iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39689

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Deep Learning Based TTS-STT Model with Transliteration for Indic Languages

Kartik Tiwari[2], Shruti Nikam[2], Kalyani Vidhate[3], Astitva Ghanmode[4], Omkar Dalwai[5], Ranjana Jadhav[6]
[1, 2, 3, 4, 5, 6]*Department of Information Technology , Vishwakarma Institute of Technology, Pune*

Abstract: ***This paper introduces a new text-to-speech presentation from end-to-end (E2E-TTS) using toolkit called ESPnet-TTS, which is an open source extension. ESPnet speech processing tools kit. Various models come under ESPnet TTS TacoTron 2, Transformer TTS, and Fast Speech. This also provides recipes recommended by the Kaldi speech recognition tool kit (ASR). Recipes based on the composition combined with the ESPnet ASR recipe, which provides high performance. This toolkit also provides pre-trained models and samples of all recipes for users to use as a base .It works on TTS-STT and translation features for various indicator languages, with a strong focus on English, Marathi and Hindi. This paper also shows that neural sequence-to-sequence models find the state of the art or near the effects of the art state on existing databases. We also analyze some of the key design challenges that contribute to the development of a multilingual business translation system, which includes processing bilingual business data sets and evaluating multiple translation methods. The test result can be obtained using tokens and these test results show that our models can achieve modern performance compared to the latest LJ Speech tool kit data.***

Terms of Reference — Open source, end-to-end, text-to-speech

I. INTRODUCTION

Language translation and transliteration is increasingly being used in all aspects of daily life, this because the access to communication is faster, cheaper and readily available, enabling people all over the world to communicate directly with each other also can officially understand one another, with the meaning distinctly clear and succinct. [1]

TTS or Text-to-Speech technology translates text into spoken speech . TTS can make life effortless and make you more efficient. It helps illiterate students and other disabled learners to learn by removing the pressure to read and present information appropriately. This project aims to tackle the issue with translating the contents of the page in the selected language. It also has a speech to text feature, where the browser speaks in the selected language.

Speech is basically, natural form of communication. Speech is very common way for humans to interact because it is the most effective way to communicate, it can be also extended further to interact with the system. If any part of the society is not able to communicate because of Speech disability, technology can be utilized to make this happen. Besides Speech Recognition can be adapt for the individual having Speech disability, If Speech Technologies is a built-in communication system, it can help individuals, who can use a common type of communication to produce the right Speech. Speech Recognition is the capability of machine/program to recognize words and phrases in spoken language and convert them into machine-readable format ,it is a field of research aimed at converting audio into a series of related words.

Transliteration — the translation of precise nouns from one orthographic system to another — is an important function in multilingual text processing, useful for programs such as online mapping and as part of typewriters. It is a challenge to translate words and technical terms from user input into all languages in different alphabets and phonics lists. One of the most common problems translators have to deal with is translating the correct words and technical terms into a user's input. These objects are often translated, that is, they are replaced by phonetic equivalent. For example, the local name "कंचनपुर" in the Devanagari language is translated "Kanchanpur" into English. Translating such material from English into Devanagari is even more challenging, as the translated material creates a vast amount of text that is not available in bilingual dictionaries.

II. LITERATURE REVIEW

This paper proposes the idea of having Text-to-Speech and Speech –to-Text for some Indian languages. Mihaela Rosco and Thomas Breuel [15] has proposed a sequence to sequence neural network models for transliteration in which model based on epsilon insertions , CTC alignment and attentional sequence to sequence model used in end to end machine translation also achieved high performance on cross script transliteration

Anand Arokia Raj and TanujaSarkar[2] have discussed paper that addressing the issues of Font-to-Akshara mapping, pronunciation rules for Aksharas in content and text normalization in the context of building text-to-speech systems in Indian languages. But as Indian language is rich in inflectional and derivative morphology errors occurred due to new words found in the context.
Then also The final performance of the text normalization are 96.60%,96.65% and 93.38% for Telugu, Hindi and Tamil languages respectively and in future scope Syllable level features could be used to build a text normalization system whose performance is significantly better than the word-level feature.
Nimisha Srivastava and Rudrabha Mukhopadhyay[3] have discussed text-to-speech for three different Indian languages namely Hindi, Malayalam and Bengali and trained a state-of-the-art TTS system for these languages and reported their performances. The advantage is that collected corpus, code, and trained models are made publicly available and Facilitates the training of neutral network based text-to-speech models.
Mano Ranjith Kumar M, Karthik Pandia, Jom and Hema proposed an model where The audio from lecture is transcribed using automatic speech recognition also HMM-GMM are used.[4] The advantage is A novel technique improves the naturalness of auto-trans created videos but Professional dubbing is an expensive and labor intensive process. Result indicate that accurate alignment at the syllable level is crucial for lip-syncing and Using visual speech units alongside syllable segmentation may further improve the observed results for lip-syncing.
Hardik and Sanket has discussed Speech recognition using neural networks which focused on different types of neural networks used for automatic speech recognition and Modern ways are used for recognizing speech although the Capacity of storing the inputs are less. It can be more useful for disabled people.
Song W and Cai J has also proposed End-to-End Deep Neural Network for Automatic Speech Recognition that Developed end to end speech recognition using hybrid CNN and RNN in which hybrid convolutional neural networks are used for phoneme recognition and HMM for word decoding but Much of errors occurs due to incapability of classifiers then also Best model achieved an accuracy of 26.3% frame error on standard core test dataset and planning to investigate thoroughly the reasons for RNN's poor convergence, and perform more hyper parameter tuning.
S.FuruiMusashino[7] proposes a paper based on new isolated word recognition technique based on a combination of instantaneous and dynamic features of the speech spectrum.
Shen, J., Pang, R., Weiss, RJ, Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A [8] discussed formation of network to integrate speech into text. The system is built on a predictive network of continuous sequence-to-sequence feature that puts maps embedding on mel-scale spectrograms, following the modified WaveNet model.
PH Rathod, ML Dhore and RM Dhore [16] has demonstrated machine transliteration for Hindi to English and Marathi to English language using Support Vector Machine i.e. SVM. This approach uses phonetic of the source language and n-gram as a characteristic for transliteration and gives good results also It can further be expanded for foreign names, organization names. As English is non phonetic language, no need to carried out back transliteration , therefore there is also future scope to perform the back transliteration.

III. METHEDOLOGY AND IMPLEMENTATION

The project will also outline real-time E2E-ASR deployment using ESPnet2-ASR, ESPnet2-TTS and ParallelWaveGAN repo. We have chosen to work with ESPnet because of its extensive documentation, community support, and end-to-end framework.

A. ESPnet

ESPnet launched in December 2017 is basically an open source speech processing tool kit that combines speech recognition and integration. Its main idea is to Accelerate future research towards the end of many speech researchers. ESPnet-ST is specially designed to identify ST activity. ESPnet was originally designed for ASR work (Watanabe et al., 2018), and was recently expanded to text-to-speech (TTS) function (Hayashi et al., 2020). Engineers have been very supportive of ESPnet development .This provides the formation of an integrated neural model that leads to the development of specific Mechanical Learning Software . ESPnet has the first Chainer but later became the PyTorch based dynamic neural network toolkit as the engine that led to the development of a neural network novel development. ESPnet has a built-in Automatic Speech Recognition (ASR) mode based on the popular Kaldi project.
ESPnet Functions -1) Has a pre-calculated Kaldi style data that compares the performance of the Kaldi hybrid systems. It uses preliminary data processing developed in the Kaldi recipe. 2) Attribute-based encoder -i) Sampled BLSTM and / or VGG-encoder ii) location-based attention (+10 attention) 3) CTC- i) WarpCTC, label-synchronous recording 4) Hybrid CTC / attention i) Multi-

functional learning, code-sharing 4) Use of language models i) A combination of RNNLM and CTC is an integrated label label / attention design. It has Extensive Support algorithm with End-to-End TTS Algorithms and a working community of TTS Extremely Open with More language support. It has Novel ASR Transformers that provide improved performance.

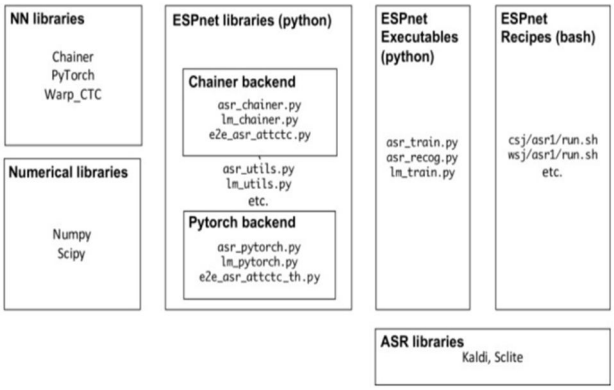


Figure 1 : Software Architecture of ESPnet[18]

ESPnet1 to ESPnet2 - Digging around ESPnet repo, you will probably be confused as it holds ESPnet1 and ESPnet2. Various opportunities and updates from ESPnet1 to ESPnet2. Major Changes in Internal Framework From ESPnet1 to ESPnet2. ESPnet2 is a new DNN training program. Various and major updates dealing with distributed training, in the release of the fly feature, downgrade, software efficiency by developing continuous integration, document enrichment, booth support, plumbing installation, and model zoo functions.

ESPnet2 is the latest release of DNN training and has the following enhancements:

1) It is now independent of Kaldi and Chainer, unlike ESPnet1.
2) Release feature and compliant text processing during training
3) For beginners, it is best to use ESPnet2, and in my opinion, ESPnet1 can be customized.

Although ESPnet has supported the implementation of ASR and TTS as the first speech applications, we recently began to support speech translation (ST) work. Both the standard pipeline method, in which ASR and ASR modules and text-based machine translation (MT) are played, and end-to-end (E2E) modes, in which source speech is translated directly into another language, are readily available at ESPnet ST. ESPnet-TTS consists of two main components: the E2E-TTS neural network model models and recipes that cover all the tests to complete the test. Part of the library is written via Python using PyTorch as a neural network library. The recipes are all-in-one style texts written in Bash and follow the Kaldi style.

B. Text-to-Speech

Text-to-speech (TTS) is a form of assistive technology that reads digital text aloud. It is also called "reading aloud" technology.[2]With the click of a button or the touch of a finger, TTS can take words from a computer or other digital device and convert them into sound. TTS is very useful for children with learning disabilities. But it can also help user with writing and editing, even focusing.

How text-to-speech works

TTS works with almost all personal digital devices, including computers, smartphones and tablets. All types of text files can be read aloud, including Word and Page texts. Even online web pages can be read text. The voice in Text To Speech is generated in computer, and the reading speed can usually be accelerated or reduced. The voice level varies, but some words sound human. There are also many computer-generated voice notes that sound like person is talking. Many Text To Speech tools focus on words as they are read aloud. This allows children to see the text and hear it at the same time. Many Text To Speech tools also have a technology called optical character recognition (OCR). Optical character recognition promotes TTS tools to read text aloud in pictures or frame. For example, user may take a picture of a road sign and the words of the sign will be turned into a sound.

Models-

1) We support three types of E2E-TTS2

TacoTron 2, Transformer TTS and Fast Speech.

Input for each model is a sequence of letters or phonemes and the output is a sequence of acoustic features (e.g. log features of Mel filter bank).Centaur is a handmade model of OpenSeq2Seq.Let us count things in the official storehouse, not including the fork. The text in the acoustic characteristic conversion system is mainly refer to as E2E-TTS . Part of the voucher is not included unless clearly stated .Tacotron2 is a RNN-based model for sequencing and sequencing.[17]

Contains a bi-directional LSTM-based encoder and a unidirectional LSTM-based decoder with sensitive location sensitivity. Unlike the original TacoTron 2, we also support it further attention w / or w / o change agent, helpful to learn diagonal attention. [9] In the case of Transformer TTS, it accepts a multi-headed attention-seeking approach. By replacing the RNN.

A parallelizable self-attention structure, it works faster and faster effective training while maintaining a high level of thinking compared to TacoTron 2. With Fast Speech,[17] design a feed forward Transformer architecture for an uncontrolled generation under the pipeline for training teachers and students.

In addition, to provide TTS performance for multiple speakers, we support the use of speaker embedding as our auxiliary inputE2E-TTS models We use a pre-trained x-vector provided by Kaldi as an embedded speaker.

Types of Text To Speech Architecture - We need to understand the different types of structures we can use to integrate speech and current change.

Concatenative - Old School - A traditional old school method that uses a database of speech where the speech is drawn in specific words. Although with some words included in the map, you may be able to produce understandable Sounds, the outgoing speech will not include natural voice sounds, "prosody, emotions, etc .."

Main stream 2 stage : An integrated TTS parameter method based on the Deep Neural Network that combines an acoustic model with a neural vocoder to achieve parameters and the relationship between input text and the speech form that forms the speech Pre-Text Processing and Customization:

Simply the previous step of text input. It will be converted into vernacular language features embedded in the acoustic model. Convert text into a format that ESPnet can translate. This is done normally (e.g. Aug to August) and converted to phonemes by grapheme-to-phoneme conversion (e.g. August to 2fadaset).

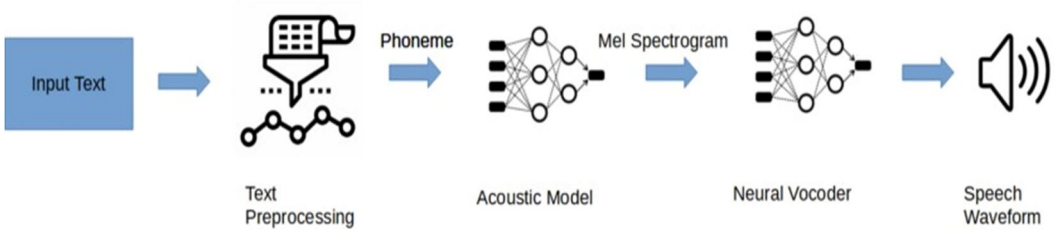


Figure 2 : A basic high-level overview of mainstream 2- Stage TTS System[3]

Acoustic Model - Algorithms have been developed to convert the text that is standard and pre-processed into Mel-spectrograms as final result. With most algorithms you need to convert language features into acoustic features into Mel-Spectrogram. [12] Spectrogram confirms that we have now calculated all the appropriate audio features.

Neural Vocoder - Last step input Mel-Spectrograms are interpreted into a waveform with help of Neural Vocoder. Although there are many different types of neural vocoders, modern ones have a GAN base.

2) Next Generation end to end Text Wave Model

Recent papers on Audio TTS look at this case. A single acoustic model that does not produce Mel-spectrograms that feed on neural vocoders is used.

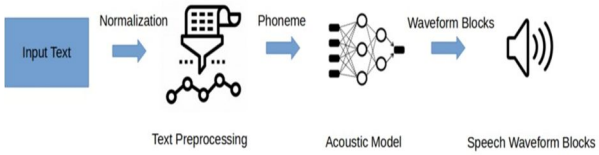


Figure 3: Next generation end-to-end architecture[3]

a) *Training:* In training, we use several training methods: L1 loss and L2 loss in accordance with the predicted element and Sigmoid cross entropy with the weight of the sequence of the stop token sequence. Additionally, we support loss of directed attention, which forces the attentional weights to be diagonal and speeds up the reading of diagonal attention. Credit to PyTorch, it support multi-GPU training efficiently by reducing the training time, especially in the case of Transformer TTS. This is because it requires a larger batch size (e.g.,> 64) to train Transformer steadily. However, this means that Transformer training requires multiple GPUs, which is a bad thing for them light users. To avoid this problem, we support the creation of a flexible pile and gradient collection. In making a flexible collection, the size of the collection is adjusted automatically depending on the length of the input and / or results. By using this program, we can prevent memory loss GPU error created by a very long sentence, thus improved GPU usage. The gradient overlap enables background streaming in several clusters and is reviewed for model parameters as well. This allows us to use the size of a large pseudo-collection and as a result, thus can successfully train Transformer with only one GPU.

b) *Synthesis:* In compilation, we first produce the log Mel filter bank sequence feature using trained E2E-TTS models and then using the Griffin – Lim (GL) algorithm, the WaveNet vocoder (WNV). Comparison between ASR and TTS recipe flowor Parallel WaveGAN (PWG) to generate speech from the sequence of elements. In the case of GL, we change the sequence of the features of the log Mel filter bank into a line spectrogram and use GL in spectrogram. Conversions are done using the opposite Mel base or GRU bank highway network network(CBHG) network. In the case of WNV and PWG, we use the generated Mel filter bank sequence log as the auxiliary network input for production of waveform. We support two types of WNV: one is to use a 16-bit combination of logistics (MOL) and one is that you use 8-bit SoftMax with timeless sound, which can reduce visual acuity in the high frequency band. WNV can greatly enhances the nature of speech produced but in need long production. On the other hand, since PWG is a non autoregressive model, it can produce much faster than real time while maintaining quality compared to WNV.

C. *Speech-to-Text*

Speech recognition is part of the sub-fields of computer science and computer languages that develop methods and technologies that allow the recognition and translation of spoken language into computerized text. This is feature that accurately and quickly transcribes audio into text. Speech Recognition is nowadays regarded as one of the most promising technologies of the future.

ESPnet actually gives bash recipes and python library for speech recognition .Using ESPnet in bash consist of some stages like feature extraction, training and decoding.

Stage 1: Stage 1 consist of data preparation which performs feature extraction, normalization, and text formatting as same as TTS using kaldi. The Feature Extraction is aim to extract the voice features to distinguish different phonemes of a language.

Stage 2: In stage 2 dump dataset of the speech feature and transcription pairs into the datasets in some specific format. A speech database is a collection of recorded speech accessible on a computer and supported with the necessary annotations and transcription. The database collects the observations required for parameter estimations.

Training Stages: Next stage is Training which consists as a backend and built from a large number of different correspondences. Training contains some predefined models like RNN (LSTM, GRU, VGG, etc) with Attention + CTC (dot, location, multi-head, etc),Transformer + CTC, RNN Transducer with Attention and for speech enhancement like joint training beamformer, dereverb (WPE, DNN-WPE), speech separation has done.

The encoder is a deep Convolutional Neural Network (CNN) based on the VGG network.[7] The CTC network sits on top of the encoder and is jointly trained with the attention-based decoder. At the time of beam search process, we integrated the CTC assumption, the attention-based decoder predictions and a individually trained LSTM language model.

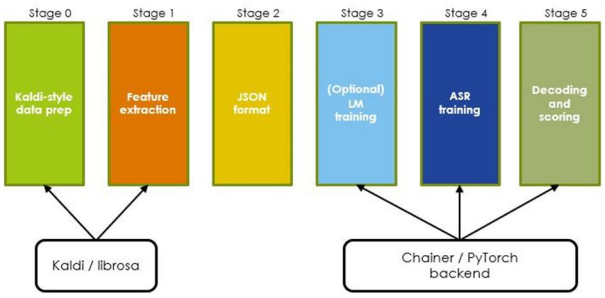


Figure 4 : End-to-End Speech Processes Stages.[13]

Recurrent Neural Networks (RNNs) have been used successfully in many tasks that include sequential data such as machine translation, emotional analysis, photo captions, timeline predictions etc. RNN models developed as Short-Term Memory Networks (LSTMs) enable training.[11]

Attention is an integrated approach to RNN that allows it to focus on specific parts of the input sequence when predicting a specific output sequence, allowing for easy reading and high quality. The combination of monitoring methods
has enabled improved performance across multiple functions making it an integral part of modern RNN networks.

It trained with the embedded and decoder using back propagation, RNN predictive error parameters are broadcast back through decoder then from there to the encoder. This machine enables the decoder to determine which parts of the input sequence to pay attention to. By allowing the decoder have an attention mechanism, we free the encoder from encrypting all the information in the input sequence into one vector.

D. Transliteration

Transliteration is determined by the collection of historical risks, principles, mathematical principles: many language pairs have adopted different rules of translation over time and most translations depend on the origin of the word.[14] These structures make it desirable to look for high-quality, automated learning solutions to the problem.

The goal of text conversion is to represent the source text accurately in the target text, without losing phonetic information. It is useful for reading aloud manuscripts, signposts, etc. It can serve as a useful tool for linguists, NLP researchers, etc. their research is multilingual. Script conversion allows you to read external text accurately in the user's native script. On the other hand, [13] transliteration is intended to conform to the phonology of the target language, while it is closer to the phonetic language of the source language. Transliteration is required for phonetic input systems, multilingual retrieval, answer questions, machine translation and other multilingual applications.

1) MT5 Model

MT5 model introduced with mT5: Text-to-speech translator pre-trained by Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, Colin Raffel [14].

The latest "Text-to-Text Transfer Transformer" (T5) has used the integrated text-to-text format and scales to obtain modern results in various NLP activities in the English language. In this paper, we introduce mT5, a multilingual variation of the pre-trained T5 in a new Common Crawl database covering 101 languages. We describe in detail the design and modified training of mT5 and demonstrate its modern capabilities in multilingual specification. We also explain an easy way to prevent "mistranslation" in the zero shot setting, where the productive model chooses (partially) to translate its prediction into the wrong language. All codes and test models used in this project are publicly available.[15]

Tokenization - It is one of the most ordinary function when it comes to working with text data. Creating tokens to break a sentence, sentence, paragraph, or document into complete text into smaller units, such as individual words or phrases. Each of these episodes is called tokens.

[16]Auto-Tokenization -In most cases, the architecture you want to use can be guessed from the name or method of the pre-trained model you provide in the from pretrained () method. Auto Classes is here to perform the task for you so that you can automatically retrieve the appropriate model given by name / route at pre-trained weights / preparation / vocabulary.

Installing one of the AutoConfig, AutoModel, and AutoTokenizer will automatically create a category of appropriate architectures. For example

model = AutoModel.from_pretrained ('bert-base-cased') will create a BertModel model. There is one AutoModel section for each function, and one backend (PyTorch, TensorFlow, or Flax).

BERT stands for Bidirectional Encoder from Transformers a language representation model one of the model . Unlike the latest models representing the language, BERT is designed to train in-depth two-dimensional presentation of the two-dimensional direction from the labeled text by combining the position both left and right across all layers. As a result, the BERT pre-trained model can be optimized with only a single additional output layer to create high-quality models for a wide range of tasks, such as questioning and language interpretation, without much work and structural adjustments.

BERT is simple in mind and powerful. It gets new modern results in eleven native language processing activities, including pushing the GLUE score to 80.5% (7.7% point total development), MultiNLI accuracy to 86.7% (4.6% total development), SQuAD v1 .1 question to answer Test F1 went to 93.2 (1.5 point for full improvement) and SQuAD v2.0 Test F1 to 83.1 (point 5.1 for full improvement).

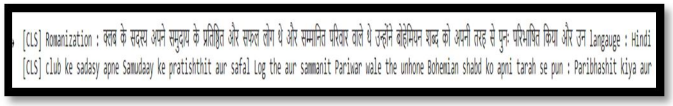


Fig-5Transliteration

2) Data Preparation

With a controlled setting test, 550 set sentences from each language were collected in a variety of ways sources from newsgroups to blogs and more web content. We made sure the sentences were selected includes most of the allowed character combinations in that language as much as possible.

IV. FUTURE WORK

It has scope in emerging voice-controlled technologies, but training algorithm is again very complex. Speech recognition has attracted many scientists and researchers and can be influential to society in emerging technologies.

V. EXPERIMENTAL EVALUATION

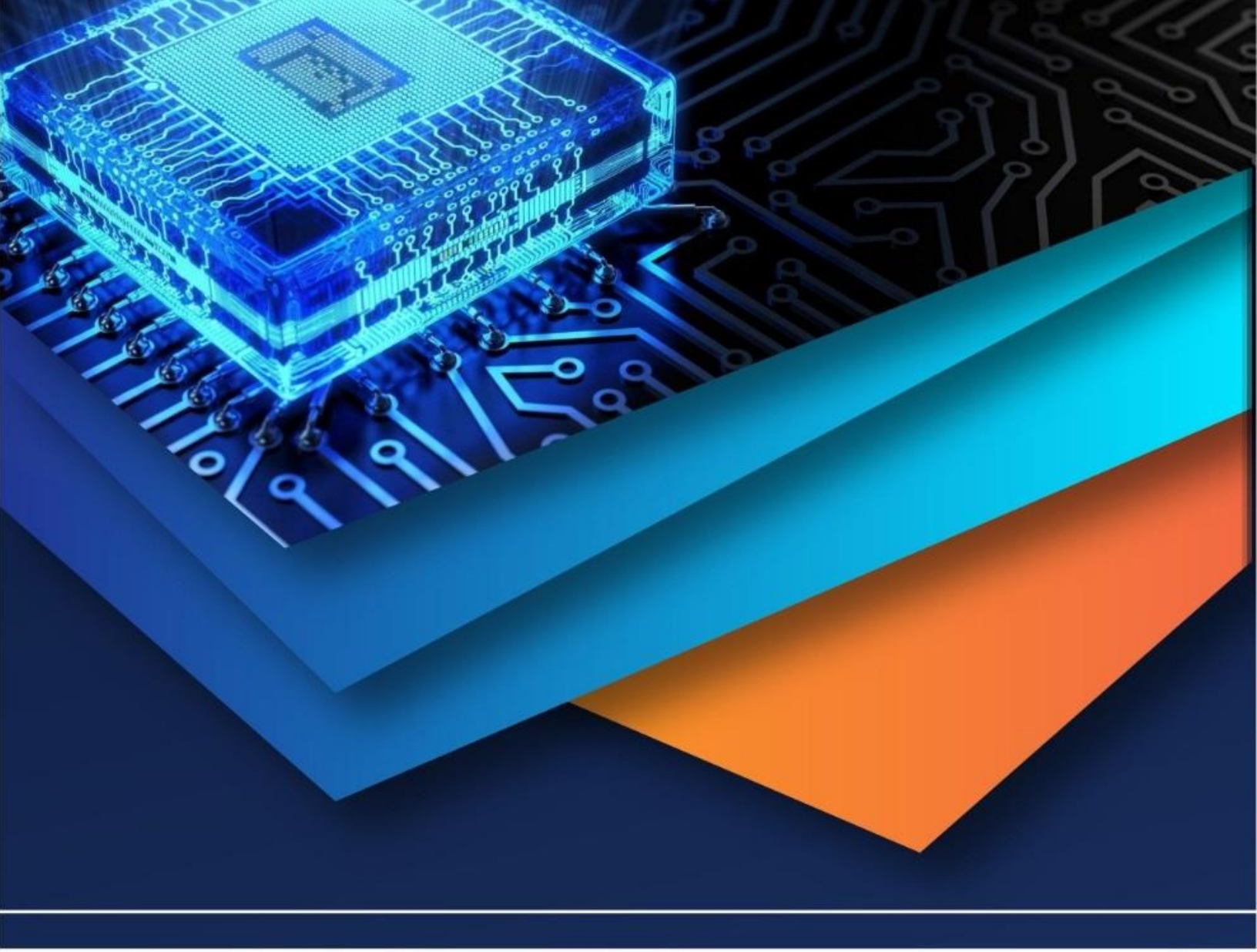
Transformer TTS is slower than Tacotron 2 but FastSpeech is much faster than other models. Basically on the GPU the FastSpeech is 30 times faster than the Tacotron 2 as well as 200 times faster than Transformer TTS ,FastSpeech is more faster. As FastSpeech is not autoregressive model, fully utilized GPU without bottleneck loop processing. Therefore, the level of development is unlocked The GPU is higher than other models.

VI. CONCLUSION

This paper introduces a new E2E-TTS toolkit called ESPnet-TTS as an extension of the ESPnet open source speech processing tool. The tool kit is designed for research purposes to be performed. The E2E-TTS systems are very friendly and speed up this research field. The tool kit not only supports advanced E2E-TTS models however and various TTS recipes whose design is integrated with ASR recipes ,to provide high productivity. Test results we have shown that our models can perform high quality work compared to other recent tools, on the LJSpeech database.

REFERENCES

[1] Chaw Su Thu Thu ,TheingiZin “Implementation of Text to Speech Conversion” Vol.3 Issue, March 2014
[2] AnandArokia Raj, TanujaSarkar , SatishPammi “Text Processing for Text-to-Speech Systems in Indian Languages” 6th ISCA Workshop on speech synthesis, Bonn, Germany, August 2007
[3] NimishaSrivastava ,RudrabhaMukhopadhyay “IndicSpeech:Text-to-Speech Corpus for Indian Languages” ,12th conference on language resources and evaluation, Marseille, May 2020
[4] Mano Ranjith Kumar M, KarthikPandia, Jom and Hema“ Lipsyncing efforts for transcreating lecture videos in Indian Languages”, August 2021
[5] Hardik, Sanket“ Speech recognition using neural networks” Vol 7 Issue 10, Oct2018
[6] Song W., Cai J “End to end deep neural network for automatic speech recognition”2015
[7] S.FuruiMusashinoIEEE Transactions on “ Acoustics, Speech, and Signal Processing” Volume: 34, Issue: 1, Feb 1986
[8] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2017). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.
[9] Yuxan Wang, Stanton, Yang, Xiao “Tacotron: Towards end-to-end speech synthesis” 6 Apr 2017
[10] Heiga Zen, Andrew Senior, Mike Schuster “ Statistical parametric speech synthesis using deep neutral networks”
[11] AdityaAmbekar, Deshmukh, Dave, Parikshit “Speech recognition using Recurrent Neutral Networks” March 2018
[12] Wu, ,Xiu, Shi, Kalini, Koehler “Transformer based Acoustic Modeling for streaming Speech Synthesis”
[13] Tanaka, Ryo, Masumara, Moriya and Aono “Neutral Speech-to-Text Language Models for Rescoring Hypotheses of DNN-HMM hybrid Automatic Speech Recognition System” 15Nov201
[14] Linting Xue ,Noah Constant, Adam Roberts, Mihir Sanjay Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, Colin Raffel “mT5: A massively multilingual pre-trained text-to-text transformer”
[15] Mihaela Rosca and Thomas “sequence to sequence neural network models for transliteration” 29Oct 2016.
[16] PHRathod ,MLDhore ,RMDhore “Hindi and Marathi to English Machine Transliteration using SVM” Vol2, No.4, August 2013.
[17] Tomoki Hayashi1,2 , Ryuichi Yamamoto3 ESPNET-TTS: UNIFIED, REPRODUCIBLE, AND INTEGRATABLE OPEN SOURCE END-TO-END TEXT-TO-SPEECH TOOLKIT.
[18] airc.aist.go.jp/seminardetail/docs/espnet_aist_v2.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY