3.1 (2010): 1-187.

[16] R. Dey and F. M. Salem, Gate-variants of gated recurrent unit (GRU) neural networks, in: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE, Boston, MA, USA, 2017, pp. 1597-1600. doi: 10.1109/MWSCAS.2017.8053243.

[17] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, Ye Jia, Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation, 2019. URL: https://arxiv.org/abs/1904.04169.

[18] B. E. D Kingsbury, N. Morgan, S. Greenberg, Robust speech recognition using the modulation spectrogram, Speech communication 25.1-3 (1998): 117-132.