symmetry

MDPI

Article

Development and Evaluation of Speech Synthesis System Based on Deep Learning Models

Alakbar Valizada [1,2], Sevil Jafarova [1,3,*], Emin Sultanov [1,4] and Samir Rustamov [3,5]

1 Artificial Intelligence Laboratory, ATL Tech, Jalil Mammadguluzadeh 102A, Baku 1022, Azerbaijan; alakbar.valizada@atltech.az (A.V.); emin.sultanov@ufaz.az (E.S.)
2 Information and Telecommunication Technologies, Azerbaijan Technical University, Hussein Javid Ave. 25, Baku 1073, Azerbaijan
3 School of Information Technologies and Engineering, ADA University, Ahmadbey Aghaoglu Str. 11, Baku 1008, Azerbaijan; srustamov@ada.edu.az
4 Department of Computer Science, French-Azerbaijani University, 183 Nizami St., Baku 1010, Azerbaijan
5 Institute of Control Systems, Bakhtiyar Vahabzadeh Str. 9, Baku 1141, Azerbaijan
* Correspondence: sevil.jafarova@atltech.az

Abstract: This study concentrates on the investigation, development, and evaluation of Text-to-Speech Synthesis systems based on Deep Learning models for the Azerbaijani Language. We have selected and compared state-of-the-art models-Tacotron and Deep Convolutional Text-to-Speech (DC TTS) systems to achieve the most optimal model. Both systems were trained on the 24 h speech dataset of the Azerbaijani language collected and processed from the news website. To analyze the quality and intelligibility of the speech signals produced by two systems, 34 listeners participated in an online survey containing subjective evaluation tests. The results of the study indicated that according to the Mean Opinion Score, Tacotron demonstrated better results for the In-Vocabulary words; however, DC TTS indicated a higher performance of the Out-Of-Vocabulary words synthesis.

Keywords: text-to-speech synthesis; Tacotron; deep convolutional text-to-speech; evaluation of speech synthesis; speech quality and intelligibility

Citation: Valizada, A.; Jafarova, S.; Sultanov, E.; Rustamov, S. Development and Evaluation of Speech Synthesis System Based on Deep Learning Models. *Symmetry* **2021**, *13*, 819. https://doi.org/10.3390/sym13050819

Academic Editor: José Carlos R. Alcantud

Received: 12 March 2021
Accepted: 21 April 2021
Published: 7 May 2021

1. Introduction

Speech is one of the simplest sources of communication, yet it is a challenging phenomenon. Spoken communication includes two distinct aspects, which are the verbal component and the prosodic component. The former component combines two systems in which the first forms words from phonemes and the second constructs sentences from words. The prosodic component of the speech aims to transmit the intonational facet of language rather than symbolic. In other words, it helps to express emotions, to stress a particular word or the end of a sentence [1]. Both components share the same speech signal; however, they play peculiar roles in formulating natural language.

With the advances in technology and the increase in the number of portable devices, modern means of communication are becoming more attractive. The main goal of a speech synthesizer or text-to-speech (TTS) is the artificial reproduction of the human voice from the given text. The system primarily focuses on achieving a clear understandable synthesis of the speech, which does not sound robotic. TTS has a wide range of applications such as call-center automation, weather announcements or schedule reports, reading text out loud for visually impaired people, and for those who lost their voice.

In this paper, we describe previous approaches of TTS development, then, we introduce the architecture of two models for a better understanding of the synthesis process. Next, we provide information about the database collected from scratch specifically for the Azerbaijani language, which is used in the training process for both TTS models. Lastly, the experiments are carried out and both speech synthesizers are evaluated to ensure the accuracy of the models.

2. Related Work

Different approaches of speech synthesis have been investigated and analyzed throughout the years. Previous research in this area has made significant progress, thus, accelerating the training process of the model and improving the quality of speech synthesis, allowing creating speech similar to humans. One of the well-known architectures represents formant synthesis, which involves the generation of artificial spectrograms to simulate formant characteristics produced by the vocal tract. This technique does not utilize instances of human speech, however, focuses on the extraction of formant parameters and transition between phonemes based on the rules derived by linguists from spectrograms. The formant synthesized speech is intelligible and is suitable for systems with limited memory since it does not store samples of human speech, however, it has not reached a high level of naturalness [2]. Articulatory synthesis produces speech by creating synthetic human articulatory systems and includes parameters such as positions of jaws, lips, and tongue [3]. The data for this model is gathered from MRI or X-ray images. Thus, one of the challenges to build an articulatory model is data acquisition considering expenses and availability of appropriate equipment [4]. Concatenative synthesis mainly represents two types of approaches, which are diphone synthesis and unit selection synthesis. Diphone synthesis is achieved by recording natural speech that involves each possible phoneme of context, which are then segmented and labeled. After, these diphones should be modified by signal processing methods to adjust prosody of the speech [2]. Synthesis based on the unit selection, however, does not require any signal processing since it stores multiple samples of units with varying prosodies in the database [5]. One problem with the rule-based systems is the storage of various occurrences of each acoustic unit, which leads to the consumption of large memory resources. In contrast, Hidden Markov Model (HMM) retrieves averages of similar sounding speech segments rather than storing each instance of speech from the database and represents relatively smooth transitions [6]. Recently, Deep Neural Network (DNN) algorithms have been applied to achieve natural and high-quality speech synthesis. Experiments indicated that DNN speech synthesis can outperform HMM-based approach given the same number of speech parameters [7].

In [8], authors from Baidu keep traditional text-to-speech pipelines and adopt the same structure, while replacing all components with neural networks and using simpler features. They propose a novel way of performing phoneme boundary detection DNN using connectionist temporal classification (CTC) loss. For the audio synthesis model, they implement WaveNet that completes training faster than the original.

Deep reinforcement learning (DRL) algorithms are also being employed in audio signal processing to learn directly from speech, music, and other sound signals in order to create audio-based autonomous systems that have many promising applications in the real world. Despite DRL applied widely in the natural language processing (NLP) tasks, there is no broad investigation of DRL for speech synthesis [9].

The encoder-decoder architecture with self-attention or bi-directional long short-term (BLSTM) units can produce high-quality speech. However, these models' synthesis speed is slower for longer inputs. Authors of [10] propose a multi-rate attention architecture that breaks the latency and RTF bottlenecks by computing a compact representation during encoding and recurrently generating the attention vector in a streaming manner during decoding.

The project GraphSpeech proposes a novel neural TTS model that is formulated under graph neural network framework [11]. GraphSpeech encodes explicitly the syntactic relation of input lexical tokens in a sentence, and incorporates such information to derive syntactically motivated character embeddings for the TTS attention mechanism. Experiments show that GraphSpeech consistently outperforms the Transformer TTS baseline in terms of spectrum and prosody rendering of utterances.

Similarly, studies have proposed different approaches for Azerbaijani Text-to-Speech synthesis. One of the studies [12] describes the process of conversion text to speech by suggesting the search of individual words in the dictionary. If the word is not in the proposed dictionary, then it searches according to the "Root and Affix" splitting

method. In case of the absence of "root forms" or affixes, conversion of letters to sounds is performed to match the best pronunciation. Another study [13] proposes the combination of concatenative and formant synthesis to achieve naturality and intelligibility of the produced speech. However, researchers have not studied the application of DNN models for Azerbaijani TTS systems. Most of the existing research on speech technologies for the Azerbaijani language is related to speech recognition [14]. In this paper, we propose the development and evaluation of Azerbaijani DNN TTS systems.

3. Lexicon of the Azerbaijani Language

The Azerbaijani language contains quite a few aspects that affect the different pronunciations of words. In addition to the most important aspects—accent and dialect—there are also basic rules of pronunciation of vowels and consonants in this language. Contrary to English that contains 44 phonemes, the Azerbaijani language comprehends 34 phonemes, 9 vowels, and 25 consonants. However, if we differentiate nine short (i,ü,e,ö,ə,a,o,u,ı) and six long (i:,e:,ö:,ə:,a:,u:) vowels, then the total number of phonemes will be 40.

Pronunciation rules for vowels:

- Two identical vowels, one following the other in a word, are pronounced as one long vowel (e.g., "saat" [sa:t] ("clock")).
- The sound [y] is included between two different consecutive vowels (e.g., "radio" [radiyo] ("radio")).
- Once the so-called vowels "əa", "üə", "üa" converge, the first one falls and the second one stretches (e.g., "müavin" [ma:vin] ("deputy")).
- In the case that o or ö are followed by the sound [v], the v drops and the vowels are stretched (e.g., "dovşan" [do:şan] ("rabbit")).
- When certain suffixes with conjunctive y at the beginning are added to polysyllabic words ending in "a" or "ə", these vowels become one of the closed vowel sounds ([ı], [i], [u], [ü]) during pronunciation (e.g "babaya" [babıya] ("to grandfather")).

Pronunciation rules for consonants:

- If the consonants "qq", "pp", "tt", "kk" merge in the middle of a word, one of them changes during pronunciation (e.g., "tappıltı" [tapbıltı] ("thud")).
- If the sounds b, d, g, c, q, k, z come at the end of a word, they change during pronunciation to p, t, k, ç, x, x', s, respectively (e.g., "almaz" [almas] ("diamond")).
- If a consonant comes after the k sound in the middle of a word, its pronunciation is k = x ' (e.g., məktəbli [məx'təbli] ("student")).

4. Methodology

Different methods have been used for the systems designed for speech synthesis. Each system has both advantages and drawbacks. For this reason, two of the most popular systems have been chosen to preserve most of the models' privileges: Tacotron and Deep Convolutional Text-to-Speech (DC TTS).

Both Tacotron and DC TTS models have a different structure and design. Therefore, the initial study of these systems allows a better understanding of their performance and their impact on the result.

4.1. Tacotron

Tacotron is a system that uses the seq2seq [15] model to produce spectrogram frames, which are needed to create waveforms, from the input characters. According to [16], the architecture of the model consists of an encoder, an attention-based decoder, and a post-processing net (Figure 1).

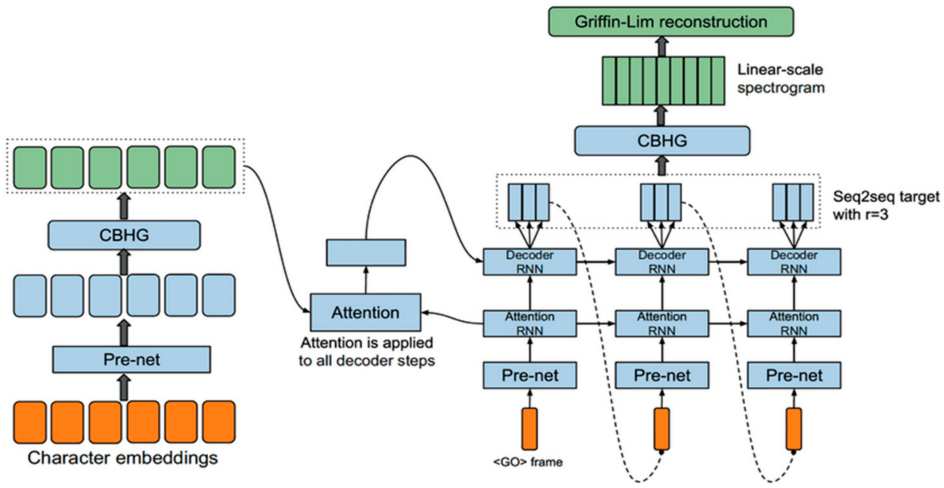

Figure 1. Tacotron architecture [16].

In addition to the basic modules, this model also uses a building block CBHG for extracting representations from sequences. This module consists of three parts:

1. 1-D convolutional filters bank, where the convolved inputs are used to generate local and contextual information. Afterward, the outputs are added together and combined over time to build increased local invariances. This sequence is then passed to the several fixed-width 1-D convolutions and added with the input sequence.
2. Highway networks (Figure 2) [17] which are used to extract high-level features.
3. Bidirectional gated recurrent unit (GRU) [19] recurrent neural network (RNN) [20] used to derive sequential features from the forward context and backward context.

- Encoder module is necessary for extracting solid, consistent text representations. The procedure starts with the embedding of the one-hot representation of each character into a continuous vector. Subsequently, the vector passes through the non-linear transformations of the bottleneck layer, also known as pre-net, which helps to improve generalization, and the transformations of the CBHG module to reduce overfitting and mispronunciations. This gives the final representation of the outputs of the encoder used as the attention module.
- Decoder is a content-based tanh attention decoder, where the input is a concatenated context vector with the attention RNN output. During each time step, it creates the attention query. In addition, the decoder is implemented with a GRU stack containing vertical residual connections as it helps to speed up convergence. Moreover, due to the highly redundant representation of the raw spectrogram, and as the seq2seq target can be highly compressed while it provides prosody information for an inversion process, the 80-band mel-scale spectrogram is used as the target of the seq2seq, that will be further converted to waveform in post-processing network.
- The decoder output layer represents a fully connected output layer. It serves to anticipate the decoder targets. As the prediction process uses prediction of r frames at once, it reduces training time together with inference time as well as increases the convergence rate.
- Post-processing net is necessary to convert the seq2seq target to target, which will be further converted into a waveform. This network learns to predict the spectral value displayed on a linear frequency scale. Information contained in the post-processing net that includes both forward and backward information is used to correct the prediction error. As a post-processing net, Tacotron uses the CBHG module.

This system uses as a synthesis from the predicted spectrogram the Griffin-Lim algorithm [21].

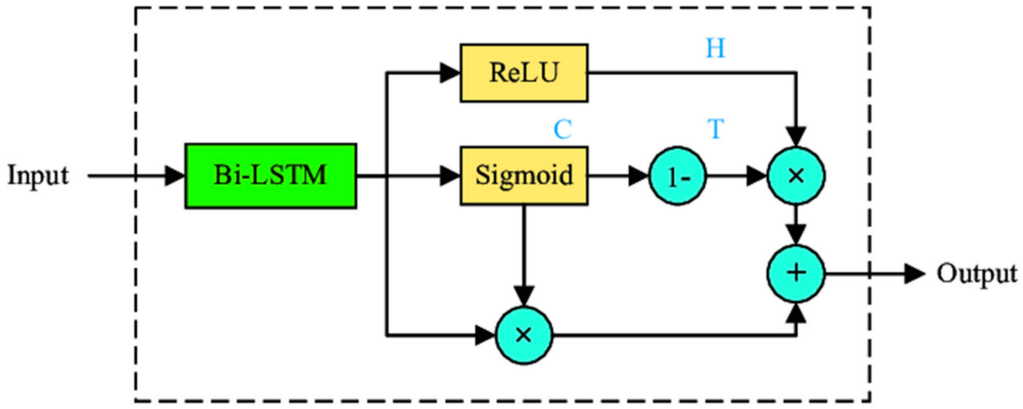

Figure 2. Highway network architecture [18].

4.2. Deep Convolutional Text-to-Speech

DC TTS is another speech synthesis technology based on convolutional seq2seq [22]. As described in [23], this model consists of two networks (Figure 3):

- Text to mel spectrogram network, which forms a mel spectrogram from an input text and is composed of four submodules:
  1. The text encoder encodes an input sentence L of N characters into the matrices of keys and values K, V of dimension $256 \times N$.

$$(K, V) = TextEnc(L) \tag{1}$$

  2. The audio encoder encodes the coarse mel spectrogram $S1{:}F,1{:}T$, considering that S is a normalized mel spectrogram with applied mel filter bank, where F is the number of frequency bins and T is the length of the previously spoken speech, into a matrix Q with dimension $256 \times T$.

$$Q = AudioEnc\ (S_{1:\ F,1:\ T}) \tag{2}$$

  3. Attention matrix A evaluates how closely the n-th character in the sentence is related with t-th time frame of the mel spectrogram.

$$A = softmax_{n\text{-}axis}\ (K^T Q/\sqrt{d}) \tag{3}$$

  where *softmax* function determines whether it is the searched character or not. In case if A_{nt} 1 with n-th character, it starts to look at l_{n+1} or characters near it or near l_n at the subsequent time $t + 1$. Assuming that these characters are encoded column V, then a seed $R \in R^{256xT}$ that is decoded to subsequent frames $S_{1:\ F,2:\ T+1}$

$$R = Attention\ (Q, K, V){:} = V\ A \tag{4}$$

  4. Audio Decoder decodes the concatenated matrix $R' = [R, Q]$ to synthesize a coarse mel spectrogram.

$$Y_{1:\ F,2:\ T+1} = AudioDec\ (R') \tag{5}$$

  Afterward, this result is compared with temporally shifted ground truth $S_{1:\ F,\ 2:\ T+1}$ by a loss function, which is the sum of $L1$ loss [24] function and binary convergence, that is calculated by:

$$D_{bin}(Y \mid S) := E_{ft}[-S_{ft}\acute{Y}_{ft} + log(1 + exp\acute{Y}_f)] + const \tag{6}$$

  where $\acute{Y}_{ft} = logit\ (\acute{Y}_{ft})$. The error is back propagated to the network parameters.

- Spectrogram Super-resolution network. The network synthesizes a full spectrogram from the obtained coarse mel spectrogram.

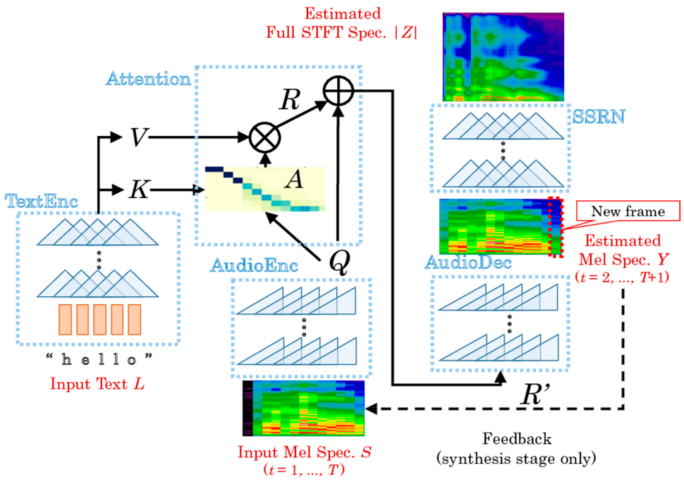

Figure 3. DC TTS model [23].

5. Data Collection and Processing

Each system based on neural networks depends on the data used during the training process. Despite the fact that large public domain speech datasets are available online, there is no Azerbaijani data for this specific purpose. Thus, the data collection process represents one of the significant stages of this project.

The data collected for the model training were obtained from the publicly available news portal consisting of around 24 h in total spoken by a single male speaker. Each acoustic data has its transcription which allows the training of models based on 16,780 <audio, text> pairs in total without phoneme-level alignment. The process of speech and text mapping was performed by applying the forced alignment technique, which, in turn, uses the Viterbi algorithm together with the linguistic model and an acoustic model [25]. Considering that longer duration of audio recordings in the data for training can lead to incorrectly trained models and also significantly slow down the training process, each audio clip varies in length from 1 to 15 s.

6. Evaluation Methods

Currently, the evaluation of text-to-speech systems occurs in different ways. However, not all methods are effective, some in turn are only suitable for certain cases such as one-word synthesis evaluation. The choice of a suitable method is, therefore, challenging.

In general, the process of model evaluation can be performed on two types of tests:

- **Subjective tests.** These are the listening tests where each listener judges speech quality and naturalness.
- **Objective tests.** Tests where measurement of voice performance estimated by applying appropriate speech signal processing algorithms.

According to [5], the evaluation of the TTS system is based on two main metrics:

1. **Intelligibility**-index of the correctness in the interpretation of the words. This metric can be evaluated using the following tests:
    - Diagnostic Rhyme Test (DRT)-subjective test, based on pairs of words with confusable rhyming and differs only in a single phonetic feature. In this test, listeners have to identify the index of the given word [26]. The percentage of the right answers is used as an intelligibility metric.
    - Modified Rhyme Test (MRT)-subjective test, similar to the previous test, except the fact that it is based on different sets of six words [27].

- Semantically Unpredictable Sentence (SUS)-subjective test, based on sentences of randomly selected words [28]. Using this method, the intelligibility can be evaluated using the formula below:

$$SUS = 100 \times \frac{C}{S \times L} \quad (7)$$

where C is the number of the correct predicted sentences, S is the total number of the tested sentences and L is the number of listeners.

2. **Quality**-index of the naturalness, fluency, clarity. This metric can be evaluated using the following methods:
   - AB test-subjective test, where listeners measure the quality of different audio files with synthesized speech produced by system A and B by comparing the results of these systems with each other (Table 1) [29]. The final result shows which of the systems is better.
   - ABX test-subjective test, where listeners make comparison between the synthesized sentence by system A and synthesized sentence by system B in terms of the closeness to the originally voiced sentence X [29].
   - Mean Opinion Score (MOS)-subjective test, where listeners evaluate each sentence synthesized by the system on a 1–5 scale from bad to excellent (Table 2) [29].

Table 1. AB test rating scale.

Rating	Quality of Synthesized Audios by Systems A and B
3	A very good
2	A better
1	A good
0	About the same
1	B good
2	B better
3	B very good

Table 2. MOS rating scale.

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Afterward, the sum of these scores is collected and divided by the total number of the evaluated sentences and the number of listeners.

- Mel Cepstral Distortion (MCD)-objective test, that measures the difference between synthesized and natural mel cepstral sequences consisting of extracted mel-frequency cepstral coefficients [30]. This difference shows how the reproduced speech is closer to the natural one. MCD can be calculated by the formula:

$$\text{MCD} = \frac{10\sqrt{2}}{ln\ 10 \times T'} \sum_{t=0}^{T-1} \sqrt{\sum_{t=s}^{D} \left(v_d^{targ}(t) - v_d^{ref}(t)\right)^2} \quad (8)$$

where v_d^{targ}, v_d^{ref} are mel frequency cepstral coefficients of the t-th frame from the reference and predicted audio, d is dimension index from 0 to 24, t is time (frame index) and T' is the number of non-silence frames.

- Segmental Signal-to-noise ratio (SNRseg) objective test, which measures noise ratio between two signals [29]. This ratio can be calculated using the formula below:

$$\text{SNRseg} = \frac{10}{M} \sum_{m=0}^{M-1} lg \frac{\sum_{n=Nm}^{Nm+N-1} x^2(n)}{\sum_{n=Nm}^{Nm+N-1} (x(n) - \overline{x}(n))^2} \quad (9)$$

  where $x(n)$ is the original signal, $\overline{x}(n)$ is the synthesized signal, N is the frame length, M is the number of frames in the speech signal.
- Perceptual evaluation of speech quality (PESQ)-objective test, which allows to predict the results of MOS evaluation [29]. This helps to automate the MOS evaluation and make this process faster. PESQ is calculated by the following formula:

$$\text{PESQ} = \text{a0} + \text{a1} \cdot \text{dsym} + \text{a2} \cdot \text{dasym} \quad (10)$$

  where $a_0 = 4.5$, + $a_1 = -0.1$, $a_2 = -0.0309$, d_{sym} is the average disturbance and d_{asym} is the average asymmetrical disturbance value.

However, the evaluation tests mentioned above cannot guarantee an accurate result. All subjective tests require a large number of listeners for more accurate evaluation. Nevertheless, the cognitive factor, such as attention, dependency on the mood of the subject, an environment of listening tests, etc., often plays a role and affects the results. On the other hand, objective tests cannot give a full assessment of the model either, since the acoustic properties of natural and synthesized speech are different as well as prosody. These differences can contribute to the naturalness and an intelligibility loss for synthesized speech.

Despite all the inaccuracies, many systems use the most popular metric MOS for evaluating synthesized speech, taking into account only the quality of the speech, since the intelligibility metrics are not important in commerce.

7. Experiments

To evaluate two main attributes, which are the quality and the intelligibility of the speech signal, subjective assessments were carried out. Online survey was conducted among 34 participants, distinguishing 32 native and two nonnative speakers. It was recommended to participants to be in a quiet room and use headphones while evaluating sounds.

Test data was classified based on the presence of Out of Vocabulary (OOV) words and source type with a total of 53 sentences (Table 3). Specifically, 5 sentences without any OOV words were selected to evaluate the AB test, 14 sentences for each system were chosen for MOS assessment, where 3 of them include 1 OOV word, 6–2 OOV words, 4–3 OOV words, and 1 sentence contained 4 OOV words. Moreover, 10 sentences for each system were organized to assess the SUS test accordingly: 3 sentences with no OOV words, 4 sentences with 1 OOV word, and 3 sentences included 2 OOV words. In addition, it is noteworthy to mention that 25 sentences were taken from news resources, 8 sentences were selected from various storybooks and the rest of the 20 sentences were formed with random words specifically for the SUS test.

Table 3. Classification of test sentences depending on the number of OOV words.

Test Type	No of Test Sentences	No of Sent. without OOV Words	No of Sent. with 1 OOV Word	No of Sent. with 2 OOV Words	No of Sent. with 3 OOV Words	No of Sent. with 4 OOV Words
AB	5	5				
MOS for system A	14		3	6	4	1
MOS for system B	14		3	6	4	1
SUS for system A	10	3	4	3		
SUS for system B	10	3	4	3		

To assess the naturalness and clarity of synthesized speech produced by two systems, we used the AB metric that examines participants' relative preference for either system A (Tacotron) or system B (DC TTS). In this test, listeners are asked to rate synthesized audios on a 3-point scale, where 0 indicated no significant difference between system A and system B, whereas the selection of 3 following by the system type (A or B) demonstrated a strong preference, and the selection of 1 illustrated a slight preference of participants for that particular system over another (Table 1). After, all scores are summed and divided to the total number of AB test sentences and to the number of participants.

In contrast to the AB test, which was used to compare generated audios of both systems at the same time, in the MOS test we presented synthesized audios of each system separately to let listeners assess speech quality and level of distortions (Table 2). Despite the fact that different sets of sentences were picked for each system, the number of OOV words in sentences was distributed equally to avoid possible future bias. Finally, the quality of the test signal produced by the system was obtained by averaging the ratings obtained from all listeners then, summing these averages and dividing the obtained sum by the number of total test sentences.

As the name suggests, for the SUS test we have selected random words that form grammatically correct sentences, however, they are unpredictable and do not have any semantic meaning. The SUS test measures the intelligibility of test sentences since it challenges listeners to type what they hear instead of predicting words based on the meaning of a whole sentence. At the beginning of the survey, participants were asked to listen to each synthesized recording up to two times and to write down the words that they heard in a sequence. Lastly, the SUS score was calculated for each system A and B by (Formula (7)).

8. Discussion and Results

Findings suggest inconsiderable dominance of Tacotron in the comparison of average results of evaluation metrics and relatively higher individual evaluation scores of DC TTS for sentences that contain more OOV words. Firstly, while assessing the quality of the two systems by the AB score, we have observed the value of 0.90 for Tacotron and 0.59 for DC TTS. Figure 4 illustrates that Tacotron outperforms DC TTS in the AB test and the AB score is close to 1, which indicates that the quality of audios produced by Tacotron is "good" according to the 0–3 scale. Moreover, the score obtained by DC TTS indicates that the quality of synthesized sounds is between "almost the same with another system" and "good".

Table 4 demonstrates the comparison of MOS (95% Confidence Interval) for all sentences for both systems. The table shows that Tacotron achieves a MOS of 3.49 ± 0.193 and slightly outperforms DC TTS, which gets a MOS of 3.36 ± 0.187. In other words, both systems synthesize audios with "fair" sound quality and the level of distortion is perceptible, however, not considerably annoying. Furthermore, besides considering only the average MOS of all sentences without considering the number of OOV, we analyzed the average MOS of the sentences grouped by OOV words (Figure 5). Moreover, individual mean opinion scores were examined depending on the number of OOV words presented in separate test sentences. Surprisingly, although Tacotron demonstrated higher average

MOS for the sentences containing few OOV words, it achieved lower values (<3.0) on the last two sentences containing three and four OOV words. On the other hand, DC TTS got higher scores ($\leq$3.0) during the synthesis of sentences including three or four OOV words. These results may imply the future usage of a hybrid model for the Azerbaijani TTS system that combines Tacotron for the synthesis of sentences with words in the train data and DC TTS for sentences containing more OOV words.

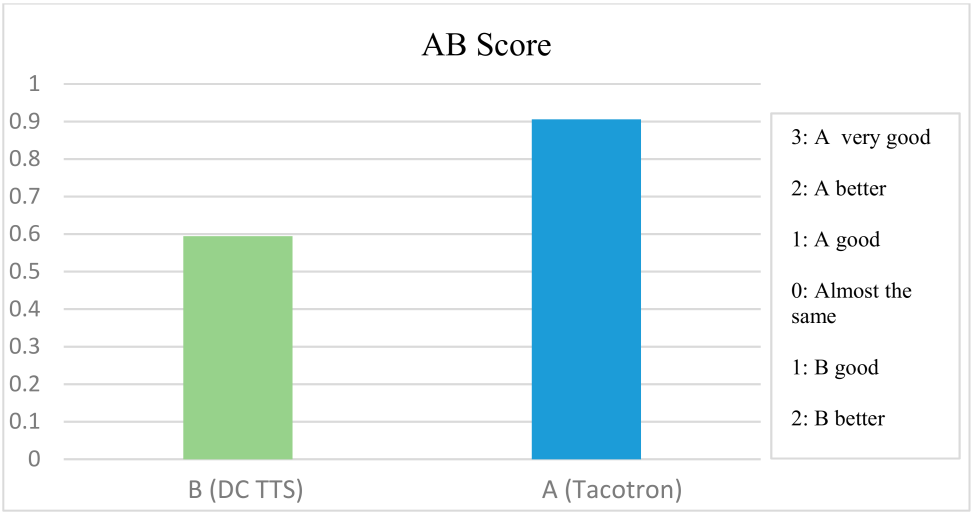

Figure 4. AB Score for A (Tacotron) and B (DC TTS).

Table 4. Comparison of Mean Opinion Score (MOS) for Tacotron and DC TTS.

System	MOS (95% CI)
Tacotron	3.49 $\pm$ 0.193
DC TTS	3.36 $\pm$ 0.187

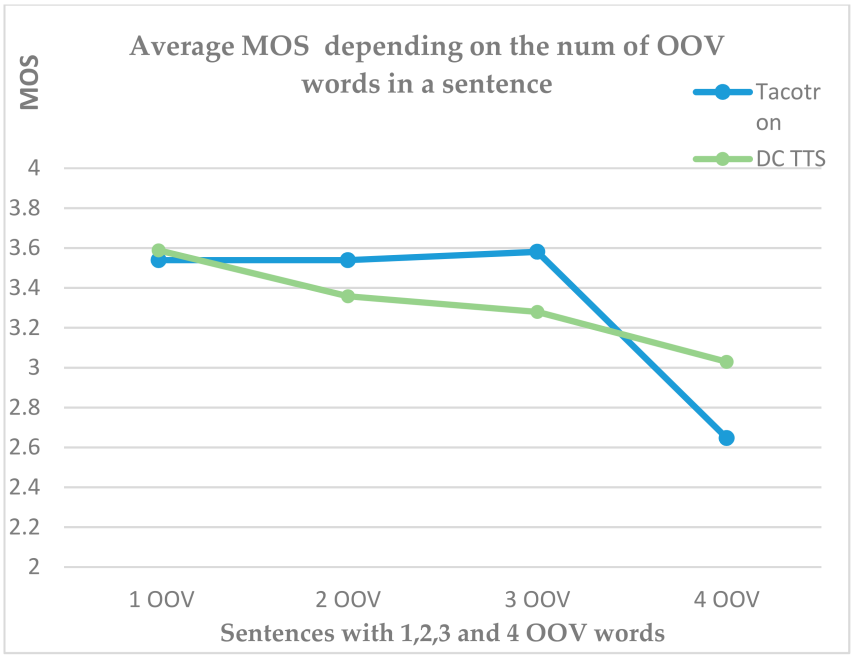

Figure 5. Average MOS for Tacotron and DC TTS and sentences with OOV words.

Lastly, the percent of the SUS test for the systems is demonstrated in Table 5. The study shows that more sentences without semantic meaning were predicted correctly for

audios produced by Tacotron and resulting in 49% of correct test sentences compared to DC TTS having 45%.

Table 5. Comparison of SUS Score.

System	SUS Score F%
Tacotron	49%
DC TTS	45%

An apparent limitation of the study is the insufficient sample size of the included test sentences in surveys since 53 sentences took more than 40 min for listeners to evaluate synthesized audios. A possible solution could be dividing the survey questions according to the type of evaluation metrics.

9. Conclusions

The current research examines the architecture and evaluation of two speech synthesis systems based on DNN for the Azerbaijani language. DC TTS and Tacotron models were trained on the 24-h mono speaker data. Test sentences were classified and divided based on the number of OOV words to compare the results of the proposed models. We described and applied subjective evaluation metrics to measure the intelligence and the quality of the synthesized speech produced by systems. The AB, MOS, and SUS metrics indicated higher results for all test sentences synthesized by Tacotron. The study demonstrated that sentences prevailing words from train vocabulary achieved higher MOS metrics for Tacotron and sentences containing a higher number of OOV words obtained better MOS results for DC TTS. Thus, future research should consider the application of a hybrid model that optimizes benefits of both systems to achieve more natural and intelligible speech.

Author Contributions: Conceptualization, S.R.; methodology S.J., E.S., A.V. and S.R.; software S.J. and E.S.; validation A.V.; writing—original draft preparation E.S. and S.J.; writing—review and editing S.J., S.R. and A.V.; visualization S.J.; investigation A.V.; project administration S.R.; resources A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ATL Tech LLC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Research data will remain confidential due to commercial reasons.

Acknowledgments: This work was carried out in the Center for Data Analytics Research at ADA University and in Artificial Intelligence Laboratory at ATL Tech.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Taylor, P. Communication and Language. In *Text-to-Speech Synthesis*; Cambridge University Press: Cambridge, UK, 2009; pp. 13–15.
2. Tabet, Y.; Boughazi, M. Speech Synthesis Techniques. A Survey. In Proceedings of the International Workshop on Systems, Signal Processing and Their Applications (WOSSPA), Tipaza, Algeria, 9–11 May 2011.
3. Kaur, G.; Singh, P. Formant Text to Speech Synthesis Using Artificial Neural Networks. In Proceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, 25–28 February 2019.
4. Tsukanova, A.; Elie, B.; Laprie, Y. Articulatory Speech Synthesis from Static Context-Aware Articulatory Targets. In *International Seminar on Speech Production*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 37–47. Available online: https://hal.archives-ouvertes.fr/hal-01937950/document (accessed on 30 September 2020).
5. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 2nd ed.; Prentice Hall: Hoboken, NJ, USA, 2008; pp. 249–284.
6. Jeon, K.M.; Kim, H.K. HMM-Based Distributed Text-to-Speech Synthesis Incorporating Speaker-Adaptive Training. 2012. Available online: https://www.researchgate.net/publication/303917802_HMM-Based_Distributed_Text-to-Speech_Synthesis_Incorporating_Speaker-Adaptive_Training (accessed on 30 September 2020).

7. Qian, Y.; Fan, Y.; Hu, W.; Soong, F.K. On the Training Aspects of Deep Neural Network (DNN) for Parametric TTS Synthesis. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014. Available online: https://ieeexplore.ieee.org/document/6854318 (accessed on 30 September 2020).
8. Arık, S.Ö.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Li, X.; Miller, J.; Raiman, J.; Sengupta, S.; et al. Deep Voice: Real-time Neural Text-to-Speech. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
9. Latif, S.; Cuayahuitl, H.; Pervez, F.; Shamshad, F.; Ali, H.S.; Cambria, E. A survey on deep reinforcement learning for audio-based applications. *arXiv* **2021**, arXiv:2101.00240.
10. He, Q.; Xiu, Z.; Koehler, T.; Wu, J. Multi-rate attention architecture for fast streamable Text-to-speech spectrum modeling. *arXiv* **2021**, arXiv:2104.00705.
11. Liu, R.; Sisman, B.; Li, H. Graphspeech: Syntax-aware graph attention network for neural speech synthesis. *arXiv* **2020**, arXiv:2104.00705.
12. Rustamov, S.; Saadova, A. On an Approach to Computer Synthesis of Azerbaijan speech. In Proceedings of the Conference: Problems of Cybernetics and İnformatics, Baku, Azerbaijan, 12–14 September 2014.
13. Aida–Zade, K.R.; Ardil, C.; Sharifova, A.M. The Main Principles of Text-to-Speech Synthesis System. *Int. J. Signal Process.* **2013**, *6*, 13–19.
14. Valizada, A.; Akhundova, N.; Rustamov, S. Development of Speech Recognition Systems in Emergency Call Centers. *Symmetry* **2021**, *13*, 634. [CrossRef]
15. Sutskever, I.; Vinyals, O.; Quoc, V.L. Sequence to Sequence Learning with Neural Networks. *arXiv* **2014**, arXiv:1409.3215.
16. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, J.R.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards End-to-End Speech Synthesis. *arXiv* **2017**, arXiv:1703.10135.
17. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Highway networks. *arXiv* **2015**, arXiv:1505.00387.
18. Jin, Y.; Xie, J.; Guo, W. LSTM-CRF Neural Network with Gated Self Attention for Chinese NER. *IEEE Access* **2019**, *7*, 136694–136703. [CrossRef]
19. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
20. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *arXiv* **2018**, arXiv:1409.3215.
21. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
22. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
23. Tachibana, H.; Uenoyama, K.; Aihara, S. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. *arXiv* **2017**, arXiv:1710.08969.
24. Janocha, K.; Czarnecki, W.M. On Loss Functions for Deep Neural Networks in Classification. *arXiv* **2017**, arXiv:1702.05659.
25. ReadBeyond. Aeneas. 2020. Available online: https://github.com/readbeyond/aeneas (accessed on 13 May 2020).
26. Voiers, W.; Sharpley, A.; Hehmsoth, C. Diagnostic Evaluation of Intelligibility in Present-Day Digital. In *Research on Diagnostic Evaluation of Speech Intelligibility*; National Technical Information Service: Springfield, VA, USA, 1975; pp. 87–92. Available online: https://apps.dtic.mil/dtic/tr/fulltext/u2/755918.pdf (accessed on 12 June 2020).
27. House, A.; Williams, C.; Heker, M.; Kryter, K. Articulation testing methods: Consonantal differentiation with a closed response set. *J. Acoust. Soc. Am.* **1965**, *37*, 158–166. [CrossRef] [PubMed]
28. Benoît, C.; Griceb, M.; Hazanc, V. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Commun.* **1988**, *18*, 381–392. [CrossRef]
29. Loizou, P.C. Speech Quality Assessment. In *Multimedia Analysis, Processing and Communications*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 623–654. Available online: https://ecs.utdallas.edu/loizou/cimplants/quality_assessment_chapter.pdf (accessed on 15 August 2020).
30. Kominek, J.; Schultz, T.; Black, A.W. Synthesizer Voice Quality of New Languages Calibrated with Mean Mel Cepstral Distortion. In Proceedings of the SLTU-2008—First International Workshop on Spoken Languages Technologies for Under-Resourced Languages, Hanoi, Vietnam, 5–7 May 2008. Available online: https://www.cs.cmu.edu/~{}awb/papers/sltu2008/kominek_black.sltu_2008.pdf (accessed on 18 April 2021).