| **1** |
| --- |
| **Reference in APA format** | Expressive Neural Voice Cloning | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| <https://www.researchgate.net/publication/348958580_Expressive_Neural_Voice_Cloning> | Paarth Neekhara  Shehzeen Hussain,Shlomo Dubnov  Farinaz Koushanfar  Julian McAuley | | | | Transfer learning, speaker verification, multi-speaker text-to-speech synthesis, Deep Voice 3 |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| Expressive Neural Voice Cloning | To synthesize the voice of a speaker with fine control over various style aspects of speech.  It aims to address the limitation of current voice cloning methods that lack the ability to control the expressiveness of synthesized audio. | | | | Speaker Encoding  Style Conditioning  Global Style Tokens (GST)  Multi-Speaker Text-to-Speech (TTS) Model |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| The "Expressive Neural Voice Cloning" solution aims to synthesize the voice of a speaker using only a few reference audio samples.  The steps involved in the process are as follows:   |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Speaker Encoding:  This encoding captures the unique characteristics of the speaker's voice. | Precise Control:  The solution allows for precise control over various style aspects of the synthesized speech, such as tone, speaking rate, and emphasis. | Chance of Overfitting | | **2** | Training the Model:  A multi-speaker Text-to-Speech (TTS) model is trained using the speaker encodings and other conditioning factors. | Few Samples Required:  The system can generate a speaker's voice using only a few reference audio samples, making it applicable in scenarios where limited data is available. | None | | **3** | Cloning Tasks:  The trained model can be used for different voice cloning tasks. These tasks include synthesizing speech directly from text and style control. | Expressive Cloning:  The model can capture and reproduce the unique characteristics of a speaker's voice, enabling expressive voice cloning. | Ethical Concerns:  The technology can be potentially misused for unethical purposes | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | Mel spectrogram  Attention map | Text  Speaker encoding  Pitch contour | None | None | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | | Mel spectogram and attention map are the visualization techniques whose result is affected by the independent variables like text, speaker encoding and Pitch contour. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Text, multi-speaker dataset. | Mel-spectrogram, synthesised audio. | | | Controllable voice cloning  Speech synthesis from text  Precise style control | | | The contribution of this solution is the ability to control the expressiveness of the synthesized audio, including variation in tone, speaking rate, emphasis, and emotions. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| The system can generate speech that is not only accurate but also expressive, capturing variations in tone, speaking rate, emphasis, and emotions.  This capability has several applications, such as voice-overs in animated films, synthesizing realistic and expressive speech for videos. | | | | Misuse and unethical use:  The technology can be abused for creating inappropriate content, spreading misinformation, or generating voice-overs for DeepFake videos. This can have negative consequences for individuals and society. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| This work demonstrates a promising approach to expressive neural voice cloning, but further research and comparisons with existing methods are needed to establish its effectiveness and applicability in different contexts. | | | Style-MOS | | 1.Introduction  2.Related Work  3.Methodology  4.Experiments  5.Results  6.Discussion  7.Conclusion  8.Future Work  9.References |
| **Diagram/Flowchart** | | | | | |
| Figure 1: Expressive voice cloning model: Tacotron-2 TTS model conditioned on speaker and style characteristics derived from the target audio of a given text. At interface time, the model can be provided independent references for style and speaker encodings to achieve expressive voice cloning. | | | | | |

**---End of Paper 1---**

| **2** |
| --- |
| **Reference in APA format** | Neural Voice Cloning with a Few Samples | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| [1802.06006v3.pdf (arxiv.org)](https://arxiv.org/pdf/1802.06006v3.pdf) | Sercan Ö. Arık - [sercanarik@baidu.com](mailto:sercanarik@baidu.com)  Jitong Chen - [chenjitong01@baidu.com](mailto:chenjitong01@baidu.com)  Kainan Peng - [pengkainan@baidu.com](mailto:pengkainan@baidu.com)  Wei Ping - [pingwei01@baidu.com](mailto:pingwei01@baidu.com)  Yanqi Zhou - [yanqiz@baidu.com](mailto:yanqiz@baidu.com) | | | | Voice cloning  Sequence-to-sequence neural speech synthesis systems  Speaker adaptation  Speaker encoding  Voice morphing |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| voice cloning in sequence-to-sequence neural speech synthesis systems | The goal of the voice cloning system is to synthesize a person's voice from only a few audio samples.  The system aims to solve the problem of voice cloning, which involves learning the voice of a speaker from a limited amount of data and generating speech that sounds like it is pronounced by the target speaker. | | | | Speaker adaptation: This approach involves fine-tuning a multi-speaker generative model for a speaker using a few audio-text pairs.  Speaker encoding: In this approach, a separate model called the speaker encoder is trained to directly estimate the speaker embedding from audio samples of an unseen speaker. This model does not require fine-tuning during voice cloning. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| The voice cloning system developed by Baidu Research addresses the challenge of learning speaker characteristics from limited data and generating voice for unseen texts through two approaches: speaker adaptation and speaker encoding.   |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Speaker Adaptation:  This approach involves fine-tuning a pre-trained multi-speaker model using a few samples from an unseen speaker.  By adapting the model to the specific characteristics of the new speaker, the system can generate voice that closely resembles the target speaker. | Good cloning quality  Naturalness  Fast cloning | Limited speaker diversity | | **2** | Speaker Encoding:  This encoding model takes only a few audio samples from the target speaker as input.  This approach offers a more efficient and resource-friendly solution for voice cloning. | Compact representation  Better performance with more cloning audios | Lower naturalness compared to speaker adaptation | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | Evaluation methods | Cloning audios  Training dataset  Model architecture and hyperparameters | Sample Count  Voice cloning approach  Cloning Time | Computational resource requirements | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | | The independent variables I.e, cloning audios, training dataset, model architecture and hyper parameters have an effect on Evaluation methods. This effect is increased or decreased by the moderating variables(sample count, voice cloning approach and cloning time). Mediating variables such as computational resource requirements act as a bridge between the independent and dependent variables. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution in This Work** |
| | **Input** | **Output** | | --- | --- | | cloning audios and the text that needs to be synthesized. | synthesized audio that mimics the voice of the speaker in the cloning audios. | | | Speaker Adaptation: It allows fine-tuning a pre-trained multi-speaker model for an unseen speaker using a few samples.  Automated Evaluation Methods: The solution introduces automated evaluation methods for voice cloning. | | | They show that fine-tuning a pre-trained multi-speaker model with a few samples from an unseen speaker can achieve good cloning quality.  The authors propose automated evaluation methods for voice cloning. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| With this solution, it becomes possible to synthesize a person's voice from just a few audio samples. This can be beneficial in various applications such as voice assistants and audiobooks | | | | Misuse of Cloned Voices: Voice cloning technology can be misused for fraudulent activities such as impersonation, identity theft, or creating fake audio evidence. This can have serious legal and ethical implications. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The paper presents an approach to produce synthesized audio outputs from limited number of samples. It proposes automated evaluation methods. These contributions contribute to the advancement of voice cloning technology. | | | Human evaluations | | Introduction  Related Work  Speaker Cloning Approaches  Automated Evaluation Methods  Voice Morphing  Experimental Setup  Results and Analysis  Conclusion |
| **Diagram/Flowchart** | | | | | |
| None | | | | | |

**--End of Paper 2--**

| **3** |
| --- |
| **Reference in APA format** | Deep Voice: Real-time Neural Text-to-Speech | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| <https://arxiv.org/pdf/1702.07825.pdf> | Sercan O. Arık ¨ † [SERCANARIK@BAIDU.COM](mailto:SERCANARIK@BAIDU.COM)  Mike Chrzanowski† [MIKECHRZANOWSKI@BAIDU.COM](mailto:MIKECHRZANOWSKI@BAIDU.COM)  Adam Coates† [ADAMCOATES@BAIDU.COM](mailto:ADAMCOATES@BAIDU.COM)  Gregory Diamos† [GREGDIAMOS@BAIDU.COM](mailto:GREGDIAMOS@BAIDU.COM)  Andrew Gibiansky† [GIBIANSKYANDREW@BAIDU.COM](mailto:GIBIANSKYANDREW@BAIDU.COM)  Yongguo Kang† [KANGYONGGUO@BAIDU.COM](mailto:KANGYONGGUO@BAIDU.COM)  Xian Li† [LIXIAN05@BAIDU.COM](mailto:LIXIAN05@BAIDU.COM)  John Miller† [MILLERJOHN@BAIDU.COM](mailto:MILLERJOHN@BAIDU.COM)  Andrew Ng† [ANDREWNG@BAIDU.COM](mailto:ANDREWNG@BAIDU.COM)  Jonathan Raiman† [JONATHANRAIMAN@BAIDU.COM](mailto:JONATHANRAIMAN@BAIDU.COM)  Shubho Sengupta† [SSENGUPTA@BAIDU.COM](mailto:SSENGUPTA@BAIDU.COM)  Mohammad Shoeybi† [MOHAMMAD@BAIDU.COM](mailto:MOHAMMAD@BAIDU.COM) | | | | Deep Voice  Real time  Neural Text-to-Speech  Grapheme-to-phoneme conversion  Audio synthesis  Neural networks  WaveNet |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| Deep Voice: Real-time Neural TTS | To develop a text-to-speech (TTS) system that solves the problem of generating high-quality and natural-sounding speech from text in real-time. It focuses on optimizing the different components of a TTS system. | | | | Grapheme-to-phoneme model  Segmentation model  Phoneme duration model  Fundamental frequency model  Audio synthesis model |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Training the Model:  The model is trained using a dataset of short audio clips and corresponding textual transcripts. This step allows the model to learn the mapping between text and speech. | It minimizes the use of hand-engineered features, making it easier to train and reproduce the system. | The synthesized audio quality may be slightly lower than the original audio. | | **2** | WaveNet Inference:  WaveNet is an autoregressive model that generates speech waveform samples one at a time. | Real-time synthesis, making the system usable for various applications. | Inference with WaveNet is computationally intensive. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | Phoneme durations  Audio synthesis quality  Mean Opinion Score (MOS) ratings | Model architecture  Phoneme duration | None | None | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | | The model architecture is expected to have an impact on the quality of audio synthesis. The different components of the model work together to generate the final output, and the quality of each component can affect the overall audio synthesis quality. | | | | | | |
| **Input and Output** | | **Features of this Solution** | | | **Contribution & The Value of This Work** |
| | **Input** | **Output** | | --- | --- | | Dataset of short audio clips  Corresponding textual transcripts | High-quality synthesized audio | | | Standalone System  Real-time Inference  Efficient WaveNet Inference | | | The authors built a fully neural system that can generate speech in real-time, offering adjustable trade-off between synthesis speed and audio quality. They optimized the inference process to achieve faster-than-real-time speeds, making the system usable for various applications. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| It offers real-time inference, which is essential for practical applications of TTS.  It simplifies the process of creating TTS systems and opens up new possibilities for exploration in the field. | | | | The use of AI-generated voices could raise ethical concerns, particularly in cases where the technology is used to manipulate individuals. For example, malicious actors could use the technology to create fake audio recordings for fraudulent purposes. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| One of the key contributions of this work is the development of efficient, real-time WaveNet inference.  While the work demonstrates impressive results, there are still areas for further exploration.  Overall, the work presents a significant advancement in the field of TTS systems, showcasing the potential of neural networks for generating high-quality speech in real-time. | | | MOS (Mean Opinion Score)  Blizzard 2013 Dataset  Performance Model | | 1.Introduction  2.Related Work  3.Model Architecture  a)WaveNet Model  b)Conditioning Network  4.Training  a)Segmentation Results  b)Grapheme-to-Phoneme Results  c)Phoneme Duration and Fundamental Frequency Results  d)Audio Synthesis Results  5.Conclusion  6.References |
| **Diagram/Flowchart** | | | | | |
| Figure 2. system diagram depicting (a) training procedure and (b) inference procedure, with inputs on the left and outputs on the right. | | | | | |

**--End of Paper 3--**

| **4** |
| --- |
| **Reference in APA format** | AN OVERVIEW OF REAL-TIME CHAT APPLICATION | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| <https://www.ijrti.org/papers/IJRTI2206316.pdf> | Akshata D Vhandale, Sayam N Gandhak,Saundarya A Karhale, Sandipkumar R Prasad, Prof. Sudhesh A Bachwani | | | | Chat application  Python  MongoDB  Express JS  Node.js |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| Real-time Chat application | The goal of the current solution is to enable users to communicate with each other seamlessly.  The chat application provides a platform for users to connect and exchange messages in real-time, regardless of their location.  It is designed to facilitate instant messaging and improve communication among users. | | | | Server  Database  User Interface  Direct Messages |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| The MERN stack is a combination of four technologies: MongoDB, Express.js, React, and Node.js.   |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | MongoDB: MongoDB is a NoSQL database used for storing and managing data. It offers flexibility in handling large amounts of data by using a document-oriented structure. | Ability to handle unstructured data efficiently | Integrating MongoDB with other databases can be challenging. | | **2** | Express.js: Express.js is a web application framework for Node.js. It simplifies the process of building the backend of a web application by providing a set of tools and middleware. | Allows for asynchronous programming and follows a single-threaded architecture, which improves performance. | May require additional configuration for complex applications. | | **3** | React: React is a JavaScript library used for building user interfaces. It breaks down the UI into reusable components, making it easier to develop and maintain complex applications. | React offers a virtual DOM, which enhances performance by efficiently updating only the necessary parts of the UI. | It is a bit difficult for beginners to learn. | |  | Node.js: Node.js is a JavaScript runtime environment that allows developers to run JavaScript on the server-side. | Suitable for building scalable and high-performance applications. | Node.js may not be suitable for CPU-intensive tasks. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | None | None | None | None | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | | NA | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| | **Input** | **Output** | | --- | --- | | User details, messages, user actions | User interface of the chat application | | | User-friendly GUI  Real-time messaging  Contextual enrichment  E2EE secured chats | | | The contribution of this work is the development of a web-based chat application using the MERN stack (MongoDB, Express.js, React.js, and Node.js).  This work contributes to the field of web development by showcasing the capabilities of the MERN stack in building real-time chat applications. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| User friendliness  Secure communication  Instant communication  Scalability | | | | The increased usage of chat applications may lead to Social isolation and reduced face-to-face interaction. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The overview emphasizes the need for a real-time chat application with multi-platform support.  The choice of the MERN stack reflects a strategic decision, leveraging MongoDB, Express.js, React, and Node.js for a comprehensive and efficient development process. | | | MongoDB  Express.js  React.js | | The paper is structured as follows:  Introduction  Node.js  Implementation of Real-Time Chat Application  MongoDB  MERN Stack  Performance Optimization  Conclusion |
| **Diagram/Flowchart** | | | | | |
| None | | | | | |

**--End of Paper 4—**

| **5** |
| --- |
| **Reference in APA format** | EmoSpeech: Guiding FastSpeech2 Towards Emotional Text to Speech | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| <https://arxiv.org/pdf/2307.00024.pdf> | Daria Diatlova - [d.dyatlova@vk.team](mailto:d.dyatlova@vk.team)  Vitaly Shutov - [vi.shutov@corp.vk.com](mailto:vi.shutov@corp.vk.com) | | | | EmoSpeech  Speech synthesis  Emotion  Multi-speaker  Lightweight solution |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| EmoSpeech | It aims to solve the problem of synthesizing speech with desired emotions.  The EmoSpeech model extends the FastSpeech2 architecture with modifications that enable conditioning on a given emotion while maintaining fast inference speed. | | | | This step involves extracting phonemes and punctuation from text using the grapheme-to-phoneme (GTP) model. It also includes extracting durations of phonemes and pitch from waveforms. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| The process involves training a speech synthesis model called EmoSpeech using a dataset of emotional speech recordings.  Here are the steps involved:   |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Data Preprocessing:  The dataset of emotional speech recordings is preprocessed to extract features such as mel spectrograms and eGeMAPS features. These features capture the acoustic characteristics of the speech. | Helps in capturing the necessary information for synthesizing emotional speech. | Requires additional computational resources | | **2** | Model Architecture:  The EmoSpeech model is designed with a multi-speaker and multi-emotional setup.  It uses a conditioning discriminator which helps in controlling the emotions of the generated speech. | Helps in controlling the emotions of the generated speech. | Relies on the quality of training data. | | **3** | Training: The EmoSpeech model is trained using reconstruction loss and an adversarial loss. | Reconstruction loss ensures that the generated speech resembles the original input.  Adversarial loss helps in improving the quality and emotional expression of the generated speech. | May require longer training time. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | MOS and NIQSA scores | Emotion-unlabeled dataset  Style control techniques | None | None | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | | MOS score is affected by the quality of the synthesised speech whereas NIQSA score is affected by the naturalness of the speech. | | | | | | |
| **Input and Output** | | **Feature of this Solution** | | | **Contribution & The Value of This Work** |
| | **Input** | **Output** | | --- | --- | | Input text | Synthesised audio with emotions | | | -Extension of FastSpeech2  -Conditioning Mechanism  -Lightweight Solution | | | The contribution of this work is the development of a text-to-speech (TTS) system that can generate expressive speech with controllable emotions.  They also introduce a style control mechanism that allows users to specify the desired emotion in the synthesized speech.  The value of this work lies in its potential applications in various domains such as virtual assistants, audiobooks etc. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| -Improved Naturalness  -Potential applications such as virtual assistants, audiobooks etc. | | | | Overtraining could result in synthesized speech that sounds exaggerated or unnatural. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The work proposes a great approach to enhance emotional speech synthesis.  The experimental results demonstrate its effectiveness, but further research is needed to find the potential for improvement. | | | openSMILE toolkit, CatBoost classifier | | Introduction  Previous approach  Methodology  Experimental Setup  Results and Analysis  Limitations  Conclusion  References |
| **Diagram/Flowchart** | | | | | |
| None | | | | | |

|  |
| --- |
| **6** |
| **Reference in APA format** | Efficient Neural Audio Synthesis | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://arxiv.org/pdf/1802.08435.pdf | Nal Kalchbrenner , Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, Koray Kavukcuoglu | | | | WaveRNN, Weight Pruning, Subscale Dependency Scheme |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| Sparse WaveRNN | The goal of the Sparse WaveRNN solution is to reduce the computational requirements for audio synthesis while maintaining high audio quality. It aims to solve the problem of real-time or faster audio synthesis on GPUs, as well as enabling audio synthesis on low-power mobile CPUs. | | | | Wave Recurrent Neural Networks (WaveRNN)  Sparse WaveRNN  Subscale Dependency Scheme  Batched Sampling |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| The Sparse WaveRNN solution is a method for efficient audio synthesis using neural networks. It addresses the problem of high computational requirements and memory bandwidth limitations in traditional WaveRNN models.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Sparsity: The first step is to introduce sparsity in the model by reducing the number of connections between neurons. This reduces the overall parameter count and computational requirements. | It allows for larger models with better audio quality within the same computational budget | It requires specialized techniques and algorithms to handle sparse matrices efficiently | | **2** | Block-Sparse Matrix-Vector Product: To handle the sparse matrices, the solution uses high-performance block-sparse matrix-vector product operations. These operations optimize the computation and memory access patterns, improving the efficiency of the model. | It reduces the memory bandwidth requirements and speeds up the computation. | Implementing these operations can be complex and requires specialized knowledge. | | **3** | Subscale Dependency Scheme: The solution introduces a subscale dependency scheme to generate multiple samples per step. This allows for batched computation and parallelization, increasing the overall sampling speed. | It improves the throughput and efficiency of the model, especially when using multiple GPU devices. | It requires careful management of dependencies and may introduce a slight sampling lag. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | the audio samples generated by the model. | the sparsity level of the weight matrices. | The sparsity of the weight matrices. | The subscale dependency scheme | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In the Sparse WaveRNN solution for audio synthesis discussed in this article, the dependent variable is the audio samples that are stuff synthesized. The independent variable is the sparsity level of the weight matrices, which refers to the stratum of sparsity or density of the weight matrices used in the model. The moderating variable is moreover the sparsity of weight matrices, which influences the relationship between the independent variable (sparsity level) and the dependent variable (audio samples). The mediating or intervening variable is the subscale dependency scheme, which is a generation process based on subscaling. | |  | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | a sequence of audio samples | generated sequence of audio samples | | | the Sparse WaveRNN solution offers an efficient and effective approach to neural audio synthesis. | | | Overall, the Sparse WaveRNN solution provides a high-quality, efficient, and scalable approach to audio synthesis, making it a valuable contribution to the field. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| the Sparse WaveRNN enables real-time on-device audio synthesis with high quality, making it a significant advancement in the field. | | | | While the Sparse WaveRNN solution offers advantages in terms of efficiency and resource usage, there are potential drawbacks in terms of audio quality, limited applicability, implementation complexity, and compatibility. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| Sparse WaveRNN offers real-time audio synthesis with high quality, but it may have limitations in capturing long-range dependencies and requires specific hardware for optimal performance. | | | Negative Log-Likelihood (NLL) metric and the Mean Opinion Score (MOS) metric. These metrics were used to evaluate the performance and quality of the Sparse WaveRNN model in comparison to other models. | | Abstract   1. Introduction 2. Wave Recurrent Neural Networks 3. WaveRNN Sampling on GPU 4. Sparse WaveRNN 5. Subscale Dependency Scheme 6. Batched Sampling 7. Results 8. Conclusion |
| **Diagram/Flowchart** | | | | | |
|  | | | | | |

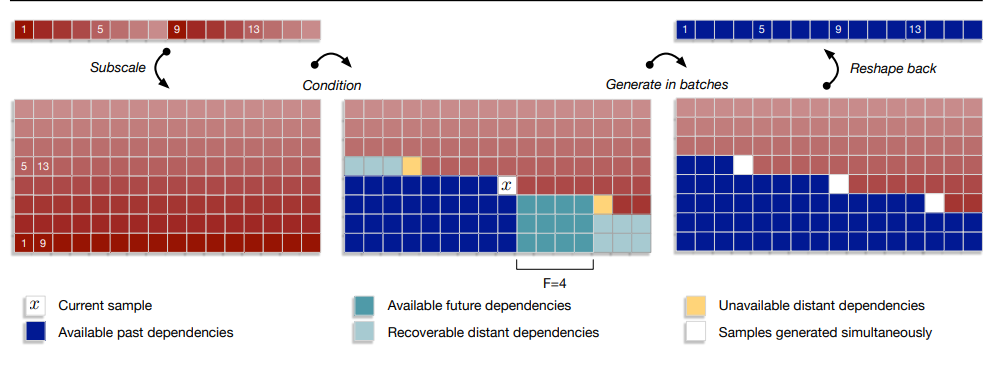
****

Figure 3.working of Sparse WaveRNN

**---End of Paper 6**

|  |
| --- |
| **7** |
| **Reference in APA format** | TACOTRON: TOWARDS END-TO-END SPEECH SYNETHESIS | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://arxiv.org/pdf/1703.10135.pdf | Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. | | | | Tacotron, end-to-end, speech synthesis, text-to-speech, generative model, mean opinion score, parametric system, concatenative system, CBHG module, attention mechanism, encoder, decoder, spectrogram, waveform synthesis. |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| Tacotron: Towards End-to-End Speech Synthesis | The goal of the Tacotron solution is to develop an end-to-end generative text-to-speech (TTS) model that can synthesize speech directly from characters. The problem it aims to solve is the complexity and laborious nature of traditional TTS pipelines, which typically consist of multiple stages and require extensive domain expertise. Tacotron aims to simplify the TTS process by training a single model on <text, audio> pairs, eliminating the need for feature engineering and allowing for rich conditioning on various attributes. | | | | 1. Text analysis frontend: This component extracts various linguistic features from the input text. 2. Acoustic model: The acoustic model predicts the acoustic features of the speech based on the linguistic features extracted by the frontend. 3. Audio synthesis module: This module synthesizes the speech waveform from the predicted acoustic features. 4. Sequence-to-sequence framework: Tacotron uses a sequence-to-sequence model to generate the speech spectrogram directly from the input characters. 5. Waveform synthesis: Tacotron uses a simple waveform synthesis technique to convert the generated spectrogram into a speech waveform. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Encoder: The encoder takes characters as input and converts them into a high-level representation. It uses a sequence of convolutional layers followed by a bidirectional gated recurrent unit (GRU) to capture the contextual information of the input characters. The advantage of the encoder is that it can effectively extract meaningful representations from the character-level inputs. | End-to-end Training: Tacotron can be trained completely from scratch with random initialization, eliminating the need for extensive domain expertise and manual feature engineering. | Training Data Requirements: Tacotron requires a large amount of paired <text, audio> data for training, which may be challenging to obtain in some cases. | | **2** | Attention-based Decoder: The attention-based decoder takes the high-level representation from the encoder and generates the spectrogram frames. It uses an attention mechanism to align the generated frames with the input characters. This allows the model to focus on relevant parts of the input during the synthesis process. The advantage of the attention mechanism is that it enables the model to handle long input sequences and capture the dependencies between characters and spectrogram frames | Faster Generation: Since Tacotron generates speech at the frame level, it is substantially faster than sample-level autoregressive methods like WaveNet. | Alignment Learning: The alignment between characters and spectrogram frames is learned by the model, which can be challenging and may result in imperfect alignments. | | **3** | Post-processing Network: The post-processing network takes the generated spectrogram frames and converts them into waveforms. It applies a series of convolutional layers to refine the spectrogram frames and then uses an inverse Fourier transform to obtain the waveforms. The advantage of the post-processing network is that it improves the quality and naturalness of the synthesized speech. | Robustness: The end-to-end nature of Tacotron makes it more robust compared to multi-stage models, as errors from each component do not compound. | Limited Experimental Results: The experimental results of Tacotron are mainly evaluated on US English, and the performance on other languages or datasets may vary | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | spectrogram frames | the character sequence inputs | CBHG module | The post-processing net | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The encoder, attention-based decoder, post-processing net, and CBHG module work together in the Tacotron model to convert text input into synthesized speech. The encoder extracts sequential representations of the text, the attention-based decoder generates spectrogram frames, the post-processing net converts the frames into waveforms, and the CBHG module helps improve the capability of the model to extract features from the text input. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Characters | spectrogram frames | | | Tacotron offers a simpler and more efficient approach to text-to-speech synthesis, with improved naturalness and faster generation speed. | | | It eliminates the need for multiple stages in traditional TTS systems, such as a text analysis frontend, an acoustic model, and an audio synthesis module. Tacotron can be trained completely from scratch with random initialization, making it easier to build TTS systems without extensive domain expertise. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| It provides an end-to-end generative text-to-speech model that can synthesize speech directly from characters. This eliminates the need for multiple stages and complex components in traditional TTS pipelines, reducing the engineering efforts required to build a new system. | | | | * Computational Resources: Tacotron, like other deep learning models, can be computationally expensive to train and deploy. It may require powerful hardware and infrastructure to achieve real-time performance or to scale for large-scale applications. * Generalization to Other Languages and Accents: The Tacotron model described in the context is trained on US English data. It may not generalize well to other languages or accents, requiring additional training data and modifications to the model. * Training Data Requirements: Tacotron requires a large amount of training data, specifically <text, audio> pairs, to achieve optimal performance. Collecting and curating such datasets can be time-consuming and resource-intensive. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| Tacotron is an integrated end-to-end generative TTS model that shows promising results in terms of naturalness and speed. It eliminates the need for hand-engineered linguistic features and complex components like an HMM aligner. However, there are still areas for improvement, such as the output layer, attention module, loss function, and waveform synthesizer. The authors are working on improving the quality of the waveform synthesis and exploring advancements in learned text normalization. | | | Mean opinion score (MOS) tests. | | Abstract   1. Introduction 2. Related Work 3. Model Architecture 4. Model details 5. Experiment Results 6. Discussions 7. References |
| **Diagram/Flowchart** | | | | | |
| Figure 4. Model architecture. The model takes characters as input and outputs the corresponding raw spectrogram, which is then fed to the Griffin-Lim reconstruction algorithm to synthesize speech. | | | | | |

**--End of Paper 7--**

|  |
| --- |
| **8** |
| **Reference in APA format** | WAVENET: A GENERATIVE MODEL FOR RAW AUDIO | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://arxiv.org/pdf/1609.03499.pdf | Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. | | | | Pulse Code Modulation, language modeling, hidden Markov models, linear prediction, text-to-speech conversion, pitch-adaptive time-frequency smoothing, speech synthesis, WaveNet, multi-speaker speech generation, mean opinion score, statistical parametric speech synthesis, vocoder, deep convolutional nets, fully connected CRFs, acoustic theory of speech production |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| WaveNet | The goal of WaveNet is to generate raw audio waveforms using a deep neural network. It aims to solve the problem of generating high-quality and natural-sounding audio, such as speech and music. WaveNet is designed to be fully probabilistic and autoregressive, meaning that the prediction for each audio sample is conditioned on all previous samples. This allows WaveNet to capture the characteristics of different speakers and generate realistic audio with smooth intonations. Additionally, WaveNet can be used for tasks like text-to-speech and phoneme recognition. | | | | Dilated Causal Convolutions, Autoregressive , Modeling Conditional Modeling , Global Conditioning ,Local Conditioning are the components that allow WaveNet to generate raw audio waveforms with high fidelity and capture the characteristics of different speakers or audio. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| WaveNet is a generative model that operates directly on the raw audio waveform. It models the joint probability of a waveform by factorizing it into conditional probabilities. The model consists of a stack of convolutional layers, with no pooling layers, and the output has the same time dimensionality as the input.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | 1. Dilated Causal Convolutions: 2. WaveNet uses dilated causal convolutions to capture long-range temporal dependencies in audio signals. These convolutions have exponentially growing receptive fields, allowing the model to capture information from a large context window.. | * WaveNet operates directly on the raw audio waveform, allowing it to capture fine-grained details and produce high-quality audio samples. | * Designing the filters for WaveNet can be challenging, as it requires training the model from data to learn the optimal filters. | | **2** | 1. Conditioning: 2. WaveNet can be conditioned on other inputs in a global or local way. Global conditioning involves providing additional inputs, such as speaker identity, to the model. Local conditioning involves transforming a separate time series, such as linguistic features, using transposed convolutional networks or 1x1 convolutions. | * The use of dilated causal convolutions enables WaveNet to model long-range temporal dependencies, which is crucial for generating realistic audio. | * WaveNet can be computationally expensive, especially when using larger context stacks or conditioning on multiple inputs. | | **3** | 1. Context Stacks: 2. To further increase the receptive field size, WaveNet can use separate context stacks that process a long part of the audio signal and locally condition a larger WaveNet. Multiple context stacks with varying lengths and numbers of hidden units can be used. | WaveNet can be conditioned on other inputs, such as speaker identity or linguistic features, allowing for control over the generated audio | * The training process for WaveNet can be time-consuming, as it requires a large amount of audio data to learn the complex patterns in the waveform. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| The major impact factors in this work include the use of WaveNet, a deep generative model for raw audio, for tasks such as multi-speaker speech generation, text-to-speech synthesis, and music audio modeling. The authors also explore the use of conditioning techniques to control the output of the model based on different inputs, such as speaker ID or linguistic features. Additionally, the use of context stacks to increase the receptive field size of the model is another important factor in this work.   |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | categorical distribution | Raw audio waveform | None | None | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | These variables are interconnected in the equations and operations . | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Raw audio waveform | Categorical distribution over the next audio sample, indicating the probability of each possible value for the next sample. | | | Learning the speech front-end with raw waveform Deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes  Postfilters to modify the modulation spectrum for statistical parametric speech synthesis.  Generative image modeling using spatial LSTMs  Speech parameter generation algorithm considering global variance for HMM-based speech synthesis | | | The value of this work is in the advancement of speech synthesis techniques and the potential for generating high-quality, customizable music samples. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| it improves the quality and naturalness of speech synthesis. | | | | Computational complexity, training data requirements, or performance issues. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| These observations highlight the strengths and limitations of WaveNet in speech synthesis, particularly in terms of naturalness, prosody, and speaker modeling. | | | Subjective paired comparison tests and mean opinion score (MOS) tests were conducted.  These evaluations were used to compare the performance of the WaveNet TTS system with baseline statistical parametric and concatenative speech synthesizers. | | Abstract   * Introduction * Background * Methodology * Evaluation * Results * Discussion * Conclusion * References |
| **Diagram/Flowchart** | | | | | |
| Figure 5. Outline of statistical parametric speech synthesis. | | | | | |

**--End of Paper 8--**

|  |
| --- |
| **9** |
| **Reference in APA format** | DEEP VOICE 3: SCALING TEXT-TO-SPEECH WITHCONVOLUTIONAL SEQUENCE LEARNING | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://in.docworkspace.com/d/sILbDueGLAdbWw6oG | Sercan Ö. Arık, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. | | | | Deep Voice 3, neural text-to-speech system, fully-convolutional sequence-to-sequence acoustic model, position-augmented attention mechanism, waveform synthesis, Grifﬁn-Lim spectrogram inversion, WaveNet, WORLD vocoder, multi-speaker speech synthesis, trainable speaker embeddings, text normalization, performance characteristics, MOS evaluations. |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| Deep Voice 3. | The goal of Deep Voice 3 is to develop a high-quality and efficient text-to-speech (TTS) system. It aims to solve the problem of synthesizing natural and intelligible speech from text input. Deep Voice 3 addresses the limitations of previous TTS systems by using an attention-based sequence-to-sequence model, which allows for a more compact architecture and avoids common attention errors. It also focuses on improving training speed and inference efficiency to make TTS feasible for production systems. | | | | Encoder: This component is responsible for converting textual features into an internal learned representation. It uses a fully-convolutional architecture.  Decoder: The decoder decodes the learned representation using a multi-hop convolutional attention mechanism. It generates a low-dimensional audio representation.  Converter: The converter is a post-processing network that predicts the final vocoder parameters from the hidden states of the decoder. Unlike the decoder, the converter is non-causal and can depend on future context information. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| Deep Voice 3 is a fully-convolutional attention-based neural text-to-speech (TTS) system that converts written language into human speech. It consists of three main components: the encoder, the decoder, and the converter.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Encoder: The encoder takes the input text and converts it into an internal learned representation using a fully-convolutional architecture. This allows for parallel computation and faster training compared to recurrent architectures. | Ability to process text quickly and efficiently. | 1. It may not capture long-range dependencies as effectively as recurrent models. | | **2** | Decoder: The decoder uses a multi-hop convolutional attention mechanism to decode the learned representation from the encoder into a low-dimensional audio representation, specifically mel-scale spectrograms. It does this in an autoregressive manner, predicting one timestep at a time | Ability to generate high-quality spectrograms. | 1. It may suffer from attention errors, such as repeated words, mispronunciations, or skipped words. 2. artifacts or distortions in the synthesized speech. | | **3** | Converter: The converter network takes the hidden states from the decoder and predicts the vocoder parameters for waveform synthesis. It is a non-causal network, meaning it can depend on future context information. | Ability to generate the final waveform parameters | It may introduce artifacts or distortions in the synthesized speech. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | mel-scale log magnitude spectrograms, | waveform synthesis method. | Speaker embeddings | Converter | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | mel-scale log magnitude spectrograms are generated by the decoder, the waveform synthesis method is determined by the chosen technique, the speaker embedding allows for multi-speaker synthesis, and the converter network predicts the parameters required for the waveform synthesis method. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | textual features | low-dimensional audio representation in the form of mel-scale spectrograms | | | 1. Encoder,Decoder,Converter, work together to optimize the overall objective function, which is a combination of losses from the decoder and the converter. The model also includes text preprocessing steps to normalize the input text for better performance. | | | Deep Voice 3 contributes to the field of speech synthesis by introducing a more efficient and scalable architecture, demonstrating the effectiveness of attention mechanisms in TTS, and providing high-quality speech synthesis with the ability to handle large-scale deployment. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| Deep Voice 3 contributes to the advancement of TTS technology by addressing common errors, providing flexibility in waveform synthesis, supporting multispeaker synthesis, and delivering superior audio quality. | | | | It's important to note that Deep Voice 3 addresses some of these issues, such as monotonic attention behavior, and offers scalability and faster training compared to previous systems. However, these potential negative impacts should be considered when using Deep Voice 3 or any other TTS system. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The work presents a novel approach to neural TTS and demonstrates promising results in terms of speech quality and scalability. However, it also acknowledges areas for future improvement, such as optimizing inference further and exploring smaller models and fixed-precision arithmetic. | | | CrowdMOS toolkit for Mean Opinion Score (MOS) ratings and the Gentle toolkit for preprocessing and splitting long utterances. | | Abstract   * Introduction * Related Work * Deep Voice 3 Model * Error Modes in Sequence-to-Sequence Speech Synthesis * Waveform Synthesis Methods * Multispeaker Speech Synthesis * Deep Voice 3 System * Experimental Setup * Results * Conclusion * References |
| **Diagram/Flowchart**  Deep Voice 3 uses residual convolutional layers to encode text into per-timestep key and value vectors for an attention-based decoder. The decoder usesthese to predict the mel-scale log magnitude spectrograms that correspond to the output audio. (Light blue dotted arrows depict the autoregressive process during inference.) The hidden states of the decoder are then fed to a converter network to predict the vocoder parameters for waveform synthesis  .  Figure 6. Deep Voice 3 | | | | | |
|  | | | | | |

**--End of Paper 9—**

|  |
| --- |
| **10** |
| **Reference in APA format** | Blockchain-enabled End-to-End Encryption for Instant MessagingApplications | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://in.docworkspace.com/d/sIKjDueGLAcbhw6oG | Raman Singh and Hitesh Tewari. | | | | blockchain-enabled, end-to-end encryption, instant messaging applications. |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| blockchain-based end-to-end encryption (E2EE) framework. | The goal of the proposed blockchain-based end-to-end encryption (E2EE) framework is to provide a real end-to-end encrypted messaging service for instant messaging applications. | | | | Mobile User: The user's device generates a public/private key pair during the application installation phase.  Mobile Network Operator (MNO): The MNO acts as a trusted third-party and issues a digital certificate for the user based on the information provided by them.  Blockchain: The public-key digital certificates are stored on the blockchain. The blockchain maintains the validity of the certificates and allows users to fetch certificates for other users.  Instant Messaging (IM) Server: The IM server verifies the user's certificate from the blockchain network. It does not control the encryption/decryption process and does not store any keys.  Encryption/Decryption: Users encrypt and decrypt messages using their own public/private keys. The sender can fetch the receiver's digital certificate from the blockchain before encrypting a message.  Backup and Restoration: Users generate their backup decryption key using a known secret and store the backup data on their cloud drive. When changing devices, users can download their backups and decrypt them using the backup key. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| The proposed blockchain-based end-to-end encryption (E2EE) framework aims to address the privacy issues associated with current messaging applications by providing a secure and confidential communication environment.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | User Key Generation: During the installation phase of the messaging application, a mobile user creates a public/private key pair on their device. This step ensures that the user has control over their encryption keys. | Users have full control over their encryption keys, enhancing the security and privacy of their messages. | If the user loses their private key, they may lose access to their encrypted messages.  process. | | **2** | Digital Certificate Creation: A trusted third-party, such as a mobile network operator, creates a digital certificate for the user based on the information provided by them. This certificate serves as proof of the user's identity. | The involvement of a trusted third-party adds an additional layer of authentication and verification to the user's identity. | If the trusted third-party is compromised, it could lead to the compromise of the user's digital certificate and potentially their privacy. | | **3** | Storing Certificates on the Blockchain: The user's public-key digital certificate is stored on the blockchain, ensuring its immutability and accessibility. | Storing certificates on the blockchain provides a decentralized and tamper-proof storage solution, enhancing the security and reliability of the certificates. | Storing large amounts of data on the blockchain can be resource-intensive and may impact scalability. | | **4** | Certificate Verification: The messaging application can verify a user's certificate from the blockchain network before establishing a secure communication channel. | Certificate verification from the blockchain ensures the authenticity and integrity of the user's certificate, reducing the risk of impersonation or unauthorized access. | The verification process may introduce additional latency in establishing secure communication channels. | | **5** | End-to-End Encryption: Once the sender fetches the receiver's digital certificate from the blockchain, they can encrypt the message using the receiver's public key. | End-to-end encryption ensures that only the intended recipient can decrypt and read the message, providing strong confidentiality. | If the sender's private key is compromised, an attacker may be able to decrypt the messages. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | |  | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Mobile user's device: The user creates a public/private key pair on their device during the application installation phase.  Digital certificate: A trusted third-party, such as a mobile network operator, creates a digital certificate for the user based on the information provided by them.  Blockchain network: The user sends their public-key digital certificate to the blockchain network. | Encrypted message: The sender fetches the receiver's digital certificate from the blockchain and encrypts the message using the recipient's public key.  Decrypted message: The receiver decrypts the message using their private key. | | | Real End-to-End Encryption: Unlike some messaging apps that claim to provide end-to-end encryption but still involve servers in the encryption/decryption process, this framework ensures true end-to-end encryption. The server does not store any keys or participate in the encryption/decryption process.  Public Key Infrastructure (PKI): The framework utilizes a blockchain-based PKI system to provide secure and scalable digital certificates for users. This eliminates the need for a centralized authority and reduces the cost of issuing and maintaining digital certificates.  Secure Messaging: The framework allows users to securely exchange messages by generating message keys using cryptographic mechanisms. The message keys are derived from chain keys, ensuring secure communication between users.  Backup and Restoration: The framework provides a mechanism for users to backup and restore their data. Unlike traditional messaging apps where backup decryption keys are stored on application servers, this framework advises users to generate their own backup keys, enhancing the security of user data.  Group Messaging: The proposed framework also supports secure group messaging. Group administrators can generate group keys and share them with group members over encrypted channels. Group messages are encrypted using the group key, ensuring confidentiality and authentication. | | | The contribution of this work is the proposal of a blockchain-enabled end-to-end encryption framework for instant messaging applications. The framework aims to provide secure and private communication by leveraging the decentralized nature of blockchain technology. It eliminates the need for a centralized server to store encryption keys and ensures that only the intended recipients can decrypt the messages. The use of blockchain allows for the implementation of a large-scale public key infrastructure (PKI) system at a low cost. The value of this work lies in its potential to enhance the security and privacy of instant messaging applications, offering users a more secure communication experience. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| the blockchain-enabled end-to-end encryption framework has the potential to significantly improve the security and privacy of instant messaging applications, providing users with a more secure and trustworthy communication platform. | | | | Scalability: Blockchain technology is known to have scalability limitations, especially when it comes to handling a large number of transactions and blocks. The proposed framework would need to be tested for its scalability in a blockchain-based environment.  Performance: Blockchain transactions can be slower compared to traditional centralized systems. The encryption and decryption processes may experience delays due to the involvement of blockchain operations, potentially impacting the real-time nature of instant messaging applications.  User Experience: Implementing a blockchain-based framework may introduce additional complexity for users. They may need to understand and manage public/private key pairs, digital certificates, and blockchain operations, which could potentially impact the user experience and adoption of the application.  Governance and Trust: While blockchain technology provides decentralized and transparent mechanisms, the governance and trust aspects of the proposed framework need to be carefully considered. The trustworthiness of the mobile network operator as a trusted third-party and the security of the blockchain network itself are crucial factors for the success of the framework.  Compatibility and Interoperability: Integrating the proposed framework with existing messaging applications and platforms may pose compatibility and interoperability challenges. Ensuring seamless communication between users of different messaging applications could be a complex task. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The work presents a critical analysis of the privacy concerns in online communications and proposes a blockchain-enabled solution to provide real E2EE. It addresses the limitations of the current PKI model and highlights the importance of user privacy in messaging applications. | | | Google Firebase: Used to implement the IM server in the proposed framework.  Ethereum: Implemented on the Docker platform to provide blockchain functionality.  Docker Container: Used as a platform to implement the Ethereum blockchain.  Android Emulators: Used to measure the performance of the AES256 encryption, HMAC calculation, and total encryption time in the Android application. | | Abstarct   * Introduction * Related Work * System Architecture * Phases of the Security Framework * Implementation Details * Evaluation * Conclusion |
| **Diagram/Flowchart** | | | | | |
| Figure 7. Certificate Generation and Registration | | | | | |

**--End of Paper 10—**

|  |
| --- |
| **11** |
| **Reference in APA format** | REAL-TIME VOICE CLONING USING DEEP LEARNING: A CASE STUDY | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://ijcrt.org/papers/IJCRT2305749.pdf | Hruthik B Gowda,Karun Datta Ramakumar,Sheethal.V,Sushma M,Dr. Madhusudhan G K | | | | Voice Cloning, Deep Learning, Text to Speech Synthesis, Dimension Reduction ,Voice Morphing |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| The approach includes three stages: voice cloning, text-to-speech synthesis, and speaker encoding | The aim is to develop a system that can produce speech in any given speaker's voice from given text input data. | | | | Author used TTS Synthesis, WaveNet Vocoder, Neural Network Training to create a voice cloning system that can replicate speech in any given speaker's voice from supplied text input |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Speaker Encoding: In this step, an independent speaker encoder network is used to generate fixed-dimensional embeddings from reference audio of the target speaker. This is used to capture the unique characteristics of the speaker's voice. | The speaker encoder model can generate embeddings with just seconds of reference audio or speech | The quality of the synthesized speech depends on the accuracy of the speaker encoder in capturing the speaker's voice characteristics.  The speaker encoder model may not perform well with very limited or low-quality reference audio. | | **2** | TTS Synthesis: Inspired by Tacotron 2, a sequence-to-sequence TTS synthesis network is used in this step. The fixed-dimensional embeddings from the speaker encoder are used as conditions to generate the synthetic speech. | The TTS synthesis model can produce high-quality synthetic speech that resembles the reference speaker. | The quality of the synthetic speech depends on the accuracy of the speaker encoder in generating appropriate embeddings. | | **3** | Vocoder: The final step is to use a WaveNet powered neural vocoder. This vocoder takes the spectrogram representation from the previous step and generates time-domain waveform samples, which are the final output of the system. | The WaveNet vocoder can generate high-quality and natural-sounding speech waveforms.  Used in converting the spectrograms into speech waveforms. | The vocoder generation process can be computationally expensive and time-consuming.  The quality of the synthesized speech depends on the accuracy of the Mel spectrograms generated by the TTS synthesis model. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The dependent variables in this work are the synthesized voice products | The independent variables in this work are the speaker encoder network, the Tacotron 2 inspired sequence to sequence TTS synthesis network, and the Wave Net powered neural vocoder. These components are combined to create high-quality multispeaker TTS output. | none | none |  |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | | | | | |
| |  | | --- | | The synthesized voice products are dependent variables, influenced by and responsive to changes in the independent variables, which consist of the speaker encoder network, the Tacotron 2 inspired sequence to sequence TTS synthesis network, and the WaveNet-powered neural vocoder, as they collectively contribute to the creation of high-quality TTS output. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Textual data and sample audio | Speaker's voice generated from the supplied text input. | | | The key features of this solution include its ability to produce high-quality text-to-speech (TTS) output | | | This work is mostly lies in advancing the field of TTS by combining Sequence to sequence synthesis network. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| The positive impact of this solution is there is a better scope in advancement of technology on applications like voice assistants. | | | | The major negative aspect of this solution is this model may give unsatisfactory results if the given audio has noise it. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| This work is good, as they tried to develop a voice cloning technique by utilizing Text to speech synthesis for speech generation | | | automated evaluation,Tacotron2 | | Abstract   1. Introduction 2. Literature review 3. Methodology 4. Conclusion 5. Reference |
| **Diagram/Flowchart** | | | | | |
|  | | | | | |

**Figure 8. Text to speech Synthesizer**

**---End of Paper 11-**

|  |
| --- |
| **12** |
| **Reference in APA format** | Text To Speech Conversion Using Different Speech Synthesis | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| file:///C:/Users/yogir/Downloads/Text-To-Speech-Conversion-Using-Different-Speech-Synthesis%20(1).pdf | Fahima Khanam , Farha Akhter Munmun , Nadia Afrin Ritu, Aloke Kumar Saha , | | | | Text to Speech (TTS), Domain specific synthesis, Phoneme based synthesis, Unit selection synthesis. |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| TTS(Text to Speech) Synthesizer for Speech Generation | The goal is to convert written text into natural and intelligible speech by using TTS. | | | | NLP-The NLP includes text analysis, phonetic conversion, and prosodic phrasing. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Natural Language Processing (NLP):  Text Analysis: The input sentence is segmented into tokens, and each word is determined as part of speech (POS) using techniques like segmentation, text normalization, and POS tagging.  Phonetic Conversion: Phonetic transcription is assigned to each word. This can be done using rule-based or dictionary-based approaches.  Prosodic Analysis: This step determines the intonation, amplitude, and duration modelling of speech, which describes the speaker's emotion. | NLP helps in analysing and understanding the input text, enabling accurate phonetic conversion and prosodic analysis. | NLP can be challenging for complex sentences, and errors in text analysis can affect the quality of speech. | | **2** | Digital Signal Processing (DSP):  Speech Synthesis: This part focuses on generating the actual speech waveform. There are different technologies for this:  Concatenative Synthesis: This technology uses a database of pre-recorded natural sounds and concatenates them to produce speech.  Formant Synthesis: This technique does not have any database of speech samples so it sounds artificial and robotic.  Articulatory Synthesis: This technique synthesizes speech based on models of the human vocal tract. It can produce more natural speech but requires complex modelling. | DSP techniques like concatenative synthesis can produce natural-sounding speech | Concatenative synthesis may require a large database and memory capacity, while formant synthesis can sound artificial. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The dependent variable discussed in this article are the output speech quality | The independent variables discussed in this article are:  Text analysis, Phonetic conversion  And Prosodic phrasing | none | none | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The variations in text analysis, phonetic conversion, and prosodic phrasing are expected to influence the output speech quality | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Textual Data | Synthesized speech waveform that corresponds to the input text. | | | One of the features of this solution is the use of unit selection speech synthesis, which selects an optimum set of sounding units from a speech database | | | The contribution of this work is the development of a text-to-speech (TTS) system that can generate natural and intelligible speech for numbers, words, and sentences. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| One positive impact of this solution in the project domain is that the output speech is natural and intelligible, making it easier for blind people to access information through speech. Additionally, the system can be used in various applications such as teaching aids, text reading, and talking books/toys. | | | | One negative impact of this solution in the project domain is that the output speech for words may have discontinuities between transitions of phonemes. This can affect the overall smoothness and naturalness of the speech output. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The work presents an analysis of different synthesis methods for TTS systems and discusses the strengths and weaknesses of each approach. | | | Domain specific synthesis, Phoneme based speech synthesis and Unit selection synthesis | | Abstract   1. Introduction 2. Methodology 3. Implementation 4. Simulation Result 5. Conclusion |
| **Diagram/Flowchart** | | | | | |
| **---End of Paper 12-** | | | | | |
| Figure 9. Flowchart of phoneme based text to speech | | | | | |
|  | | | | | |

**--End of Paper 2--**

|  |
| --- |
| **13** |
| **Reference in APA format** | REAL TIME VOICE CLONING | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://ijrti.org/papers/IJRTI2206107.pdf | Dr.T NAnitha, Amilio Dsouza, Ashutosh , Akshay Gole | | | | Neural Networking, 3-stage pipeline, Text-to-Speech, Deep Learning. |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| The current solution is called Real Time Voice Cloning. | The goal of the "Real Time Voice Cloning" is to develop a three-stage deep learning system that can clone a person's voice in real-time. | | | | Three-level pipeline,Deep learning models,vocoder,reference speech,training data |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
|  | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Feature Extraction: The first step is to extract features from the reference audio, which better represent the sound of the voice produced by the model. | The feature extractor provides data that better represents the voice produced by the model. It helps in capturing the important characteristics of the voice. | The feature extractor might not catch every little detail in the voice, which could mean that some information gets lost along the way. | | **2** | Acoustic Model:  The acoustic model is like a smart tool that figures out how the features we calculate are connected to the actual sounds produced for a given piece of text. | It helps in generating speech that closely resembles the target speaker's voice. | The acoustic model may not be able to perfectly capture all the variations and nuances of the target speaker's voice, leading to some discrepancies in the cloned voice. | | **3** | Vocoder:  The vocoder is a system that reconstructs audio waveforms from the acoustic features generated by the acoustic model. It takes the acoustic features as input and generates the corresponding audio waveform, which closely resembles the voice of the target speaker. | The vocoder is responsible for reconstructing audio waveforms from the acoustic features generated by the acoustic model. It helps in generating high-quality and natural-sounding speech. | The vocoder may introduce some distortions in the reconstructed audio waveforms, affecting the overall quality of the cloned voice. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The dependent variable mentioned in this paper is naturalness. | The Independent variables mentioned in this paper are TTS Pipeline,Training data | none | none | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The study explores the impact of TTS pipeline, and training data as independent variables on the perceived naturalness, the dependent variable, of the synthesized audio. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Reference audio | Cloned voice of reference audio | | | One of the feature of this solution is Personalised voice interface | | | This work's key contribution lies in its development of a real-time voice cloning framework via text-to-speech synthesis, offering the valuable potential to advance voice cloning technology. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| One positive impact of solution in this project domain is that it allows for the replication of unseen voices from just a few seconds of reference speech. This means that with minimal input, the system is able to generate highly natural sounding cloned voices. This can be beneficial in various applications such as personalized voice interfaces. | | | | One negative impact of solution in this project domain is that the cloned voice may lack naturalness and accent. This means that the cloned voice may not sound as natural or have the same accent as the original human voice. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| This work on voice cloning appears to be a promising and innovative approach in the field of text-to-speech synthesis. The authors have identified the limitations of existing models and have proposed a three-stage pipeline to address these limitations. | | | The tools used to assess this work include the SV2TTS database,Google Colab. | | Abstract   1. Introduction 2. Literature survey 3. Proposed system 4. Implementation 5. Result and Conclusion |
| **Diagram/Flowchart** | | | | | |
| Figure 10. System Block Diagram | | | | | |

**--End of Paper 13--**

|  |
| --- |
| **4** |
| **Reference in APA format** | One Model, Many Languages: Meta-learning for Multilingual Text-to-Speech | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://paperswithcode.com/paper/one-model-many-languages-meta-learning-for | Tomáš Nekvinda, Ondˇrej Dušek | | | | text-to-speech, speech synthesis, multilinguality, code-switching, meta-learning, domain-adversarial training |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| The current solution discussed in the article is a multilingual text-to-speech (TTS) model. | The goal of the multilingual text-to-speech (TTS) model discussed in the document is to enable natural-sounding speech synthesis in multiple languages using less training data. | | | | The multilingual text-to-speech (TTS) model discussed in the document consists of the following components : Input Text Encoder,Decoder,Convolutional Encoder,Parameter Generator Network |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Model Architecture: The researchers base their experiments on Tacotron 2, a TTS model. They introduced some changes to the architecture, including the use of convolutional encoders, parameter generation conditioned on language embeddings, and training with multilingual batches. | The advantage of using Tacotron 2 as the base model is that it provides a solid foundation for TTS | The disadvantage is that it may not be optimized for multilingual scenarios. | | **2** | Convolutional Encoders: Instead of having separate encoders for each language, the researchers use multiple language-specific input text encoders as they are used for better Batch Normalization | The advantage of using convolutional encoders is that they can effectively process the input text | The disadvantage is that they may not capture long-term dependencies | | **3** | Encoder Parameter Generation: To enable cross-lingual knowledge sharing, parameters of the encoders are generated using a separate network conditioned on language embeddings. This allows for controllable cross-lingual parameter sharing. | The advantage of generating parameters conditioned on language embeddings is that it enables cross-lingual knowledge sharing. | The disadvantage is that it may limit the ability to capture highly language-specific parameters. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The dependent variables discussed in this article are pronunciation accuracy, voice quality. | Grapheme Encoder Network(GEN) | none | none | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In this article, the dependent variable is the performance of the multilingual text-to-speech (TTS) model, which includes voice fluency and pronunciation accuracy. The independent variables are the different models and approaches used in the experiments, such as the generated (GEN) model .The article explain how these independent variables affect the dependent variable by comparing the performance of the different models and approaches. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Text in specific language | Natural sounding Multilingual Speech | | | One of the features of this solution is the use of multilingual training batches to fully utilize the potential of the architecture. | | | This work contributes to the advancement of multilingual TTS synthesis and provides insights into the effectiveness of different approaches for handling multilingual data. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| This solution enables better knowledge sharing, expands voice cloning capabilities, and facilitates code switching, leading to improved performance and versatility in the project domain. | | | | The major negative aspect in this solution is this model is prone to error generation of speech while working with complex language scripts like Chinese | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The authors highlight the need for cross-lingual knowledge-sharing in multilingual text-to-speech (TTS) systems. They mention that previous work in this area is limited, and they aim to address this gap by proposing a scalable grapheme-based model. | | | The tools used to assess this work include character error rate (CER) evaluation and Google Cloud Platform. | | Abstract   1. Introduction 2. Related Work 3. Model Architecture 4. Dataset 5. Experiments 6. Conclusion and Acknowledgment |
| **Diagram/Flowchart** | | | | | |
| Figure 11. The meta-network generates parameters of language-specific convolutional text encoders. | | | | | |

**--End of Paper 14—**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Version 2.0 Week 2** | | | | | | |
| **15** |
| **Reference in APA format** | Development of chat application | | | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** | | |
| https://www.ijraset.com/research-paper/development-of-chat-application | Dr. Abhay Kasetwar, Ritik Gajbhiye, Gopal Papewar, Rohan Nikhare, Priya Warade | | | | Javascript , React.js, Internet | | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | | |
| To Build a chat Application | The goal of the chat application is to provide a reliable and flexible chat system that allows users to communicate in real-time. | | | | Java script,Internet,Application Registration Page,Message editing field with keyboard | | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Define the Purpose and Goals | This helps to determine the purpose of the app |  | | **2** | Plan the features and functionalities | This Helps to make a list of features and functionalities that the app will have. Prioritize them based on their importance and feasibility. |  | | **3** | Design the User Interface:  Design the user interface (UI) of the app. | A well-designed and intuitive UI enhances the overall user experience, making the app more enjoyable and engaging. |  | | **4** | Backend development:  Set up the server, database, and APIs required for the app's functionality. | Backend development provide overall Functionalities for Smooth operation of app | More Complex and time Taking | | **5** | Front end Development:  Develop the frontend of your app using programming languages like HTML, CSS, and JavaScript. Implement the UI design and integrate it with the backend | Frontend development enhances user experience and engagement by creating visually appealing and interactive interfaces for seamless interaction with the app. | Need to be compatible with both app and the browser | | **6** | Test and debug  Conduct thorough testing to identify and fix any bugs or issues in the app. | Testing and debugging are used for identifying and fixing potential issues early in the development process, ensuring a more reliable and robust application. | Complex and time consuming | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | none | none | none | none | |  |  |  |  | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | N/A | | | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** | |
| |  |  | | --- | --- | | **Input** | **Output** | | User registration details, login credentials, text messages, and search queries | Display of messages, and the transmission of text messages to other users | | | This is a Two way communication system which includes basic app functionalities like Notifications and statistics for unread messages, Saving past messages etc. | | | | The contribution of this work is the development of a real-time messaging app using modern web technologies. Unlike most chat apps available in the market, this app focuses on developers and aims to increase their productivity. | |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | | |
| The positive impact of this chat application in the project domain is that it focuses on developers and aims to increase their productivity. By providing a real-time messaging platform, it allows for easier and faster communication between developers | | | | Development of Chatting application is Complex and Time taking and another limitation is Internet is required in order to use the application | | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** | |
| The work seems to be a well-planned project that addresses the needs of developers and aims to provide a reliable and secure chat system. | | | JavaScript, React.js, MongoDB, Express.js, and Node.js | | | Abstract   1. Introduction 2. Problem Statements 3. Aim 4. Objective 5. Components for App development 6. Conclusion and Future Scope | |
| **Diagram/Flowchart** | | | | | | | |
| Figure 12. Shows the relationship between clients(server and receiver) and the chat services. | | | | | | | |

**--End of Paper 15—**

| **16** |
| --- |
| **Reference in APA format** | Real Time Voice Cloning | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| <https://ijirt.org/Article?manuscript=151003> | Sakith Nalluri, A.Rohan Sai, M.Saraswati | | | | Text-to-speech synthesis, Natural Language Processing, Digital Signal Processing*.* |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| TextToSpeech Robot | The TextToSpeech Robot solution addresses accessibility, aiding visually impaired or reading-disabled users by converting text to speech. It also allows saving converted text as audio files locally. | | | | The Main Application Module manages GUI, basic operations, and parameter input through file, keyboard, or browser. Integrated with it, the Conversion Engine Module utilizes the free TTS API for text-to-speech synthesis. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| The TextToSpeech Robot solution converts text into speech using a text-to-speech (TTS) engine. The process involves two main phases: text analysis and speech waveform generation.   |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Text Input: The user can input text either by typing it into the text field or by copying it from an external document and pasting it into the application. | 1. Accessibility: It provides a useful tool for people with visual impairments, allowing them to convert text into speech and listen to it. | 1. Naturalness of speech: The quality and naturalness of the synthesized speech may vary depending on the TTS API used. Some TTS systems may produce speech that sounds less natural or robotic. | | **2** | Conversion: The input text is converted into speech using the TTS (Text-to-Speech) functionality. The application uses the open-source API called c# for this conversion. | Ease of use: The application has a simple and user-friendly interface, making it easy for users to input text and convert it into speech. | Limitations in handling punctuation and abbreviations: The conversion process may not accurately handle punctuation and abbreviations, leading to potential errors or unnatural speech output | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | |  |  |  |  | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | |  | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| | **Input** | **Output** | | --- | --- | | The input of the TextToSpeech Robot solution is text, which can be entered directly into the application's text field or copied from an external document. | The output of the solution is speech, where the text is converted into synthesized speech that can be heard by the user. | | | Our software is called the TextToSpeech Robot, a simple application with the text to speech functionality.The main features of the TextToSpeech Robot solution include:   * 1. Text-to-speech functionality   2. User-friendly interface   3. Input options   4. Saving audio files | | | The TextToSpeech Robot solution aims to improve the similarity and naturalness of generated speech, and it can be further enhanced and expanded upon in future developments.  The solution utilizes deep learning networks and advanced audio processing techniques to generate natural-sounding speech |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| This can be beneficial for individuals with visual impairments as it allows them to access and consume large volumes of text more easily | | | | The generated speech may not sound completely natural or human-like, which can affect the user experience and make it less engaging or enjoyable. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| It mainly discusses the development of a framework for real-time voice cloning using deep learning networks and text-to-speech synthesis. It mentions the successful implementation of the framework and the potential for further improvement | | | NONE | | Abstract   1. INTRODUCTION 2. LITERATURER SURVEY 3. STRUCTURE OF A TEXT-TO-SPEECH 4. PROPOSED WORK 5. METHODOLOGY 6. CONCLUSION |
| **Diagram/Flowchart** | | | | | |
|  | | | | | |

**---End of Paper 16---**

| **17** |
| --- |
| **Reference in APA format** | VOICE CLONING: A MULTI-SPEAKER TEXT-TO-SPEECH SYNTHESIS APPROACH BASED ON TRANSFER LEARNING | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | Keywords in this Reference |
| https://arxiv.org/pdf/2102.05630.pdf | Giuseppe Ruggiero, Enrico Zovato, Luigi Di Caro, Vincent Pollet | | | | End-to-end text-dependent speaker verification, deep neural networks, 1D convolution neural networks, gated recurrent neural networks, neural audio synthesis. |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| Advanced Gru Network. | The goal of the "Advanced Gru Network" solution is to build a Text-to-Speech (TTS) system that can generate natural speech for a wide variety of speakers. | | | | * 1 Conv1D layer * 3 GRU layers * Linear projection layers |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Input Processing: The input texts are converted into phoneme sequences, and the target mel spectrogram features are computed on 50 ms signal windows, shifted by 12.5 ms, and passed through an 80-channel mel-scale filterbank. | Faster Training: Compared to the GRU network, the "Advanced Gru Network" is faster during training, which can save computational time. | The linear projection layer may introduce additional computational complexity to the model and increase the number of parameters that need to be trained. | | **2** | Model Architecture: The Advanced Gru Network consists of 1 Conv1D layer and 3 GRU layers, each followed by a linear projection layer. The Conv1D layer performs convolutional operations on the input features, and the GRU layers capture the temporal dependencies in the data. | Improved Speaker Verification: The "Advanced Gru Network" achieved the best Speaker Verification Equal Error Rate (SV-EER) on the test set, indicating its effectiveness in accurately verifying speakers. | The main disadvantage of the Conv1D layer is that it may not capture long-range dependencies effectively, as it operates on local regions of the input. | | **3** | GRU Layers: The network then includes three GRU layers, which are recurrent neural network layers that can capture long-term dependencies in sequential data. Each GRU layer is followed by a linear projection layer. |  | GRU layers may struggle with capturing very long-term dependencies in the data. They may also suffer from the vanishing gradient problem, which can affect the training process. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | synthesized speech waveform. | mel spectrogram feature vector | None | None | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | | The dependent variable is the synthesized speech waveform, this means that in study the synthesized speech waveform is the output or result that you are specifically interested in analyzing, measuring, or evaluating.  The independent variable is the mel spectrogram feature vector, this means that the mel spectrogram feature vector is the input data or parameter that you are manipulating or varying in order to observe its effect on the synthesized speech waveform, which is the dependent variable, commonly used in speech processing tasks.  The moderating variable is the speaker encoder, which computes an embedding vector that characterizes the voice of the speaker and conditions the synthesis process.  The mediating (intervening) variable is the synthesizer, which takes the input text and the embedding vector and predicts the mel spectrogram. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution in This Work** |
| | **Input** | **Output** | | --- | --- | | The input of the "Advanced Gru Network" model is a sequence of mel spectrogram frames | the output is a fixed-dimensional embedding vector that represents the speaker characteristics in the transformed space. | | | The "Advanced GRU Network" solution is a speaker encoder model that combines the advantages of convolution and GRU networks. It consists of 1 Conv1D layer and 3 GRU layers, each followed by a linear projection layer. It achieved the best Speaker Verification Equal Error Rate (SV-EER) on the test set compared to other models. | | | This model was found to be faster during training compared to the GRU network and achieved the best Speaker Verification Equal Error Rate (SV-EER) on the test set. The advanced GRU network was able to create a robust space of internal features that effectively separated speakers based on their utterances. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| The goal of this work is to build a TTS system which can generate in a data efficient manner natural speech for a wide variety of speakers, not necessarily seen during the training phase. | | | | Lack of human-level naturalness.  Inability to reproduce speaker prosody.  Data limitations | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The authors have provided detailed information about the methodology, including the use of deep neural networks and attention mechanisms | | | UMAP  Mean Opinion Score (MOS) | | * Introduction * Model Architecture * Experiments and Results * Conclusions * References |
| **Diagram/Flowchart** | | | | | |
| Figure 13. Voice cloning: A Multi-Speaker Text to Speech Synthesis approach based on transfer learning. | | | | | |

**--End of Paper 17--**

| **18** |
| --- |
| **Reference in APA format** | Text to Speech Conversion with Emotion Detection | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://www.ripublication.com/ijaer18/ijaerv13n14\_23.pdf | Anita , Srinivasan. | | | | Overall, the techniques used involve analyzing sentence patterns, assigning emotion constants to words, and matching words with audio files in the multimedia database based on their assigned emotions. |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| data collection, pattern recognition, output audio generation and validation. | To create a text-to-speech conversion, analyze the emotions present in textual data. | | | | Emotion Detection, Grammar Identification, Text-to-Speech Conversion, Pattern Recognition. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Emotion Detection, Grammar Identification, Text-to-Speech Conversion, Audio Arrangement | Helps in understanding the emotional context of the textual data.  Ensures proper sentence structure and improves the quality of the speech output. | May not accurately detect complex emotions or subtle nuances.  May not handle all grammar rules or variations in sentence structure. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | The emotion associated with a word or sentence. | patterns of emotion, the strength of the words in the sentence, the position of the words. | The process of emotion detection from text and converting it into speech. | None | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | | The article discusses the relationship among the variables of emotion categories, sentence pattern, emotion strength, text analysis, machine learning, and the accuracy and performance of predictive models. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| | **Input** | **Output** | | --- | --- | | Takes input in the form of a sentence with an emotional word | output is a synthesized speech that conveys the emotional content of the input sentence | | | Accurate and Reliable Data, User-Friendly Interface, Handling Complex Sentences and Emotions. | | | This work efficiently handles complex sentences, saving time in text-to-speech conversion, providing accurate information with emotional detection. It's valuable for the information age's accuracy expectations. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| Improved data understanding, user interaction, robotics applications, accuracy, time efficiency, and future enhancements. | | | | Limits in emotional range, database dependency, context understanding, and training | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| This paper explores emotion detection in text, converting it to speech for human-like machine interaction. Utilizing complex algorithms, the system achieves 100% accuracy with ample training data, offering potential enhancements for emotional voice outputs in robotics, promising more engaging machines. | | | Vector Space Model (VSM), Naïve Bayes classifier, Support Vector Machine (SVM), unsupervised machine learning approach. | | Abstract   1. Introduction 2. Literature Review 3. Methodology 4. Results and Discussion 5. Conclusion 6. References |
| **Diagram/Flowchart** | | | | | |
| Figure 14. Text to Speech Conversion with Emotion Detection | | | | | |

**--End of Paper 18--**

| **19** |
| --- |
| **Reference in APA format** |  | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://www.researchgate.net/publication/322509087\_Developing\_an\_End-to-End\_Secure\_Chat\_Application | Noor Sabah  Jamal M. Kadhim  Ban N. Dhannoon | | | | Developing an End-to-End Secure Chat Application,  Two-step verification, MongoDB ..etc |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** |
| Chat application that aims to preserve the security and privacy of the chat communication | The goal is to solve the problem of security and privacy concerns, To provide end-to-end security for users to safely exchange private information without worrying about data leakage. | | | | Client-side  Server-side |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| * Registration * Login * Message Encryption | | | | | |
| |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Users need to register an account by providing their name, email, and password. | Provides a secure container to store the local storage key, making it difficult for unauthorized access. | The encryption algorithm used for password encryption may have vulnerabilities that could be exploited by attackers. | | **2** | Users authenticate themselves by providing their email and password. The password is encrypted and sent to the server for validation. | Ensures user authentication and generates a JWT for secure communication. | If the JWT is compromised, an attacker could impersonate the user and gain unauthorized access. | | **3** | Each message has its own separate key and nonce for better security. | Ensures the confidentiality and integrity of messages by encrypting them and computing a MAC. | The encryption algorithm used may have vulnerabilities that could be exploited by attackers. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| * Security and Privacy * Encryption Algorithms * Keystore and Local Storage * Client-Server Architecture  | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | * Security and Privacy of the Chat Application * Message Encryption * End-to-end Security | * Registration Information * Encryption Algorithm * Authentication Information | * End-to-End Security * Secure Storage | * MongoDB * Node.js | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | | In this variable there is a strong relationship among security and privacy measures, storage protection, speed and performance, and user interface and user experience. | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| | **Input** | **Output** | | --- | --- | | Text messages, images, or files that they want to send to another user. | encrypts the input messages to ensure security and privacy. | | | Secure chat: End-to-End Encryption, secure authentication, local storage, TLS, no server storage, two-factor authentication. | | | The contribution of this work is the development of an end-to-end secure chat application, ensures the security and privacy of user communications by implementing various encryption algorithms and secure storage mechanisms. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| It ensures that messages exchanged between users are encrypted and can only be read by the sender and receiver, without the involvement of any third party. | | | | Compatibility issues, open-source evaluation absence, limited features, user adoption challenges, performance issues, security and privacy concerns. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| The importance of security and privacy in chat applications that addresses these concerns by implementing modern methods and lightweight algorithms and providing confidence to clients that their messages are protected even if their mobile phones are compromised. | | | None | | Abstract   1. Introduction 2. Mobile Chat Applications 3. Proposed architecture 4. Analysis the Proposed Chat 5. Conclusion |
| **Diagram/Flowchart** | | | | | |
| Figure 15. providing security using end to end encryption. | | | | | |

**--End of Paper 19—**

| **20** |
| --- |
| **Reference in APA format** | Voice Generation Using Deep Learning | | | | |
| **URL of the Reference** | **Authors Names and Emails** | | | | **Keywords in this Reference** |
| https://www.academia.edu/91460398/Voice\_generation\_using\_deep\_learning | Gonzalo Gómez Sánchez | | | | * Text-to-Speech Systems, Accessibility Tools, Multimedia Production, Language Learning and Education. |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | **The Goal (Objective) of this Solution & What is the to problem that need be solved** | | | | **What are the components of it?** |
| Deep Neural Networks | The solution is to develop a system for voice generation using deep learning | | | | deep neural networks (DNNs), LSTM. |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | |
| * **Acoustic Feature Extraction** * **DNN Training** * **Audio Waveform Generation** | | | | | |
| |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | --- | --- | --- | --- | | **1** | Acoustic features, extracted from speech waveform and text transcriptions, serve as DNN input capturing speech characteristics. | DNN system enhances voice with natural, intelligible quality surpassing tradition | Training and optimizing deep neural networks can be computationally expensive, requiring significant computational resources. | | **2** | DNN training optimizes network to learn mapping between acoustic features and audio waveform | DNNs optimize synthesis system, eliminating separate steps like feature extraction for efficiency | System produces intelligible audio, but voice quality may lag behind state-of-the-art | | **3** | If the DNN system is trained, it can be used to generate the audio waveform directly from text input. | Success could yield a text-to-voice Deep Learning system, bypassing pre-recorded data. | Research needed to enhance proposed system's performance and quality. | | | | | | |
| **Major Impact Factors in this Work** | | | | | |
| | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | --- | --- | --- | --- | | Audio Waveforms | Acoustic Features | None | None | | Intelligibility of Audio | Datasets | None | None | | Text-to-Speech System | Pseudo-Quadrature Mirror Filters | None | None | | | | | | |
| | **Relationship Among The Above 4 Variables in This article** | | --- | |  | | | | | | |
| **Input and Output** | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| | **Input** | **Output** | | --- | --- | | Text generates audio waveform parameters. | System outputs audio waveform via parameter-based vocoder. | | | Use of Deep Neural Networks (DNN), Investigation of different architectures, Generation of audio waveform | | | The thesis contributes to speech synthesis through deep learning, proposing models for intelligible audio, introducing Pseudo-Quadrature Mirror Filter banks for efficiency, and emphasizing computational cost and audio quality considerations for future advancements. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | |
| The audio signal in speech synthesis can lead to improved quality, simplified system architecture, parallelization, and the potential for more advanced text-to-speech systems. | | | | Proposed solution faces high computational cost, audio distortion, limiting scalability. | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| It explores deep learning for voice generation, proposing architectures with limitations. Parallelization reduces computational costs, suggesting avenues for future improvements in audio quality and alternative architectures. | | | PQMF architecture, Deep Convolutional Neural Network (CNN) | | Abstract   * Introduction * State of the art * Background * Data Preparation * Proposed models for Voice Generation * Conclusions |
| **Diagram/Flowchart** | | | | | |
| Figure 16.Voice Generation Using Deep Learning | | | | | |

**--End of Paper 20--**