Voice Assisted Smart Notes Application using Deep learning synthesizers along with “Deep Neural Networks (DNNs)

Suksham Gupta¹
Assistant Professor
Lovely Professional University

Rajendra ²
Student
Lovely Professional University

ABSTRACT

“Smart notes deals with the conversion of text to speech and speech to text. This application helps the users to communicate with each other easily and create a fun environment. Smart Notes Talking Application deals with the conversion of text to speech and speech to text. This application helps the users to communicate with each other easily and create a fun environment. In today world, around the world, huge number of users are using mobile phones and tablets that run on the base of android operating system. Android has gained this popularity because of its multitasking capacity, diverse, device options, ease of access. An android application called “Java Programming Speech Recognition Application” specially designed for handicapped users who lack capacity to type on keyboard. By using this application, that user can write a compute program by pronouncing the words without the help of a keyboard. In this case, the person would have to dictate the commands essential for the programme to run.”

1. INTRODUCTION

Make life Easy – no more struggling to write down texts on paper. For students, thinkers, bloggers, writers, busy people and anyone who prefers fast and easy typing. Smart Notes applications is a platform integrated in the form of a mobile phone application to provide people with the ease of creating their notes and managing them at one place. Through this the users will get rid of stocking piles of notebooks, can easily create and customize their virtual notes and save them.

1.1 General description

1.1.1 Functional Requirements

“It is necessary to incorporate all these functionalities into the system as part of the contract. In contrast to the nonfunctional requirements, they are the requirements issued by the user that can be identified in the
end product. For example, a doctor should be able to retrieve information from his patients in a hospital management system. Can high-level functional requirement involve multiple system-outside interactions or dialogues. To describe the functional requirements accurately, all scenarios need to be listed. There are many ways to express functional specifications, such as natural language, a standardized or formatted language with no strict syntax and a formal specification language with proper syntax.”

TTS: Text to speech. Conversion can be done by the users whenever they can by performing through add on application in the personal inbox. It converts text to speech when we give input in text.

STT: Speech to text. Conversion can be done by the users whenever they can by performing through add on application in the personal inbox. It converts speech to text where ever we give input in speech.

LC: Language Converter is a term that could refer to a compiler, assembler, or interpreter; anything that translates code into another language.

1.1.2 Non-functional requirements:

“ These are simply the performance requirements that the project contract allows the program to fulfill. The priority and magnitude of the application of these considerations varies from project to project. These are also referred to as non-compliance requirements. We deal with issues such as:”

- **Portability:** The probability of a consumer finding a bug is also known as BugProbability.
- **Security:** aims to detect device vulnerabilities and to establish that its information and assets are secured against possible intruders.

2. OBJECTIVE

“Smart notes mostly deals with the users who are disabled like handicapped and dumb. It makes the user comfortable with out making him to type. This project can run in multiple android phones so that users can get easily access to it.

The project deals with retrieval of information through an internet based mobile phone. It stores all the information from all the users of an organization and maintains files, which makes user to recall his previous text by one click go.

This application would be helpful for those who face difficulties in conversation because of language barriers.”

3. EXISTING SYSTEM

“The system starts with whenever there is a text to speech or speech to text is given as a input by the user into the application then the add application takes part while the user is texting to other user. Once the user gives the input it converts it to text to speech or speech to text and saves it into data.”

3.1 Problems in the existing system:

“If the sentence is too big it will be not able to convert. The text we are speaking should be clear
otherwise it won't read the text and its lead to time consuming in some particular cases due to improper voice input.”

3.2 Risks involved in existing system:
“The current system takes time and results in the absence of an unnecessary broad paragraph. Some of the risks involved in the present system is:”

- It might effect the disabled users while they are giving the input in large sentences.

4. PROPOSED SYSTEM

“The Application for Speech Recognition helps the users to create their own Java software through the Java framework and the application's code. After the code is saved, the software translates the spoken words into texts and imports the file into a compiler and runs the program. The user must start the program with className while using this application, followed by elements and declaration of methods. Its prepared statements and blocks, the Speech Recognition software supports the system process. For example, the consumer wants not every single specific part of them main solution to be dictated. After the program is complete, the user can dictate the word “save” which will save. This system has the following functional divisions. users (Normal Users/Disabled/Illiterates)”

5. PROBLEM ANALYSIS

“Once the data requirements are gathered the consecutive step to be performed is to analyze the data we have. It would be more challenging while we understand the context and after this we need to have a clear picture about the existing system so that we can differ our system from the existing one. The next step is to understand the requirements as we discussed earlier. Both are equally important activities, but the first activity is used to establish functional specifications and then to design the proposed system successfully. Understanding the properties and requirements of a new system is more difficult and requires creative thinking and understanding of existing running system is also difficult, improper understanding of present system can lead diversion from solution.”

5.1 PRODUCT DEFINITION

“Smart notes is a common daily usage application that enables the user to comfortably convert voice into textform and vice versa. Using this feature, visually specially abled and other people can communicate with others seamlessly.”

“We want to show the machine learning approach for the two modules in the application. Speech-to-text and text-to-speech. Due to the unavailability of the required dataset for fluent English conversation as even Google required 1 year for the collection of data from whole world throughout the internet. We can’t possibly get that much data. And we don’t have medium to save that much data. So to complete our
project we decided

To first create the application in android studio by using libraries and show in this module

How that library can be trained in machine learning via model. We also decided to create TTS and Speech-to-text in python via API as many software do.”

5.1.1 Speech-to-text

“Hello Sir, I am Jarvis? This sound must be familiar to everyone who ever watch iron man or who use google(siri). Nowadays, nobody write the whole sentence in Google search, I certainly do not. Most of the time I just say the search text and Google do the rest.

It saves me a lot of time and I can save my time and get back to work. A win-win situation for everybody! But question is how does Google understand what ever we are saying? And how does Google’s convert our question or voice into text on our phone’s display?”

And, This is where we are working on, the miracle of speech-to-text model comes in.

Google uses a mix of deep learning and Natural Language Processing (NLP) Viterbi, A.J. (1967) techniques to

go through our query, fetch the answer and show it in the form of both voice and text.

5.1.2 Components of wave

- Amplitude – It refers to the maximum displacement of air molecules from the rest position
- Crest and Trough – Crest is highest point whereas trough is lowest in the wave
- Wavelength – Distance between two consecutive crests or trough.
- Cycle – All “the audio signals in nature travel in cyclic form. On signal form of a cycle consists of one upward and one downward movement in respect to axis.

Fig 5.1 Speech modulation (Xiong, X. (2009))  
Fig 5.2 Longitudinal wave generated Xiong, X. (2009)
• Frequency – It refers to how signal is changing per unit time.”

Fig 5.3-Components of wave Vintsyuk, T.K. (1968)

5.1.3 SAMPLING

A “continuous representation of amplitude can be called as audio signal as it changes with respect of time. Here, time can even be less than microseconds. That is reason that audio signal is considered as analog signal.” Analog signals takes too much memory storage space as they have an infinite number of samples and processing them is too much computationally demanding. So, to solve this problem, we need a way to convert analog signals to digital signals so that we can work with them easily Trentin, E. and Gori, M. (2001). By selecting “a certain number of samples per second from the analog signal we can convert analog signal into digital signal and this process is” called Sampling of the signal. Do you guys understand what we are doing below? Here we are converting an audio signal to a digital signal by sampling and then it can easily be stored and processed efficiently in memory. The main thing to take away from the figure above is that even after sampling the analog signal, we can recreate an almost equal audio wave as I have chosen a high sampling speed.

Fig 5.4 Sampling of a signal Tebelskis, (J. (1995))

5.1.4 Different Feature Extraction Techniques for an Audio Signal

The “first step in recognizing speech is to extract the features from an audio signal that we will later enter into our model. So now, we're going to see the various ways to extract features from the audio signal.”

1.) Time Domain

Here “the audio signal is defined as a function of time by the amplitude. In simple words, it is a correlation of time and amplitude Xiong, X. (2009). The characteristics are the amplitudes recorded at different intervals of time.” The limitation of the time-domain analysis is that the information on the signal rate that is analyzed is completely ignored.
2.) Frequency Domain

The “audio signal is displayed in the frequency field as a frequency function with amplitude. Simply put a plot between amplitude and frequency. The characteristics are the amplitudes recorded at various frequencies.” Salvi, G. (2006)

3.) Spectrogram

Has a spectrogram ever been heard? This is a 2D image between time and frequency where each point in the picture reflects the value of a specific frequency in terms of color intensity at a certain time. Simply put, the spectrogram is a frequency spectrum (wide color range), as it varies with time.

6. IMPLEMENTATION

- Import the libraries
  Firstly, import into our notebook all the requisite libraries. The Python libraries used for the treatment of audio messages are Librosa and Scipy.

Libraries imported in the project are –

1. “Os – This gives us the opportunity to communicate with the operating system underlying our python.OS, comes with standard utility modules from Python. This device provides a compact feature based on the operating system. Include many functions” for interacting with the file system, the * os * and * os.path * modules.

2. Librosa – “Librosa is a Python package mainly for music and audio analysis. It can be used primarily for music information collection or audio processing. It offers the building blocks needed for music recovery systems Siniscalchi, S.M., Yu, D., Deng, L. and Lee, C.H. (2013).”

3. Scipy- “It is a math, science and technology library. Scipy contains a wide range of subpackages to solve the most common science computing problem.

4. Matplotlib- “The main purpose is to plot the graph. Matplotlib is a great library for 2D plots of displays in Python.

5. NumPy – This library is for high dimensional array and many mathematical function to operate on the array. NumPy is a general-purpose array-processing package. This provides a multi-dimensional high performance array object with tools to work with the arrays. It is the main kit for Python science computing.

6. “Wavfile – This library is to help to write the audio signal 9n raw format and also read the wav file from the system. WAV files contain a sequence of bits representing the raw audio data, as well as headers with metadata in RIFF (Resource Interchange File Format) format. Of CD recording, the audio sample (a single air pressure audio datapoint) is to be captured at 16 bits with 44,100 samples per second, according to an industry standard.”
6.1 Model architecture

Model building

According to the model structure we have built the model. We also define the loss function across entropy since it is a multi-classification problem. For callbacks we define check points to save the best model. For training the model we took 32 as batch-size and evaluate the performance on the holdout set.* Diagonsis To understand the performance of our model we have to diagnose it by plotting the graph. We will build the speech-to-text model using Conv1d. (Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K. (1989) Conv1d is a neural network that works together in just one dimension.

Conv1d is a neural network that works together in just one dimension.

Here is the model architecture:

Fig 6.1 Model Architecture of Conv1d (Tebelskis, J. (1995))

6.1.1 Loading the best model

Now, we have to load our model which we saved in our checkpoint during training the model.

6.1.2 Prediction

Now, comes the point where we have to predict the audio signal.

6.1.3 Record and test

To better use the model we add something here. We have allowed the user to give the commands in microphone from where it will store the audio in wav format and then we convert it into text by using the prediction of our trained model.

6.2 Text-to-Speech

If we want the artificial production of human speech then that process is called Speech synthesis. A text-to-speech (TTS) system change normal language text into speech. Naturality and intelligibility are the most important qualities of a system of voice synthesis. Naturality describe show close the output sounds, such as the human speech. This is a natural and intelligible synthesizer for speech, too. Systems of speech
synthesis usually attempt to maximize the two features.
The two principal technologies producing waveforms in synthetic speech are concatenatory synthesis and
synthesis of formants. There are strengths and weaknesses in each technology, and the intended uses of a
synthesis system typically determine the approach used.

6.3 Concatenation synthesis

The “synthesis of concatenations is based on the combination (or combination of) of the voice parts.
The most natural synthesized speech is typically produced by concatenative synthesis. Nonetheless,
inconsistencies between natural speech variations and the design of automated waveforms segmentation
techniques can sometimes lead to audible performance failures. Three main concatenative synthesis
subtypes are present.”

6.3.1 Unit Selection synthesis
Synthesis “unit selection uses large speech databases. During the creation of a database, the individual
phone calls, diphones, half-phones, syllables, morphemes, words,
Sentences and sentences of the recorded utterances are separated into some or all of those. The segments
are normally divided by the use of a specially modified speaking recognizer set to "forced alignment" with
manual correction later on by using waveform and spectrograph visual representations.

6.3.2 Domain-specific synthesis
In “order to create complete utterances, a domain-specific synthesis combines pre- registered words and
phrases. The technology is very simple to implement and has long been in commercial use in devices such
as speech timepieces and calculator systems. It is used for applications where the various texts produced on
a system are limited to a particular domain, such as announcements of transit schedules and weather
reports.

6.3.3 Diphone synthesis
Diphone synthesis uses a low speak database with all the sound to sound diphones in a language.

6.3.4 Format synthesis
Format “synthesis does not use human samples of speech during runtime. The output of a synthed voice is
instead produced using an additive synthesis and acoustic model. Parameters like the fundamental
frequency, voice and sound levels vary over time, so as to create an artificial waveform. This is sometimes
referred to as rules-based synthesis; Nonetheless, there are also several rules based elements in many
concatenative systems. Many synthesis-based systems create artificial, robotic speech which will never be
mistaken for human speech. Nevertheless, the purpose of a speech synthesis system is not always total
naturalness, and the synthesis systems have benefits over concatenative systems Format”synthesized
speech, even at high speeds, can be confidently understood, preventing acoustic crashes which often
plague concatenating systems.

![Graph](image-url)

**Fig 6.2 Format synthesized speech (Trentin, E. and Gori, M. (2001))**

### 6.3.5 Deep learning

Deep learning synthesizers are using “Deep Neural Networks (DNNs) Schmidhuber, J. (2014) trained on recorded voice information. Some DNN speakers approach the quality of the human voice. Examples include DeepMind's WaveNet, Google's Tacotron and Baidu's Deep Voice (using WaveNet technology).”


### 6.4 SAMPLE LEVEL GENERATION OF AUDIO WAVEFORM

Our target is speech generation. Through digitizing the audio waveform, audio files are interpreted in a computer. This is basically an audio time-series. So it makes more sense to generate audio samples directly instead of generating some latent parameter and then processing them to get the audio. Here, try is made towards text to speech functionality by using API(Win32com.client). The range of methods for getting excited, one of which is the library shipping process. Show method passed by the SAPI. SpVoice argument It interacts to speak what you type in from a keyboard with the Microsoft SpeechSDK. The win32com package- The "ni" module of Python and the whole package are known as "win32com" for the simplification of the structural structure. Win32com does not provide any functionality, as is normal for such packages. The following are some of the modules:

1. **win32com.pythoncom-core C++support.**

   Alternatively, the module is used by the other "helper" which derive itself from the core services fromPythoncom.

2. **win32com.client package**

   Python's support for COM customers. The dynamic usage of COM clients, the module for generating files for certain COM servers etc are possible under some of the modules in this package.

3. **win32com.serverpackage**
Aid for Python COM servers. The modules in this package provide most of the framework to convert Python classes magically into COM servers, display the right public methods, register your server in the registry, etc.

4. **win32com.axscript**

   Implementation of ActiveX Python Scripting.

5. **win32com.axdebug**

   Implementation of Python successful debugging

6. **win32com.mapi**

   MAPI and Windows Exchange Server administrators Tools

   The C++ support for all related objects is the pythoncom module. In general, this module will not be used directly by Python programmers, but will use helpers and functions from win32com. The C++ interface to this module is like COM-objects that have the "quest api)" (or "invok)" (methods are being implemented in pythoncom, just like the C++ API. If you use COM in C++, you don't call a method straight away, you use pObject->Invok(.... MethodId, argArray...). Likewise, you must use the Invoke method to call the exposed method an object if you use pythoncom directly.

   Python wrappers are available to hide this raw interface, meaning that the pythoncom module should almost never be used directly. Such helpers move to the underlying call a normal interface (e.g. obj. SomeMethod)). To use Win32com.client first, we have to install pypiwin32 in our terminal.

   Import the required model- Here to use this API first we have to import the win32com.client

   Calling the Dispatch We have to call the dispatch method which interact with Microsoft Speech SDK to speak the given input from the keyboard. Stop the program To stop the program we gave the option for user to stop by using CTRL+Z.

6.5 **Outlook of the designed interface**

   In the very first interface the app displays 2 options whether to convert voice to text or from text to voice. The user will choose the option he/she wants.”
“After selecting voice to text option, the application automatically opens the microphone of the device. The user can speak his voice now. After taking the voice as input, the application processes it and converts it into text. It will be displayed on the screen within fraction of time depending on the internet connection.”

“If the user selects the option of converting from text to voice, the application opens an interface for the user to type the text. Here, the user can type the text manually, or he/she can use the option of copying text from other application like whatsapp, chrome, Documents etc. and paste it here in the required text area.”

“After this step is done, the application reads out the text as a sound. Thus the user can listen to the text that is entered. There is also a feature of translation in the application. By using this feature, the user can convert from one language to another language, say english to hindi and vice versa In both the modules of converting voice to text, or text to voice, the feature of translation can be used. There is an extra button allotted for this purpose.”
7. CONCLUSION

“This project is mainly depend upon the organisations that are established earlier and is related to the management issue so particular institution. It covers the few of the issue that are being faced by the institutions in real time and the modules of the project will be helpful for those issues. “This project is successfully implemented with all the features mentioned in system requirements specification. The application provides appropriate information to users according to the chosen service. It can be monitored and controlled remotely.” It reduces the manpower required. All gathered and extra information can be saved and can be accessed at any time. So it is better to have a Web- Based Information Management system.”

8. REFERENCES


