A Modern Approach for Speech to Text and Text to Speech Conversion Application Using Machine Learning Techniques

[1]M.Vengateshwaran, [2]P.Sowmya

[1]Assistant Professor in CSE, [2]PG Scholar, Department of Computer Science & Engineering

Agni College of Technology, Chennai, Tamilnadu, India.

ABSTRACT - Over the past few decades, designers are considering a range of applications ranging from mobile communications to automatic machine learning. Speeches are less commonly used in the electronic and computer field due to the complexity and variety of signals and sounds. By the use of modern algorithms and methods, speech signals are processed to recognize text. In this project, we will build a online speech to text engine. The program receives speech during the run through the microphone and uses sample speech to recognize the text. The known text can be saved to a file. This is being developed on Java platform using the eclipse workbench. Our speech-to-text program directly gets and converts speech into text. It can add other great applications, giving users a different choice of data input. A text-to-speech system can also improve accessibility of the system by providing data access options for users who are blind, deaf or disabled. Voice SMS application allows the user to record and convert spoken messages into an text message. The user can send messages to the phone number which is entered. Speech recognition is done through the Internet, connecting to Google's server. The application is based on input messages in English. Speech recognition uses a technique based on hidden Markov models (HMM - Hidden Markov Model). It is currently the most effective and flexible method of speech recognition.

Keywords: machine learning, SMS, Speech recognition, hidden Markov models.

I. INTRODUCTION

Mobile phones have become an integral part of our daily lives, creating high demands for content that can be used on them. Smart phones offer advanced customer service for communicating with their phones but the natural way of communicating more remains speech. The mobile phone market offers many applications for speech recognition facility. The latest Google Voice and Siri actions are applications which control the functionality of a mobile phone, such as calling businesses and contacts, sending texts and emails, listening to music, browsing the web, and completing tasks. Both Siri and Voice features require active connection to the network to process requests and most Android phones can run on the fastest 4G network than the 3G network that the iPhone runs. There's also a problem with availability, Voice Actions which is available on all Android devices above Android 2.2, but Siri is only available for iPhone 4S owners. Siri's advantage is that it can work in a variety of phrases and applications and can understand and learn in the native language, whereas Google's Voice Actions can only work using particular voice commands. Here we have developed an application which sends SMS messages which in turn uses Google's speech recognition engine.

II. SYSTEM REVIEW

2.1 Network speech-to-text conversion and store.

Implementation of the present invention, a mobile phone includes a processor and a microphone and speaker in communication with the processor. The mobile phone also includes an actuator that communicates with the processor. The actuator is capable of outputting a haptic effect that is perceived by the user of the mobile phone. The user can use a telephone to record voice messages that the processor converts to text messages. Further, the processor is configured to analyze the voice message to determine the characteristics of the message.

For example, in one such implementation, the frequency, amplitude, and duration of the user's voice message are analyzed and the processor determines the haptic effect corresponding to each parameter.

The processor then associates the haptic effect or effects with the text version of the message and sends the message to the intended recipient. In a second implementation, user biorhythms are analyzed and the processor determines

 DOI : 10.35291/2454-9150.2020.0302

haptic effects corresponding to the user biorhythm responses. Thereafter, as described above, the processor associates the haptic effect or effects with the text version of the message and sends the message to the intended recipient. Voice and biorhythm analysis is similarly performed by the processor and is accomplished by determining a haptic effect transmitted to the intended recipient.

2.2 Converting text-to-speech and adjusting corpus.

Based on the above, an object of the present invention to provide an improved text-to-speech conversion apparatus and method to obtain better voice quality. Another object of the present invention is to provide an apparatus and method for corpus TTS adjusted to meet the needs of the target speech speed. To solve the above technical problem, the present invention provides a text-to-speech conversion method, the method comprising: text analysis step of converting the speech model generated by the first text to the corpus, the text is analyzed to obtain descriptive text annotation prosody information; prosodic prediction step, for predicting the parameters of the prosodic text-based text analysis result of the procedure; speech synthesis step of synthesizing the text based on the predicted prosodic speech text. wherein the descriptive information comprises text annotation prosody prosodic structure of the text, the method further comprises the prosody of the text data Gen target speech synthesized speech speed is adjusted. The present invention further provides a text-to-speech conversion apparatus comprising: text analysis means for generating text-based corpus by the first model to-speech conversion, text is analyzed to obtain the descriptive text annotation prosodic information, the text descriptive annotation information comprises prosody prosodic structure of the text; prosody predicting means for obtaining the information based on the text analysis apparatus prosodic text prediction; speech synthesis means, 12 for speech synthesis of the text based on the predicted prosodic text; a prosody Adjusting means for prosodic structure of the text will be adjusted to the target speed of the voice synthesized speech. According to another aspect of the present invention, the target speech speed corresponding to the speed of a second speech corpus. The prosodic structure comprises a prosodic phrase. The present invention prosodic phrase text length distribution is adjusted so that it matches the length of the second corpus prosodic phrase distribution. Prosodic phrase text so that the length distribution suitable for the target speech speed.

III. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

With the use of voice, Google's Voice Actions and iphone's Siri applications control a mobile phone functionalities such as calling businesses and contacts, sending texts and email,etc., Siri and Voice Actions require an active connection to a network to process requests. Majority of Android phones run on 4G network which is faster than the 3G network that the iPhone runs on. There is also the problem of availability; Voice actions are available on all Android devices above Android 2.2, but Siri is only available for iPhone 4S owners. Siri's will work in a variety of phrases and can understand and learn in the native language, whereas Google's Voice Actions can only work using particular voice commands.

Disadvantages

- Scalability issues and Inefficiency occurs when processing or analyzing big data.
- Sparse data problem will be occurred in text representation.
- Find proper text representation of data.

3.2 PROPOSED SYSTEM

In our proposed system, we propose a machine learning techniques in application. Voice SMS includes a direct speech input that enables the user to record information that is spoken as a text message, and sends it as an SMS message. After the application has been started, it shows the button that starts the voice recognition process. When a conversation is held the app opens a connection with the Google server and begins to communicate with it by sending speech signal blocks. At the same time a waveform image is generated on the screen. Speech recognition for the received signal is done on the server. Google has compiled the largest database of words derived from the daily input into Google search engines and the digital production of over 10 million books in the Google Book Search project. The database contains more than 230 billion words. When we use this kind of speech recognition we may have our voice stored on Google servers.

This fact provides a continuous increase in the data used for training, thus enhancing the accuracy of the program. Once the recognition process is complete, the user can see a list of possible statements. The process can be done repeatedly by clicking the button *Image Button*. To press the most accurate options, the selected result is captured in the interface of writing SMS messages. The SMS texting interface has all the common features.

The user can edit the text and enter recipient's number in the empty text box. Button *contacts* open the interface with contact numbers from the phone and enables the user to select the phone number for which message to be sent after pressing the *Send* Button. The customer receives a response through a small cloud (toast) when a message is sent.

Advantages of Proposed System:

- The present invention works by first identifying clauses in text sentences by finding the punctuation marks.
- Then analyzing the structure of each clause by locating keywords such as pronouns, prepositions and articles

 DOI : 10.35291/2454-9150.2020.0302

which provide clues to the intonation of the words within the clause.

- The sentence structure thus detected is converted, in accordance with standard rules of grammar, into prosody information, i.e. inflection, speech and pause data.
- Apply ML to improve the scalability and efficiency.

IV. SYSTEM ARCHITECTURE

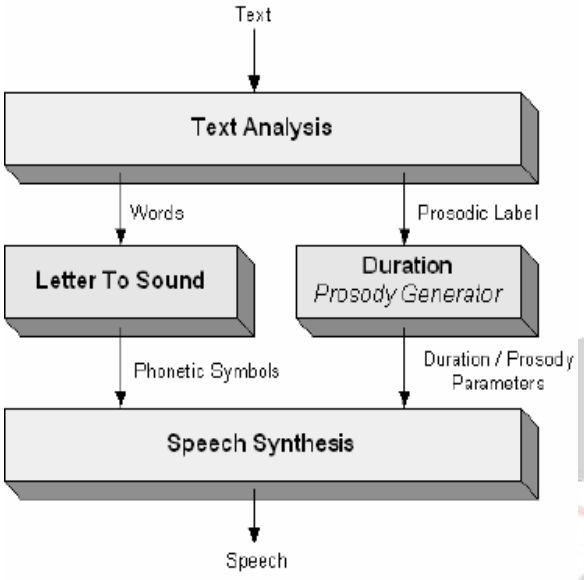

Fig. 4.1 System Architecture

We first collect the various text document in the form of datasets. we have to classify a data into a training and testing a document. Text or topic classification is processing of feature extraction by using of an Tokenization,stopword removal,stemming. By using of an TF-IDF weighting to find out the semantic feature extraction and ranking.After that indexing documents classify into two types indexed training and testing a documents. The training document is taken into input of classification model we apply Machine learning algorithms. Finally to predict the speech.

V. SYSTEM MODULES

5.1 MODULES

Module 1: Loading the Dataset

Module 2: Analysis of data and Preprocessing

Module 3: Feature extraction and Feature Selection

Module 4: Classification Model with GUI

Module 5: Predicted Speech

Module 6: Evaluation Process

5.2 MODULE DESCRIPTION

5.2.1 Loading the Dataset

In this module, we first load the text document dataset. Text were processed by tokenizing on whitespace and on punctuation subject to rules designed to keep together URL's, emoticons, usernames and hashtags. Some multiword entity names were collapsed into single tokens by using a gloss lookup derived from Wikipedia and query logs.

Fig .5.2.1 DFD for Loading the Dataset

5.2.2 Analysis of data and Preprocessing

After loading, we will analyze the dataset. Once the process is analyzed, information available in the dataset will be viewed. It is common to find that several attributes are useless. Thus, stop word removing algorithm has been applied. To initialize the algorithm, a set of stopword (such as a, a's, able, about, above, according, accordingly, and across) has set by the human before hand and hence stored in a text file. Then, the model can simply match the attributes with those preset stop word. After the stop word algorithm, a missing data checking algorithm is adopted. Missing data can be found using this algorithm and that is why it translates its value because data mining cannot operate under data condition. In preprocessing phase, the third algorithm used is the stemming. Since some words have the same meaning but in a different grammatical way, then it is necessary to combine them into one attribute. Like this, the terms are represented better in document and to acquire faster processing time, the dataset can be reduced.

5.2.3 Feature Selection and Feature extraction

Feature selection is one of the most important steps to prepare for data mining. To clear away noise feature, this is an powerful dimensionality reduction technique. The root idea of feature selection algorithm is to search for all the combinations of

 DOI : 10.35291/2454-9150.2020.0302

attributes in the data to find which set of features works best for predict. Therefore, attribute vectors can be reduced by the number where the most logical ones are stored and the wrong or repeated ones are removed. The transition from high-dimensional data to reduced feature data is known as Feature Extraction. The extracted features are expected to be relevant information from the given input dataset. There are various algorithms for feature selection and extraction. The main algorithm to be used for data reduction.

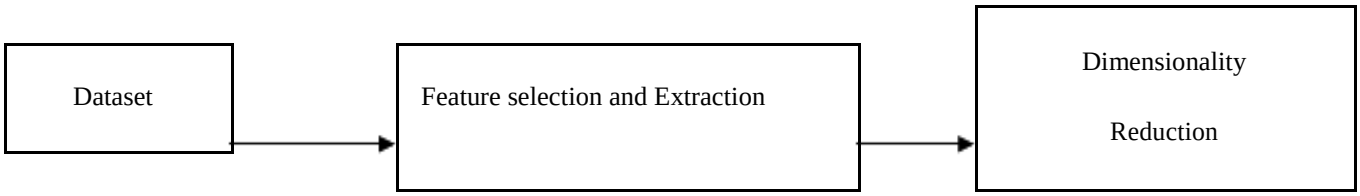

Fig .5.2.3 DFD for Feature Selection and Feature extraction

5.2.4 Classification with GUI

Home

Application Modularization is the process of using Modular Programming in our application. Modular programming is a way of building software to split functionality into independent, dynamic modules, so that they each contain everything needed to perform a specific task.

Message view

Message module is the backbone of the message stack. Enables logging and displays program events in many different use cases.

Register view

Registration is one of the main modules in any data management system. A patient's medical record management starts with registering a patient with the system. OpenMRS being the default and cost-effective solution for the management of medical records also requires a customized patient registration system. As all OpenMRS applications may differ in the type of information they may require, it is very important to keep the registration module optimized in such a way that it can be configured to capture patient registration information depending on the needs of the implementer.

Send message

The messaging module lets you send all kinds of messages, including email and SMS. It also helps you manage various aspects of messaging support, including managing addresses and scrolling past conversations. Finally, it has an easy-to-use API for other modules to add messages to their projects.

View message

Each message gateway is told to process incoming messages before sending outgoing messages. At the end of the receive Messages method, each message recipient object attached to a new message should have a status of 'received'.

Input design

The present invention provides a reference to a time domain technique which, in conjunction with a simple microprocessor, allows the construction of real-time speech sounds without a limited number of digitally recorded waveforms. The technique employed lends itself to implementation entirely by software, and permits a highly natural-sounding variation in pitch of the synthesized voice so as to eliminate the robot-like sound of early time domain devices. The software implementation of the technique of this invention requires no memory capacity or very large-scale integrated circuitry other than that commonly found in the current generation of microcomputers

Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs.

5.2.5 Evaluation process

The evaluation process takes place to predict the accuracy of the text to speech classification. The evaluation is in the form of graphs. To evaluate the performance of text document. Two types of approaches, including modifying training data and modifying the training objective of HMM, are proposed to make use of micro blog-specific information. To Apply the machine learning algorithms to improve the efficiency.

VI. CONCLUSION AND FUTURE ENHANCEMENT

Speech technology is especially interesting because of the direct support of communication between humans and computers. By using the speech tool, which works on the Internet, it allows for faster data processing. Another advantage is the larger databases used. The lack of a Voice application is its only adaptation to the English language, and the need for endless Internet connectivity. The aim and future is the development of models and multilingual data that can form the basis for daily use of these technologies. The basic goal of the program is to allow the sending of text

 DOI : 10.35291/2454-9150.2020.0302

messages based on spoken voice messages. Further work is underway to develop a speech recognition model for a different language.

Future enhancements to this work include the adaptation of the proposed model in the various domains. Even though, the model utilizes the semantics of words and phrases, the sentence level semantics are not addressed and hence there is a room for incorporating those features for improving the classification accuracy for better prediction.

REFERENCES

[1] Roth, D.L., Cohen, J.R., Johnston, D.F., Grabherr, M.G. and Porter, E.W., Voice Signal Technologies Inc, 2009. *Combined speech recognition and text-to-speech generation*. U.S. Patent 7,577,569.

[2] Bijl, D. and Hyde-Thomson, H., Speech Machines PLC, 2001. Speech to text conversion. U.S. Patent 6,173,259.

[3] Pelaez, M.B. and Verma, C., Nokia of America Corp, 2006. *Network speech-to-text conversion and store*. U.S. Patent 7,136,462.

[4] Shi, Q., Zhang, W., Zhu, W.B. and Chai, H.X., Nuance Communications Inc, 2013. *Converting text-to-speech and adjusting corpus*. U.S. Patent 8,595,011.

[5]Pelaez, M.B. and Verma, C., Nokia of America Corp, 2006. *Network speech-to-text conversion and store*. U.S. Patent 7,136,462.

[6] J. Tebelskis, Speech Recognition using Neural Networks, Pittsburgh: School of Computer Science,Carnegie Mellon University, 1995.

[7] S. J. Young et al., "The HTK Book", Cambridge University Engineering Department, 2006.

[8] K. Davis, and Mermelstein, P., "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoust., Speech, Signal Process. vol. 28, no. 4, pp. 357–366,1980.

[9] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms.", Technical standard ES 201 108, v1.1.3, 2003.

[10] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164–171, 1970.

[11] Rigutini, L. 2004. Automatic Text Processing: Machine Learning Techniques. Ph.D. Thesis, University of Siena

[12] Bernotas, M., Karklius, K., Laurutis, R., and Slotkiene, A. 2007. The peculiarities of the text document representation, using ontology and tagging-based clustering technique. Journal of Information Technology and Control. Vol. 36, pp.217 – 220.

Authors Profile:

Mr.M.Vengateshwaran M.E.,
Assistant Professor in CSE
Agni College of Technology, Chennai
Area: Machine Learning, Big Data, Data mining, IR

P.Sowmya (M.E).,
PG Scholar
Agni College of Technology, Chennai

 DOI : 10.35291/2454-9150.2020.0302