SYNTHETIC VOICE CREATION – A CONCEPTUAL MODEL

PROJECT REPORT

MARCH 28, 2021

MADE BY: **SOVAN SINHA ROY**

TABLE OF CONTENTS

1. INTRODUCTION

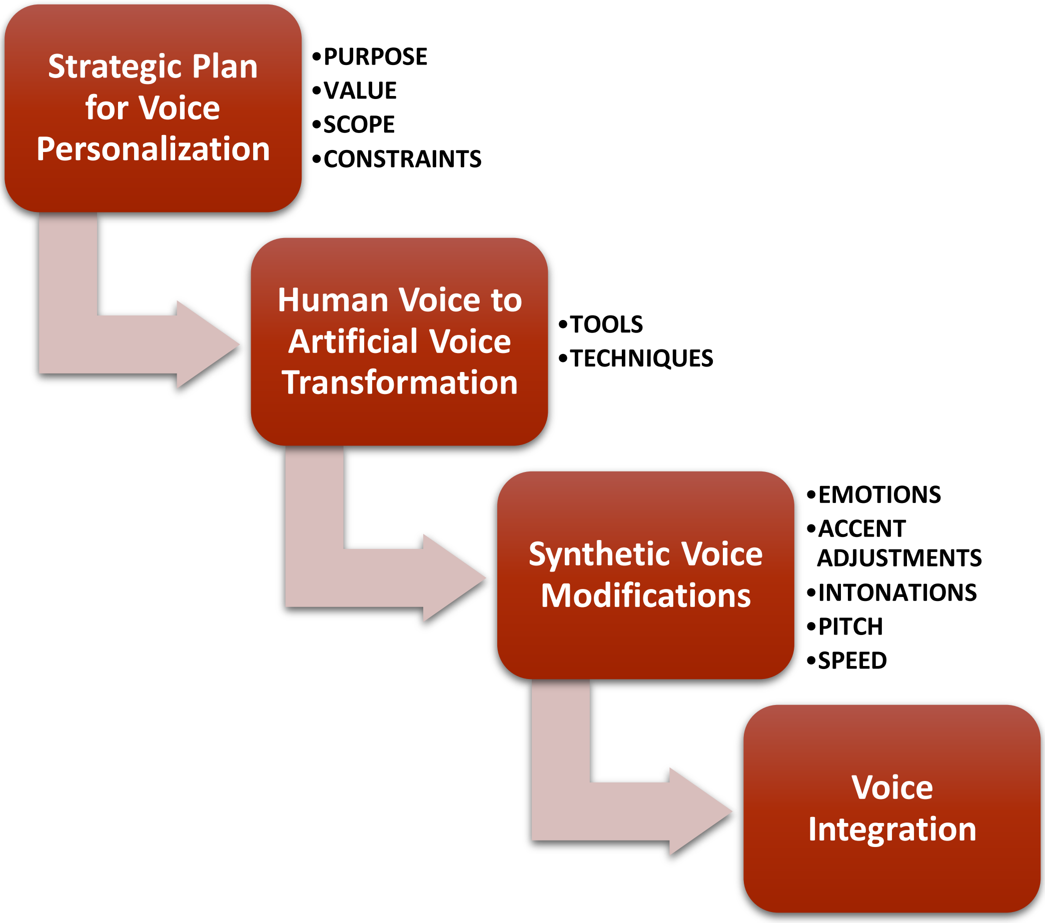
This project report titled **"Synthetic Voice Creation – A Conceptual Model"** serves as a seminal work for the practical implementation of a potential project in the field of Machine Learning (ML) & Artificial Intelligence (AI) based on the theme of Synthetic Voice Creation for Movies, TV-Series, Video Games, etc. In this endeavour, we strive to conceive an environment where the characters are all played by virtual entities, or in other words, we conceive movies & TV shows taking place without real actors rather artificial characters mimicking some real-world entities. So, this report solely emphasizes on Artificial Voice Creation whereby the dialogues and speeches generated or produced during the movies or TV series are actually done synthetically by the trained virtual actors that are inspired by some real-life characters whose voices they are replicating for the programme content. So, in this report, we will describe the needs, requirements, and the proposed working mechanism for this entire initiative.

2. PROJECT DESCRIPTION

2.1 Significance of the Synthetic Voice Creation

The prime significance of this technology is currently for providing voice assistance and facilitating prompt responses to voice commands (two-way interactions with the humans). At present, the text-to-speech (TXT) synthesis is what is used extensively with Voice-to-Voice (V2V) used sparingly for various purposes in this context but since it (Synthetic Voice Creation) relies heavily on labelled data and human expertise, and also requires statistical models & specific rules, it is only limited to generic products at the moment with little to no tailoring to any application realm. However, it is a rapidly evolving domain of IT that is improving leaps & bounds steadily whilst gaining more traction and thus it promises to grow & have a very bright future ahead. Hence, soon we can envisage a world wherein synthetic voices are widely used for movies, TV series, video games, etc rendering immaculate voices akin to real humans.

2.2 Project Block Diagram

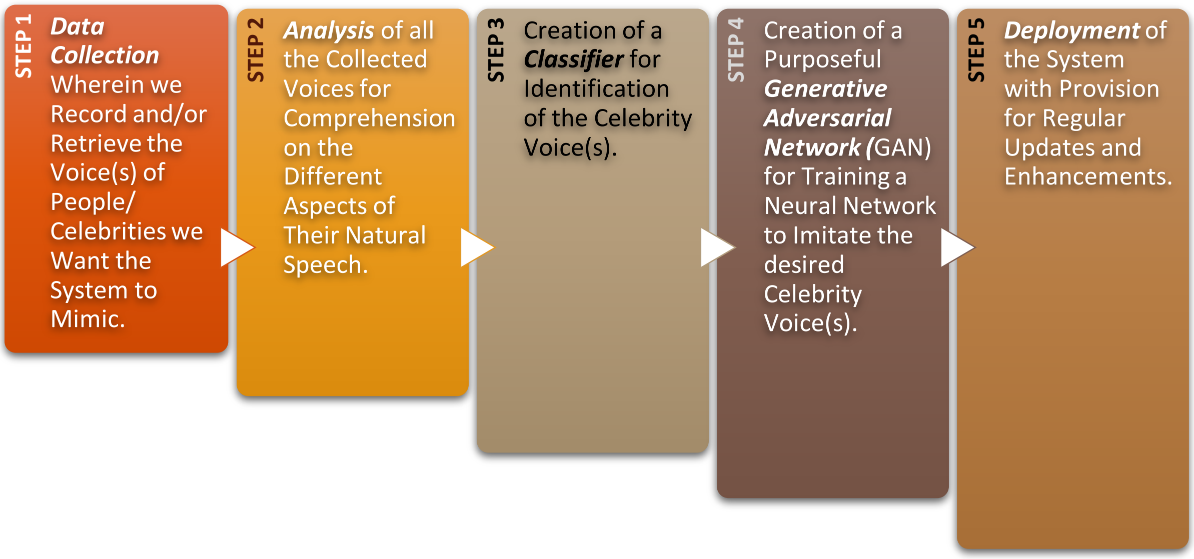


On a very high level, the flowchart above depicts the four procedural stages of this entire synthetic voice creation project. They are:

- Initially, we need to have a thorough voice personalisation strategic plan whereby we establish the main purpose which in this case is to produce content from artificial characters emulating the real-world entities that they studied during the training period. Moreover, the scope and the constraints in the overall project are also ascertained during this phase.
- Secondly, we use AI & ML tools & technologies like TensorFlow, Generative Adversarial Network (GAN), etc to transform the human voice into an artificial voice that is indistinguishable. What it does is that it facilitates an output in a particular voice that is familiar to us, but the contents to be delivered are actually created by us for the artificial voice (for it to ultimately narrate or produce).
- The synthetic voice is then modified and amended to resemble a real human voice with emotions. Apart from emotions, the accent aspect is taken care of to ensure that it is proper, clear, and comprehensible. Moreover, the rise and fall of speech aspect (intonation) is also looked into along with the pitch. Furthermore, the speed of speech delivery can be adjusted as & when required for a smooth natural flow.

- Finally, when everything (mentioned above in the first three phases) is incorporated and rigorously tested, we integrate the voice which can now replicate real-world entities across various platforms like movies, TV series, video games, etc.

2.3 The Working Mechanism of the Model



STEP 1 ***Data Collection*** Wherein we Record and/or Retrieve the Voice(s) of People/ Celebrities we Want the System to Mimic.

STEP 2 ***Analysis*** of all the Collected Voices for Comprehension on the Different Aspects of Their Natural Speech.

STEP 3 Creation of a ***Classifier*** for Identification of the Celebrity Voice(s).

STEP 4 Creation of a Purposeful ***Generative Adversarial Network*** (GAN) for Training a Neural Network to Imitate the desired Celebrity Voice(s).

STEP 5 ***Deployment*** of the System with Provision for Regular Updates and Enhancements.

The systematic working mechanism of this conceptual model can be broadly classified into five different phases: *Data Collection, Data Analysis, Classifier Creation, GAN Creation, and Deployment*. Let us explain each one of them briefly.

2.3.1 Data Collection

The initial phase consists of recording voices of the desired personalities. They can be ordinary people or perhaps celebrities. We want the system to identify whether the voice belongs to the celebrities and then interchange the voices and ultimately produce voice outputs resembling those voices learned.

2.3.2 Data Analysis

The subsequent phase of data analysis is where we analyse all the collected voice samples to ensure that we have different varieties of voices which can be processed to render dynamic artificial voices. Moreover, here we also evaluate the different aspects of the natural voices to ensure that we have high-quality and richness of input prior to synthesis.

2.3.3 Classifier Creation

This is the training phase wherein we use back-propagation to train the neural network to identify the different celebrity voices and develop the capabilities to replicate them. Superimposition of several ordinary voices can be undertaken to match the celebrity voice which is inherently done in the next phase but here we lay the foundation for it through rigorous training.

2.3.4 GAN Creation

Now that the training is complete, it is time to synthesize the voice. For which, we need to create a specific purpose-oriented Generative Adversarial Network or GAN. The prime motive here is to imitate the celebrity or real-life entity's voice now which was studied earlier in the previous phase during training. The end result should yield an output resembling the celebrity voice.

2.3.5 Deployment

Ultimately, it is time to deploy the system which is done in this final stage. We must ensure that there is room for the accommodation of further improvements in the near future. Moreover, provision for regular updates, bug fixes, etc should also be there which ought to be undertaken in a timely manner during the maintenance of the overall system.

3. NECESSARY TECHNICAL CONSIDERATIONS

The project should be created using *Agile Software Development Life Cycle* (SDLC) methodology requiring 2-3 months and an estimated budget of roughly $20,000 (AUD) including all the fixed & variable costs (subject to changes). A small Information Technology (IT) team consisting of 5-6 members is needed as the skills & expertise to internally *design, build, test,* and *implement* incrementally is vehemently required in this effort with the need for the self-management of the project to be undertaken throughout. Therefore, the recommendation is to get on board:

- **2 Developers** (one Front-End and the other Back-End or either or both Full-Stack with at least 2 years of experience in AI & ML projects and all the tools that are required for them like ***TensorFlow, Python, Data Science Algorithms, R Programming, Apache Spark***, etc.)
- **1 Tester/Quality Assurance** person well-versed with both comprehensive manual and/or automated testing with more than a years' experience with similar projects requiring ***APIs, Testing Frameworks (Verification & Validation)***, and software applications like ***Selenium***, etc.
- **1 Business Analyst** for continuous overall project analysis and management throughout the project with 1-2 years of experience of guiding similar projects in the past with strong ***Documentation, Stakeholder Management, Business Process Mapping, Requirement Analysis Skills***.
- **1 Data Analyst** for rigorous Data Analysis with a thorough understanding of ***Data Cleansing, Exploration, Visualization,*** etc. At least a years' experience along with proficiency in ***SQL, Python, Tableau, Excel,*** et al are highly desirable.
- **1 Security/Network Specialist** (Optional) for overseeing the security aspects of the project to ensure malware and unwarranted elements are kept at bay. Decent knowledge about ***Computer Networking, Data Communication, Cryptography***, et al is needed for this.

Tools & Technologies Required:

- *Deep Learning*
- *TensorFlow*
- *Voice Synthesizers (TTS like Google TTS & Amazon Polly or VTV applications)*
- *Generative Adversarial Networks (GAN)*
- *Digital Voice Processing (LyreBird)*
- *Neural Voice Cloning*

4. PRE-EXISTING TECHNOLOGIES

Currently, we do not have a fully-fledged voice synthesizer with all the comprehensive *features* and *nuances* for artificial voice synthesis that is akin to a complex or advanced human being, but we do have a few software applications that do clone the human voice to a great extent rendering realistic output**.** These services leverage the ***neural network, deep learning algorithms***, and ***artificial intelligence*** in order to generate and clone the human voice. Some noteworthy existing technologies in this regard are as follows:

1) **Lyrebird AI**: Lyrebird can be seen as the pioneer of this technology at its infancy in the contemporary world. They have a very *powerful* and *popular* AI-oriented tool which also harnesses *deep learning* and *machine learning algorithms* that are widely used to produce realistic *artificial voice messages, deepfake videos*, etc. Primarily created for *entertainment* and *creative expression*, this product is now used for many more purposes and such is its craze that it was acquired by a popular audio-editing company **Descript** in 2019.
2) **Overdub**: Another product from **Descript**, Overdub is a futuristic application still in its development stage with a beta version available at the moment for end-user testing. The tool promises to be a massive asset to the *podcast* and *video-streaming* industry due to its ability to correct voice recordings quickly via *pin-point typing*.
3) **VoiceApp**: VoiceApp is currently available to the *iOS* users on the *Apple Store*. The application is gradually gaining traction due to its high-quality output which can be used to impersonate famous celebrities. It deploys AI to recognize your voice and then creates a computer-generated impression of that which is apparently very hard to distinguish.
4) **Respeecher**: Respeecher is another alternative that is great for *filmmakers, game developers*, and other *content creators*. It captures *emotion* to some extent with a provision of *creative controls* (handling) to modify the content without re-recording. It combines classical *digital signal processing algorithms* with proprietary *deep generative modelling techniques* to produce its high-quality computer-generated voice.
5) **Screaming Bee** (products): Screaming Bee is a company that has several voice-changing technologies from **MorphVOX Pro** primarily for *gamers* with high-quality *voice morphing* and *background noise suppression* to **ScriptVOX Studio** for writers to bring stories to life via *storyboards*, etc.

SUMMARY

To wrap up in a nutshell, synthetic voice creation & recognition is the future that will assist mankind in various spheres or sectors of operations from the ingestion of documents to note-making that is clear, comprehensive, and absolutely flawless for the needs, wants, and desires of the end-users. In this report, we have discussed about the block diagram for this entire conceived project endeavour along with the working mechanism of the conceptual model in a systematic manner from data collection right down to the product deployment. We also delineated about the necessary technical considerations that are to be adhered to in order to deliver an agile project within a few months. Finally, we briefly went through some already existing products like Lyrebird AI, Overdub, VoiceApp, Respeecher, etc., that serve the purpose to some extent with a massive scope for refinement in the times ahead to incorporate complex human speech features in an emphatic manner. However, since this is a conceptual report, more empirical assessment is required in order to validate the furnished model in this report & complete a project on this subject with a detailed project timeline and well-informed budget allocation.

Nevertheless, the future looks really rosy in this regard, and with a little more research and a few advancements like Ubiquitous Computing, In-text Natural Language Generation, etc we could be looking at something that will perhaps become a household name for many quotidian tasks & activities and even a few glitzy ones like singing, acting, gaming, etc. So, all in all, it is safe to say that synthetic voice creation is a realm still in its early development phase right now that can perhaps one day transform the entire landscape of this world if unleashed to its full potential!