SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Advancing Text-To-Speech Synthesis Using Efficient Deep Learning Techniques For Enhanced Communication Recognition

[1*]M. Divya, [2]Dr. N. Srinivas, [3]rajeswari S, [4]Dr. Bharathi V, [5]m. Ganesan, [6]manikandan Rajagopal

[1*]Assistant Professor, Department of Computer Science and Applications, SRM Institute of Science and Technology, Ramapuram, Chennai, 600089, Tamil Nadu, India https://orcid.org/0000-0002-9960-8061
divyamathavan3@gmail.com
[2]Associate Professor, Department of CSE, Vignana Bharathi Institute of Technology, Aushapur(V), Ghatkesar.
https://orcid.org/0000-0002-0043-4736
Srinivas.bhaskar3@gmail.com
[3]Assistant Professor, Department of Computer Science and Engineering, Dr.N.G.P Institute of Technology, Coimbatore.
https://orcid.org/0009-0006-0269-6891
rajeswaribalaji2412@gmail.com
[4]Assistant Professor, Kristu Jayanti Institute of Technology, Kristu Jayanti University, Bengaluru https://orcid.org/0000-0002-7111-2494
bharathiramkumar@gmail.com
[5]Associate Professor, Department of Information Technology, Hindusthan College of Engineering and Technology, Coimbatore.
https://orcid.org/0000-0002-8762-2149
ganeshktg@gmail.com
[6] School of business and management, Christ university, Bangalore https://orcid.org/0000-0001-7915-0180
Manikandan.rajagopal@christuniversity.in

Abstract: Advances in deep learning have significantly improved text-to-speech (TTS) synthesis, revolutionizing voice recognition technologies and applications. This study investigates the impact of state-of-the-art deep learning approaches on TTS systems, focusing on enhancing speech naturalness, intelligibility, and usability. Specifically, this work explores advanced neural network architectures, including transformer models and generative adversarial networks (GANs), that have set new benchmarks in generating high-quality synthetic speech. Through a comparative evaluation of these methods in diverse communication scenarios, our work highlights the practical potential of modern TTS synthesis techniques. It provides insights into their future trajectory toward enabling more inclusive and effective conversational systems.

Keywords: Text-to-Speech Synthesis, Deep Learning Techniques, Communication Recognition

1. Introduction

Today, when pictures are so common, receiving information is more critical than ever. Because more than two billion people worldwide have trouble seeing, it is essential to ensure that everyone can see everything [1]. Audio description is beneficial for getting around, especially for people who are blind or have low vision. This is important for inclusion because it ensures that people can hear and see what is being shown. Sound description is essential for accessibility because of these main points: Everyone should have access to information. It's just as easy for blind or low-vision people to watch live events, TV shows, pictures, and educational movies with sound. Being a part of society: [2] The Audio gives people who are blind or have low vision a sense of community and connection by letting them enjoy cultural and social activities like art shows, sports games, and museum tours with other people. [3] This helps people understand their surroundings and makes walking safer. [4] Accessibility in the learning process: In a classroom, maps and plots can be shown visually with voice-overs to explain them. It is essential for kids who are blind or have low vision to participate fully in all of their classes.

Getting a job: Job titles must be accessible for blind or low-vision people [5]. This supports equality and similar job opportunities by letting people connect with slide shows, maps, and other work-related

materials. Changing picture data into sound data not only helps blind or low-vision people but also makes things easier for more people to use: Multiple tasks and ease of use [6]. Audio comments are helpful for people who are busy or who can't focus on visual information because they let them hear information while doing other things. Language diversity: Audio explanations are helpful for people who don't speak the same language or don't understand visual details as well as others because they are spoken and can get around language barriers [7]. Ways of learning: Different people learn differently; some may know more or prefer to hear things. Audio description is a way for people who are blind or have low vision to connect with the outside world. This is more than just easy to get to. This is the most essential thing that needs to be done to create a society where everyone, no matter their skills, can fully participate in every aspect of life. [8] People with trouble seeing can fully understand visual material and turn it into audio comments thanks to this partnership between NLP and CV (Computer Vision) technologies. Intelligent data analysis is one of the best aspects of this combination. NLP gives the language background for processing the visual information in a CV, which helps people understand pictures better and more deeply.

Text-to-speech (TTS), picture-to-speech, and the new areas of Text-to-Image and Picture-to-Speech are all looked at in terms of how AI is used now and in the future. This is shown in Fig. 1. The piece goes into great detail, is current, and is jam-packed with valuable facts [9]. There are different datasets for direct and indirect Image-to-Speech jobs. This article focuses on what makes these datasets unique and how they can be used. The piece talks about how datasets are used to train and test models and how important they are in many areas of study. [10] Researchers, practitioners, and creators in these areas can learn a lot from it.

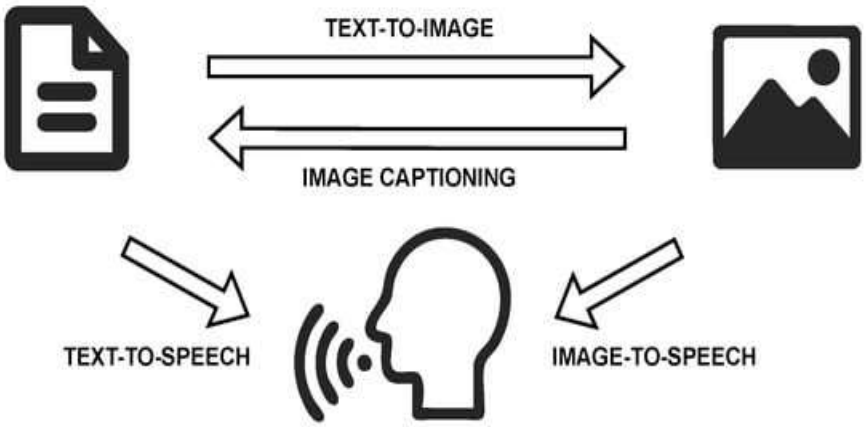

Fig. 1 Diagram showing relations between image captioning tasks, text-to-speech, image-to-speech, and text-to-image.

The initial part of this piece talks about how NLP and CV work together to turn visual data into applicable voice statements and gives examples of how AI-based virtual help has been used successfully. [11] The second part makes a well-organized list of methods by grouping them by how well they can handle different data types, like words, pictures, and sounds. A picture is discussed in the first part. Visual data represents data using standard graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand. Image captioning is the task of predicting a caption for a given image. In image-to-text, the detected text regions undergo OCR (Optical Character Recognition) to digitize the text regions and to detect and extract the text from the image. Finally, the detected text is converted to speech using a text-to-speech synthesizer. This is an essential part of AI, where words and visual information come together, next, in Text-to-Speech technologies [12], which have changed from old-fashioned composition methods to neural TTS based on deep learning methods. Recent research on Image-to-Speech systems is also discussed. TTS is a technology that converts text into spoken words. It's an assistive technology that can read text aloud on digital devices like computers, smartphones, and tablets. TTS can be helpful for many people, including those with reading disabilities, visual impairments, or light sensitivity. These systems let computers turn pictures into spoken explanations without any writing. [13] Finally, the third part compares datasets important to three critical areas of AI: tasks that mark up pictures, make text into speech, and turn photos into speech. Text-to-image is an AI that uses a text description to generate an image.

As a technical contribution to the field of Audio, speech, and language processing, the manuscript "Advancing Text-to-Speech Synthesis using Efficient Deep Learning Techniques for Enhanced

Communication Recognition" introduces a framework for advanced text-to-speech (TTS) synthesis that is based on generative adversarial networks (GANs). This Method advances state-of-the-art technology by addressing essential problems in TTS, such as making synthetic speech more natural and understandable while maintaining computing efficiency. Effective modelling of complicated prosody and reduction of artefacts is achieved by integrating GANs, resulting in high-quality speech production suited for practical applications such as interactive systems and assistive technologies. The paper uses generative adversarial networks (GANs) as a machine learning methodology. Still, it stays within the journal's purpose by applying this Method directly to make TTS systems more expressive, natural, and efficient.
The main contribution of the paper

- Introducing a GAN-based TTS synthesis framework that enhances synthetic speech's naturalness, expressiveness, and prosody while reducing artefacts, addressing key limitations in existing TTS systems.
- By emphasizing computational efficiency, the proposed Method enables real-time deployment in resource-constrained environments such as mobile devices and embedded systems, broadening its applicability.
- This advancement has potential applications in accessibility technologies, virtual assistants, and human-machine communication systems, significantly enhancing user experiences across healthcare, education, and customer service industries.

The article is structured as follows: Section 2 discusses related works. Section 3 deliberates the proposed framework. Section 4 compares the proposed model with existing models. Section 5 concludes the research study.

2 Related Work

Speech synthesis is a device that turns written words into speech waves [14]. Technology for synthesizing speech is used in many areas, including AI helpers, guidance systems, and audiobooks. Also, because speech synthesis technology is used to trick automatic speaker verification systems (ASVs), researchers are looking into voice-faking detection technology to stop attacks that use that technology. Training data, which is real synthetic speech, is required to find fake attacks that use synthetic speech. To do this, people are looking into how to make things sound natural. It is being looked into in many different areas as neural network-based speech synthesis technology grows. Speech that is put together is now a lot better than it used to be. Most models that use neural networks to make speech have two main steps. Making a model that can take text and turn it into time-frequency audio features is the first thing that needs to be done. A vocoder model turns the sounds into simple waves [15].

A voice synthesis model based on neural networks is Tacotron, of course. It moves from one order to the next based on attention [16]. End-to-end, Tacotron is a machine that makes words without any sound features in between. Attention technology in Tacotron lines up text and speech lengths, making voice synthesis more accurate. Tacotron uses an autoregressive method, meaning its manufacturing speed is slow. There are also problems with wrong speech or skipping phonemes because it depends on how long the attention module lasts. FastSpeech is a transformer-based model that gets around these problems using a non-autoregressive method that gives it a fast inference speed. With the time estimate, you can guess how well the mel-spectrogram and text match up. This lets you fix wrong pronunciation and make speech synthesis sound more natural. On the other hand, the alignment found by the autoregressive model is used as a name when learning the time prediction.

A monotonic alignment search (MAS) was added to Glow-TTS, a speech synthesis model that uses normalizing flows [17]. MAS aims to find the best orientation that doesn't change. This is possible because the text and the mel-spectrogram are always lined up similarly. This means that phonemes are present. When the MAS looks at how the text was spread out before and how it is saved, it finds the log chance. Once the log chance is known, this Method will find the most likely straight line between the text and the mel-spectrogram. If you speak this way, you can make clear speech without skipping or leaving out phonemes.

Models built on the denoising diffusion probability model (DDPM) have recently shown excellent results in several areas [18]. In DDPM, data is created using a forward process that spreads noise over the original data over time to make it easier to work with. A backwards process is then used to change the data back into the original data. A lot of different areas have shown that these DDPMs work very well. In text-to-image, DaLLE2 and steady diffusion models work very well. DiffWave models change sound features into audio waves in the audio field. If you want to give a speech with DDPM, Grad-TTS works excellently. Grad-TTS also uses a score-based decoder to show the forward and backward processes. It combines speech and random differential equations. A numerical ordinary differential equation (ODE) algorithm makes the data

based on the expected scores [19]. Grad-TTS can make a more realistic speech than other speech synthesis models. It can also take speaker ID as input and allow speech synthesis with more than one speaker. That being said, since it needs the speaker's ID, it can only make words for speakers that are seen or collected. In other words, the model can't handle the speaker as well since it only gets basic details about it. This can be used to make speaker words that you can't see. With a speaker encoder, you can get information about the speaker from the words. This is called speaker encoding.

Speaker identification is a tech tool that turns information from words into names of people. Neural network-based speaker recognition models determine who is speaking by taking speaker embeddings from speech and figuring out how similar two vectors are. Models taught with big speech datasets can also learn different speaker information, which lets unknown speakers recognize speakers. [20] The ECAPA-TDNN is a simple model for recognizing speakers. From these studies, changing how well the speaker recognition model works can improve zero-shot multi-speaker speech synthesis. When models allow zero-shot learning, they often use a speaker decoder already trained on significant speech datasets to get information about the speaker.

The Glow-TTS model is used by SC-GlowTTS [21], which can talk. A speech processor reads the sample text to find out about the person. The words of the speaker are based on what was learnt about them. You can combine words from many people, even those you can't see. YourTTS can also make words from start to finish. It is a zero-shot model for synthesizing speech from multiple speakers based on VITS and combines VAE with flow normalization. But this is one of their problems. SC-CNN suggested an excellent way to fix the problem of speech synthesis performance differences between speakers who can be seen and speakers who can't be seen. [22] It sets up speaker embeddings in a way that considers how closely two nearby phonemes are linked and gives speaker-specific local correlations. This is the best way to synthesize zero-shot multi-speaker speech to train speakers. This could happen if there is more training data or the model must be intelligent enough to learn the language.

In the same way, NANCY is a popular way to use different types of information to turn speaker information from words. That voice change works well. NANCY created information modification. [23] Changes to the pitch, formant, and frequency of speaking can be used to change details about the person. Before NANSY, the AUTOVC voice translation model had a way of changing the bottleneck factor so that it could handle information on the person and the situation. The bottleneck measurement needs to be carefully altered for this Method to work. The speech translation will only work correctly if the choke measurement is correct. Nansy offered a way to change sounds based on disturbing the voice. This Method uses information perturbation to get rid of information in speech that is specific to the speaker.

In this work, we have improved Grad-TTS and created an excellent zero-shot multi-speaker speech synthesis model. Still, this Method could be better. No one can change facts about a person; speech synthesis can only be used for people who can be seen. This work will use a speaker decoder and information change to help zero-shot multi-speaker speech synthesis. [24] Based on the Grad-TTS structure, the new Method has shifting data and a speaker encoder to help the model learn how to handle different types of speaker data. A speaker encoder removes speaker embeddings, which are used as factors in the composition process. Zero-shot multi-speaker speech synthesis can work with how it was told to do it. We also used NANSY's new information disturbance tool to make the sound models of people who can't be seen better. NANSY framework that can manipulate an arbitrary speech signal's voice, pitch, and speed. Training it with a multilingual dataset allows NANSY to extend to a multilingual setting easily. It aims to design a neural analysis and synthesis (NANSY) framework by decomposing a speech signal into analysis features that represent pronunciation, timbre, pitch, and energy. Information disturbance considers different types of people from the point of view of speech synthesis in this study. So, by changing some data, the model is trained to look at different kinds of speaker data. This lets it learn more speakers than are in the collection. It differs from Grad-TTS because of the things said about the speaker. With the Method suggested, you can get speaker embeddings from the speech source. You get the speaker ID with Grad-TTS. In this way, the model can use reference speech to show details about individuals who can be seen and those who can't. [25] There are better ways to handle speaker information than the current speaker ID, but it is a better way to model speaker traits. The model is taught to look at information about people other than the ones in the dataset by changing information. This differs from how Grad-TTS is trained now because it gives the model more chances to learn about the users. This makes the speakers it hasn't seen before more like speakers it has seen before, which makes the model better at generalization. The standard Grad-TTS method was used against the suggested Method to check how it performs for the speaker. It also examined how the secret speaker synthesized sounds by comparing SC-GlowTTS and YourTTS..

3 MIXTURE-TTS

First, this work discusses why Mixture-TTS was made. Then, it provides the architecture of Mixture-TTS

and how the loss functions are put together. Mixture-TTS is a non-autoregressive speech synthesis model based on a mixture alignment mechanism. It aims to optimize the alignment information between text sequences and the mel-spectrogram. Mixture-TTS uses a linguistic encoder based on soft phoneme-level alignment and hard word-level alignment approaches, which explicitly extract word-level semantic information and introduce pitch and energy predictors to predict the rhythmic information of the Audio optimally.

3.1 Motivation

The Montreal Forced Aligner (MFA) tells FastSpeech2 how long phonemes last, making it hard to match them properly because they get fuzzy. The main aim of this research work is Text-to-Speech Synthesis, which uses efficient deep-learning techniques for enhanced communication recognition. Several research methodologies have been introduced, but accuracy has not improved significantly. The mel-spectrogram doesn't show a clear line between the different phonemes in the existing system. To overcome the abovementioned issues, this research proposes a mixed alignment method to improve the overall system performance. Soft alignment is used for phonemes, and challenging alignment is used for words in the linguistic encoder. The proposed Method is used to provide more accurate results using practical algorithms. Picture 2 shows the language generator on Mixture-TTS, which is the same as in PortaSpeech.

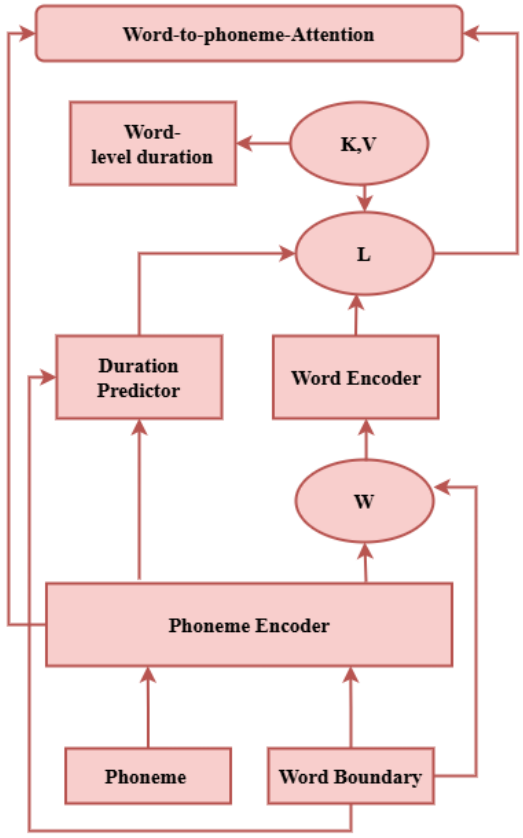

Fig. 2 Linguistic encoder architecture for PortaSpeech

A linguistic encoder consists of a phoneme encoder, a word encoder, a duration predictor, and a word-to-phoneme attention module. The phoneme encoder and the word encoder are both stacks of feed-forward Transformer layers

with relative position encoding, as shown in Fig. 2. Tacotron2 is a TTS model built on a convolutional neural network that did great things in real time. As the last step, Tacotron2 adds a post-net network comprising a 1D convolution network with five layers. Tests have shown that this makes it easier to change the mel-spectrogram. An essential part of voice synthesis is the mel-spectrogram, which changes the sound quality. This does, however, improve the prediction data for the mel-spectrogram. There is a part of the experiment where we add the same post-net network to the Mixture-TTS to ensure it works.

3.2 Basic Model Architecture

As shown in Fig. 3, it has a post-net, a transformer decoder, and a language encoder. This section analyzes the Mixture-TTS's organization and training loss in depth.

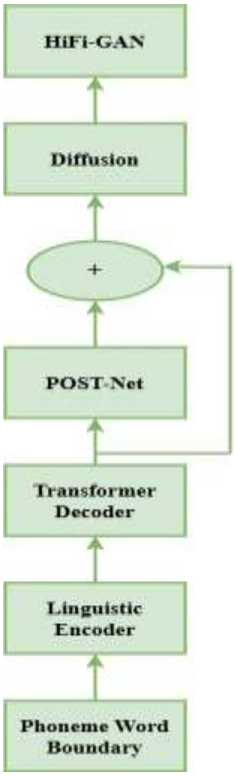

Fig. 3 (a)

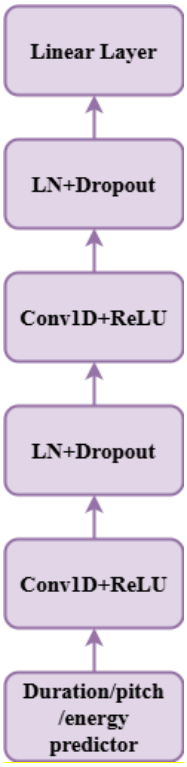

Fig. 3 (b)

Fig. 3 The structure of Mixture-TTS as a whole.

A better beat from the synthesized sound will be made if we add pitch and energy forecasts to the language used. Natural sounds teach the pitch and energy models what to do. As an extra input, this gives the reasoning stage enough knowledge about the difference. Fig. 3b shows how the language decoder is put together. The "L.R." symbol sets the length, the "W.P." symbol groups words together at the word level, and the "sinusoidal-like" symbol shows where something is about other things. For the language generator to work, we give it the phoneme patterns that show where words start and end. The encoder takes patterns of phonemes and changes them into hidden phonemes. The secret state of the phoneme is made more diverse by pitch and energy factors. It's also possible to shorten the secret word states by controlling the length at the word level. This makes them the same length as the goal mel-spectrogram. Lastly, the word-level relative position is used to keep track of the secret word-level and phoneme states. Next, these are sent to the program that pays attention to word sounds.

It comprises the same parts as the decoder, the word encoder, and the letter encoder. All three are based on the Fast Fourier Transform (FFT), and the recurrent neural network framework is no longer there. We set the decoder's FFT to 6 and the phoneme encoder's and word encoder's FFT to 4. It depends on the signal and defines the number of bins to divide the window into equal strips or bins. Hence, a bin is a spectrum sample, determining the window's frequency resolution. It is set to 9 for the size of the convolution kernel and 0.2 for the dropout in the multi-head focus engine. There are two heads inside the machine. Fig. 4a shows how the

The FFT section is put together. The multi-headed attention process is used to discover the phoneme patterns' surroundings. Multi-headed attention is a technique used in computer science to model the complex relationships between different token entities. It involves transforming an input sequence into multiple groups, each undergoing a self-attention process. It is a module for attention mechanisms that runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension. The residue network then links the incoming phoneme patterns to the data they hold after dropping out unnecessary ones and making them normal. The background data is then put together by a 1D convolution network with two layers. It is sent to the next section after failure and normalization.

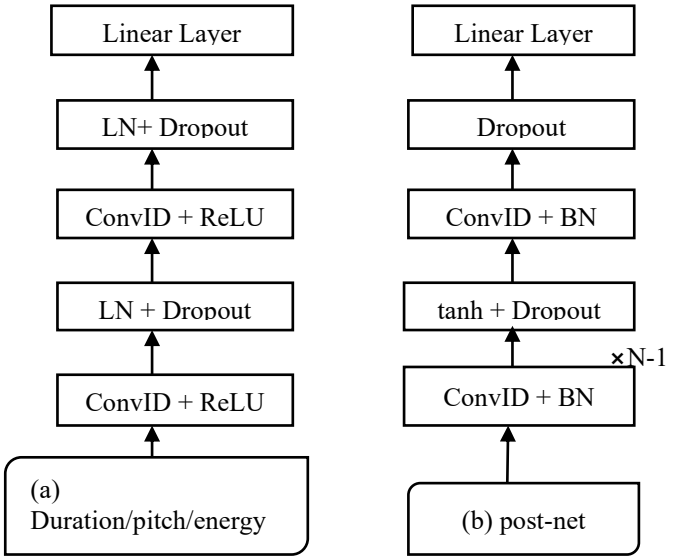

Fig. 4 (4a and 4b) Structure of FFT module and post-net

Fig. 4 illustrates that FFTs are mainly used to visualize signals. However, there are also applications where FFT results are used in calculations. Post-net is useful for machine translation, named entity recognition, and information extraction, among other things. It also works well for clearing out ambiguity in terms of numerous meanings and revealing a sentence's grammatical structure. It identifies a word's grammatical group, i.e., whether it is a noun, adjective, or verb. An extranet added by Mixture-TTS helps rebuild the mel-spectrogram better as a whole. The residue network connects the encoder's output to the post-net's production. Five levels make up the post-net, which is a 1-D neural network. You can see a picture of it in Fig. 4b. There are 512 5 × 1 convolution kernels in all but the last layer. Each layer also has batch normalization (B.N.), dropout, and a Tanh activation function. The result of the previous layer is linearly mapped to get the mel-spectrogram.

3.3 Training Loss

To train Mixture-TTS, we get the total difference between the fundamental values and the expected values as small as possible:

$$\tau\text{total} = \tau\text{met} + \tau\text{postnet} + \tau\text{duration} + \tau\text{helper} + \tau\text{pitch} + \tau\text{energy} \quad (1)$$

4 Evaluation Method

At the moment, there are different ways that text-to-speech tools are evaluated. However, not all ways work; some are only good for specific situations, like evaluating one-word synthesis. Because of this, picking the correct Method is difficult.

Model review can be done on two main types of tests most of the time:

Tests of your opinion. These are the daily tests where everyone rates how good their speech is and how good it is.

Tests with goals. These tests used the right speech signal processing methods to measure the voice's ability.

That being said, the TTS system is judged by two main criteria:

Intelligibility is a measure of how correctly the words are understood. The following tests can be used to judge this metric:

The DRT uses pairs of words that sound the same but are different in one way (phonologically). The number of correct answers indicates how easy something is to understand.

The subjective test is called the Modified Rhyme Test (MRT). It's the same as the last test, but the six words differ.

The words in the lines on the Semantically Unpredictable Sentence (SUS) test were picked at random. Using this Method, the following formula (2) can be used to check the intelligibility:

$$\text{SUS} = 100 \times \frac{C}{S \times L} \quad (2)$$

L is the number of people who heard. That's how many right words will be said. The number S shows how many lines were tried.

A quality score that shows how natural, smooth, and precise the speech is. The following ways can be used to judge this metric:

People take the A.B. test, a subjective test in which they compare the results of systems A and B to determine which audio files with synthesized speech are of higher quality (Table 1). The result shows which Method is better.

People use the ABX test to judge how similar the words made by system A and system B are to the original voiced sentence X.

People rate each line the machine wrote on a scale of 1 to 5, with five being the best and one being the worst (Table 2). The MOS stands for "Mean Opinion Score."

Table 1. A.B. test score number

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

TABLE 2. MOS MEASURE FOR RATING

Rating	Speech Quality	Level of Distortion
5	Good	Not noticeable
4	Good	This is noticeable but not bothersome.
3	Fair	Noticeable and a little annoying
2	Poor	It's annoying but not unacceptable
1	Bad	Very annoying and unpleasant

Then, the total score is added and split by the number of graded words and the number of viewers.

This objective test, called Mel Cepstral Distortion (MCD), looks at sequences of recovered mel-frequency cepstral coefficients that have been synthesized and sequences that have been found naturally. This difference shows that the copied speech sounds more like natural speech. To find MCD, use the following Method:

$$\text{MCD} = = \frac{10\sqrt{2}}{\text{In } 10 \times \text{T}} \sum_{t=0}^{T-1} \sqrt{\sum_{t=s}^{D} (\text{Vd targ (t)} - \text{Vd ref(t)})2} \quad (3)$$

There are 24 measures in all. Since d ranges from 0 to 24, the meal frequency cepstral factors are $\boldsymbol{X}tbrbd$ and $\boldsymbol{X}ref\,d$.

An objective test called SNRseg measures the noise ratio between two frequencies. The following Method (4) can be used to find this ratio:

$$= \frac{10}{\text{M}} \sum_{M=0}^{M-1} 1\text{g} \qquad = \frac{\sum_{n=NM}^{NM+N-1} x^2\,(n)}{\sum_{n=NM}^{NM+N-1} (x\,(n) - x\,(n))2} \quad (4)$$

Where $x(n)$ is the original signal, $\underline{x}(n)$ is the signal that was made, N is the length of a frame, and M is the number of frames in the speech signal.

One way to guess what the MOS test results will be is to use the objective test for perception evaluation of speech quality (PESQ). Automating the MOS review speeds up the process. To find PESQ (5), use the following Method:

$$\text{PESQ} = \text{a0} + \text{a1}.\,\text{dsym} + \text{a2}.\,\text{Dasym} \quad (5)$$

The average values for disturbance are dsym and dasym, and the values for asymmetrical disturbance are a0 = 4.5, a1 = -0.1, and a2 = -0.0309.

But the abovementioned rating tools can't promise a correct answer. For more accurate results, all subjective tests need many people to listen. However, many thinking factors can change the results, such as the person's state of mind, the setting of the hearing test, and other things. Objective tests can only partially judge the model since actual and synthesized speech have different prosody and sounds. Because of these changes, synthesized speech may sound less realistic and be more challenging to understand.

Many systems use the most common measure, MOS, to judge synthesized speech, even though it isn't always accurate. Since measures for understanding aren't crucial in business, they only look at how good the speech is.

5 EXPERIMENT

Subjective evaluations were used to rate two main aspects: the quality of the speech sound and how it could be understood. An online poll was given to 34 people, 32 of whom spoke English as their first language and 2 did not. Participants were told to listen to sounds in a quiet room with headphones.

There were 53 lines in the test, and they were grouped by source type and the number of Out of Vocabulary (OOV) words they contained (Table 3). Specifically, five sentences with no OOV words were chosen for the A.B. test. For the MOS test, 14 lines were picked. Three had one out-of-word (OOV) word, six had two or more OOV words, four had three or more OOV words, and one had four. To ensure the SUS test was fair, 10 lines were put together for each system. There were three lines with no OOV words, four with one OOV word, and three with two OOV words. There were also 20 lines of random words just for the SUS test. Twenty of the lines came from news sources, eight came from different storybooks, and the rest were made up of lines taken from news sources.

TABLE 3. CLASSIFICATION OF TEST SENTENCES DEPENDS ON THE NUMBER OF OOV WORDS

Test Type	No 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 the A-stage	14		3	6	4	1
MOS for Group B	14		3	6	4	1
SUS for plan A	10	3	4	3		
SUS for part B	10	3	4	3		

The A.B. measure is utilized to compare how natural and clear the speech made by the two systems was. This test determines how much people liked system B (DC TTS) or system A (Tacotron). People were asked to rate fake sounds on a three-point scale. When the score was 0, there wasn't a lot of difference between A and B. There was a good choice if the number was three and the type of Method was either A or B. If the number was 1, people picked that Method over another one by a small amount (Table 3). The numbers are then split between all the people who took the test and all the A.B. test lines.

The sounds made by both systems were played at the same time in the A.B. test. But, in the MOS test, the synthesized sounds of each system were played separately so that people could see how good the speech was and how much confusion there was (Table 4). Each system had a different set of lines, but the number of OOV words in each sentence was kept the same, so there would be no bias in the real world. One last thing that was done to find the system's test signal quality was to add up all the scores given by users and then divide the total by the number of test lines.

The SUS test is all about random words that can be combined to make correct grammatical lines. But these words don't mean much. This differs from the SUS test, where people guess words by reading the whole text. Instead, they type what they hear. It's clear what the test lines mean now. Last, Formula 7 was used to determine System A and System B's SUS scores.

6 RESULTS AND DISCUSSION

The data show that Tacotron needs to be out when looking at the average scores of evaluation measures. On the other hand, DC TTS gets better results for lines with more OOV words. First, we looked at the A.B. number to see how good each Method was. For Tacotron, it was 0.90; for DC TTS, it was 0.59. In the A.B. test, Tacotron did better than DC TTS, as shown in Fig. 5. Along a scale from 0 to 3, the A.B.

The number is close to 1. This means that Tacotron's sounds are "good." The score from DC TTS also says that the synthesized sounds are "good" to "almost the same as another system."

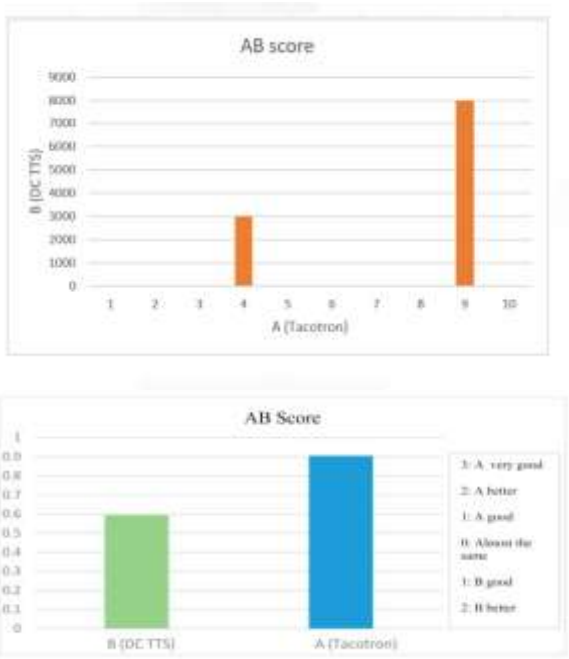

Fig. 5 Scores for A (Tacotron) and B (DC TTS)

There are two ways to find the MOS (95% Confidence Interval), and you can compare them in Table 4. There is a slight difference between the MOS for DC TTS (3.36 ± 0.187) and 3.49 ± 0.193 for Tacotron. So, both ways make synthesized sounds that are "fair," and the distortion can be heard but isn't too annoying. Not looking at the average MOS of all sentences and the average MOS of lines grouped by how many words were missing (Fig. 6). By adding up the number of OOV words used in each test phrase, the average opinion score for each person was found. When putting together lines with three or four OOV words, DC TTS did better (≤3.0). It is suggested that the Azerbaijani TTS system will most likely use a mixed model. To make sentences with words from the train data, this model would use Tacotron. It would use DC TTS to create lines with more words that don't belong together.

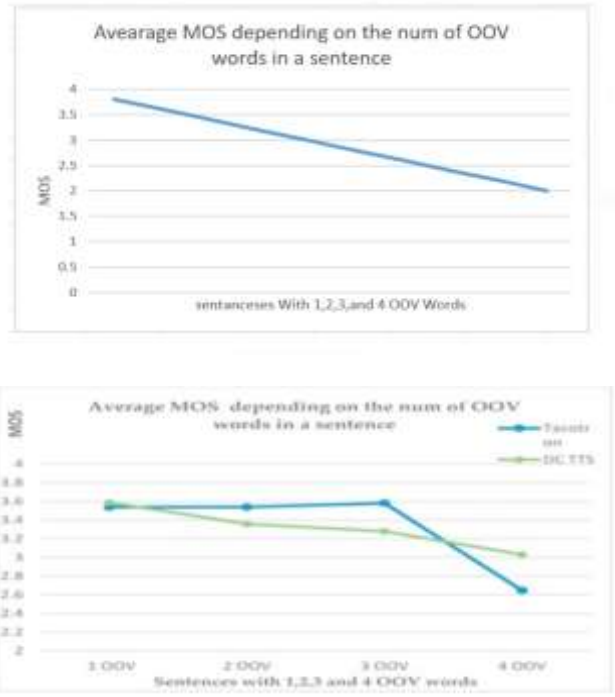

Fig. 6 Average MOS for Tacotron and Deep Convolutional (DC) TTS and sentences with OOV words.

TABLE 4. COMPARE THE MEAN OPINION SCORE (MOS) FOR DC TTS AND TACOTRON

System	MOS (95% CI)
Tacotron	3.49 ± 0.193
DC TTS	3.36 ± 0.187

Lastly, Table 5 shows how many times each system passed the SUS test. DC TTS could only get 45% of the test sentences right, but Tacotron could get 14.9% of the test sentences that didn't mean anything correct.

TABLE 5. COMPARE THE SUS SCORES

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

Limitations: The study has a flaw in that the test words used in the polls needed a bigger sample size. It took viewers more than 40 minutes to rate 53 sentences of synthesized Audio. One option could be to divide the poll questions into groups based on the type of review measures.

7 CONCLUSION

Text-to-speech synthesis uses efficient deep learning techniques for enhanced communication recognition in this research. The 24-hour single-speaker data was used to train the DC TTS and Tacotron models. Test lines were grouped based on how many out-of-words they had to compare with the suggested models. This work discusses subjective rating metrics that can be used to rate the intelligence and quality of systems that make synthesized speech. The AB, MOS, and SUS measures showed that Tacotron's test phrases had better results. Better MOS results for Tacotron were found in sentences with more words from the training vocabulary. Better MOS results for DC TTS were found in sentences with more words outside the training vocabulary. Future studies should consider using a blended model that takes the best parts of both systems to make speech sound more natural and understandable.

Acknowledgment
This work was not supported by any funding agency.

Competing Interest
The author has no conflicts of interest to declare that are relevant to the content of this article

REFERENCES

[1] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep Learning Enabled Semantic Communications With Speech Recognition and Synthesis," IEEE Transactions on Wireless Communications, vol. 22, no. 9, pp. 6227–6240, Sep. 2023, doi: 10.1109/twc.2023.3240969.

[2] K. S. Sri, C. Mounika, and K. Yamini, "Audiobooks that converts Text, Image, PDF-Audio and Speech-Text : for physically challenged and improving fluency," 2022 International Conference on Inventive Computation Technologies (ICICT), pp. 83–88, Jul. 2022, doi: 10.1109/icict54344.2022.9850872.

[3] N. Kaur and P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: a review," Artificial Intelligence Review, vol. 56, no. 7, pp. 5837–5880, Nov. 2022, doi: 10.1007/s10462-022-10315-0.

[4] H. Pashler, M. McDaniel, D. Rohrer, and R. Bjork, "Learning Styles," Psychological Science in the Public Interest, vol. 9, no. 3, pp. 105–119, Dec. 2008, doi: 10.1111/j.1539-6053.2009.01038.x.

[5] M.-F. Moens, K. Pastra, K. Saenko, and T. Tuytelaars, "Vision and Language Integration Meets Multimedia Fusion," IEEE MultiMedia, vol. 25, no. 2, pp. 7–10, Apr. 2018, doi: 10.1109/mmul.2018.023121160.
[6] J. He, "Voice emotion recognition using natural language processing deep learning", doi: 10.17760/d20476862.
[7] A. Mogadala, M. Kalimuthu, and D. Klakow, "Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods," Journal of Artificial Intelligence Research, vol. 71, pp. 1183–1317, Aug. 2021, doi: 10.1613/jair.1.11688.
[8] A. DiLodovico, "Review of More Than Meets the Eye: What Blindness Brings to Art by Georgina Kleege," Disability Studies Quarterly, vol. 41, no. 1, Mar. 2021, doi: 10.18061/dsq.v41i1.7844.
[9] J. Snyder, "Audio description: The visual made verbal," International Congress Series, vol. 1282, pp. 935–939, Sep. 2005, doi: 10.1016/j.ics.2005.05.215.
[10] J. Snyder, "Audio description in the United States," The Routledge Handbook of Audio Description, pp. 531–543, Feb. 2022, doi: 10.4324/9781003003052-42.
[11] P. Orero and A. Vilaró, "Chapter 11. Secondary elements in audio description," Audio Description, pp. 199–212, Oct. 2014, doi: 10.1075/btl.112.12ore.
[12] "Studying User Behavior on Twitter & Instagram Using Computer Vision & Natural Language Processing," 2019, doi: 10.4135/9781526493415.
[13] S. Bisser, "Introduction to the Microsoft Conversational AI Platform," Microsoft Conversational AI Platform for Developers, pp. 1–24, 2021, doi: 10.1007/978-1-4842-6837-7_1.
[14] "Turnbull, (Wilson) Mark, (1 April 1943–19 May 2016), Principal, Mark Turnbull Landscape Architect, since 1999; Chairman, Envision 3D (formerly TJP Envision, then Envision), 1999–2016," Who Was Who, Dec. 2007, doi: 10.1093/ww/9780199540884.013.u38195.
[15] R. Rastogi and D. T. V. Pawluk, "Toward an Improved Haptic Zooming Algorithm for Graphical Information Accessed by Individuals who are Blind and Visually Impaired," Assistive Technology, vol. 25, no. 1, pp. 9–15, Mar. 2013, doi: 10.1080/10400435.2012.680659.
[16] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," Interspeech 2020, Oct. 2020, doi: 10.21437/interspeech.2020-2650.
[17] J. Huh et al., "The VoxCeleb Speaker Recognition Challenge: A Retrospective," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 3850–3866, 2024, doi: 10.1109/taslp.2024.3444456.
[18] E. Cooper et al., "Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, doi: 10.1109/icassp40776.2020.9054535.
[19] E. Casanova et al., "SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," Interspeech 2021, Aug. 2021, doi: 10.21437/interspeech.2021-1774.
[20] E. Casanova et al., "SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," Interspeech 2021, Aug. 2021, doi: 10.21437/interspeech.2021-1774.
[21] Y. Chen, J. Liu, L. Peng, Y. Wu, Y. Xu, and Z. Zhang, "Auto-Encoding Variational Bayes," Cambridge Explorations in Arts and Sciences, vol. 2, no. 1, Feb. 2024, doi: 10.61603/ceas.v2i1.33.
[22] Kim, J., Kong, J., & Son, J. (2021, July). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In International Conference on Machine Learning (pp. 5530-5540). PMLR.
[23] H. Yoon, C. Kim, S. Um, H.-W. Yoon, and H.-G. Kang, "SC-CNN: Effective Speaker Conditioning Method for Zero-Shot Multi-Speaker Text-to-Speech Systems," IEEE Signal Processing Letters, vol. 30, pp. 593–597, 2023, doi: 10.1109/lsp.2023.3277786.
[24] Choi, H. S., Lee, J., Kim, W., Lee, J., Heo, H., & Lee, K. (2021). Neural analysis and synthesis: Reconstructing speech from self-supervised representations. Advances in Neural Information Processing Systems, 34, 16251-16265.
[25] Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019, May). Autovc: Zero-shot voice style transfer with only autoencoder loss. In International Conference on Machine Learning (pp. 5210-5219). PMLR.