International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 11, November - 2013

VOICEBOX – Voice Operated Mailing System

Rahul Vishwanath Shinde
Dept. of Computer Science
Sinhgad Institute of Technology
Lonavla, India

Kunal Shankar Kusalkar
Dept. of Computer Science
Sinhgad Institute of Technology
Lonavla, India

Abstract: The Internet has became an important tool for learners to acquire information and knowledge that encompasses various elements such as text, numeric, graphic, and animation for their learning process. Anyhow, the visually impaired learners have no access at all to this tool nor can it be easily taught to them as they are not able to see the links in the WebPages. This system that has the capability of access to World Wide Web by browsing in the Internet, checking, receiving and sending email, searching in the Internet, and listening to the content of the search only by giving a voice command to the system. It can be employed as an aid for the people who suffer with visual impairment. It helps us to convert written English text to audio files and play them.

Keywords: Concatenative speech Syllable units [1], Speech synthesizer [1], Speech to Text conversion [5], TTS-Text to speech [4]

I. INTRODUCTION

Today, people spend most of their time interacting with computers. Life is running at a microchip speed. Over the last decade, these electronic tiny minuscule signals have fundamentally revolutionized the way we live. People are spending more time with machines than humans. They communicate through mails across the globe for official and unofficial information. Emails have become part of our life. These mailing systems are developed by different organization with different modules and different feasibilities. But fail to reach the users who are physically disabled people. And also fail to provide better ease for the users. The frequent and prolonged computer sessions may pose physical health risks such as visual strain, posture and skeletal problems and harmful effects of radiation.

This problem can be overcome to some extent by enhancing the applications with speech synthesis and speech recognition systems. Here we propose a VOICEBOX – Voice Operated Mailing System which is a unique system developed for English language where the members of the system can exchange information and is also supported by a reader which would read the contents of the mail.

It provides the user to register him as a valid user of the VOICEBOX service and then allows the user to access the modules of the system where the valid user can compose mails, send mails, delete mails, and receive mails in the English language. The mails which the user receives are converted from text to speech, by using concatenative speech synthesis approach and an appropriate wave file is generated and then the speech of the text is played.

II. SYSTEM PREREQUISISTES

Current System

There are many current mailing servers like Gmail, Yahoo Mail, and Hot Mail. Let us discuss one by one.

A. Gmail

- All figures and tables you insert in your document are only to help you Gmail is a free, advertising-supported email service provided by Google.
- Users may access Gmail as secure webmail, as well via IMAP or POP3 protocols.
- Gmail currently provides more than 7 GB of free storage per account.
- Individual Gmail messages, with attachments, may be up to 25 MB, which is larger than many other mail services support.
- Gmail has search-oriented interface and a "conversation view" similar to an Internet forum.

B. Hotmail

- Windows Live Hotmail, previously known as MSN Hotmail and commonly referred to simply as Hotmail, is a free web-based email service operated by Microsoft.
- Hotmail integrates with Office Web Apps to allow high fidelity viewing and editing of Microsoft Office Word, Excel, and PowerPoint documents that are attached to the email messages.
- 10GB of Office documents (up to 50 MB each) attachment size.

C. Yahoo Mail

- Yahoo provides a web mail service named as Yahoo Mail.
- You can store unlimited mails in your Yahoo Mail box.
- It has 25 MB attachments size.
- Yahoo Mail provides 100 filters to automatically sort incoming messages.
- Yahoo Mail provides Protection against spam and viruses.
- Yahoo Mail provides features like POP3 support, Mail Forwarding facility, and SMTP support in some countries (but not in the US).
- If the account is not logged into for four months then it gets deactivated.

Yahoo provides latest beta version named Zimbra desktop which is allowed for all Yahoo users to use the software.

III. MATHS

ALGORITHM

A. TTS System Architecture

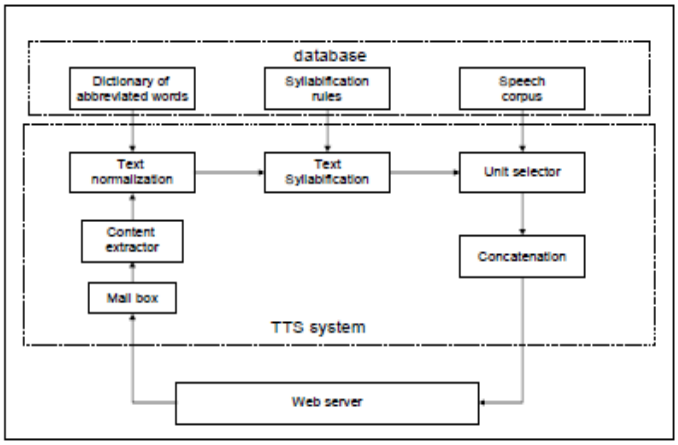

Fig. 1 TTS System Architecture

A text-to-speech (TTS) system converts normal language text into speech [7]. A TTS system has two parts: a front-end and a back-end. The front-end has two major tasks. The web server interacts with TTS system which contains different modules as shown in Fig-1. First, it extracts text from mail inbox and then normalizes the raw text containing symbols like numbers and abbreviations into the equivalent written-out words. This process is often called text normalization, pre-processing, or tokenization. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion [8]. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end often referred to as the synthesizer then converts the symbolic linguistic representation into sound.

B. Text Normalization

In practice, an input text from books, news articles consists of standard words (whose entries could be found in the pronunciation dictionary) and non-standard words such as initials, digits, symbols and abbreviations. Mapping of nonstandard words to a set of standard words depends on the context, and is a non-trivial problem. For example, number 120 has to be expanded to "nUta iravai" and rU. 300 to "mUdu vaMxala rUpAyalu" , and Ponu: 3005412 to "Ponu nambaru, mUdu sunnA sunnA, Edu nalagu okati reVMdu". Similarly punctuation characters and their combinations such as :, >, !, -, $, #, %, / which may be encountered in the cases of ratios, percentages, comparisons have to be mapped to a set of standard words according to the context. Other such situations include initials, company names, street address, initials, titles, non-native words such as bank, computer etc.

C. Text Syllabification

Speech sub units can be selected either at phone level, dyphone or triphone[9]. Our approach uses syllable as a basic unit. Text syllabification function extracts syllables from each normalized words and arranges it according to the sequence of the syllables based on Telugu phonological rules. Syllable structure is represented as C*VC* in most of Indian languages like Telugu, Tamil, Hindi. etc. The syllables in Telugu language can exist as vowel alone or as CV, VC, CVC, CCVC.

- Read the input text which is in WX notation.
- Label the characters of the normalized text as consonants and vowels using the following rule.
  - Any consonant except(y, H, M) followed by y is a single consonant, label it as C
  - Any consonant except (y, r, l, lY, lYY) followed by r is taken as single consonant
  - Consonants like (k, c, t, w, p, g, j, d, x, b, m, R, S, s) followed by l is taken as single consonant.
  - Consonant like (k, c, t, w, p, g, j, d, x, b, R, S, s, r) followed by v is taken as a single consonant.
  - Label the remaining as Vowel (V) or Consonant(C) depending on the set to which it belongs.
  - Store the attribute of the word in terms of (C*VC*)* in file1.
- For each word in the normalized text get its label attribute from file1.

If the first character is a C then the associate it to the nearest Vowel on the right.

- If the last character is a C then associate it to the nearest Vowel on the left Check.
- If sequences correspond to VV then break is as V-V.
- Else If sequence correspond to VCV then break is as V-CV.
- Else If sequence correspond to VCCV then break is as VC-CV.
- Else If sequence correspond to VCCCV then break is as VC-CCV.
- The strings separated by – are identified as syllable units.
- Repeat.
- Store the text in syllable form in file2 for synthesis process.

D. Syllable Extraction and Concatenation

This module will receive a sequence of syllables that has been properly arranged according to the raw text. Based on the list of syllable, Syllable Extraction module will search for speech units in the speech corpus [10]. The context of the syllable to be searched in speech corpus is given weights according to next, previous and the position of the syllable in the word. For context corresponding to next be given a weight age of 4, previous is given 2 and the matching of position in the word is given 1.

- Read the syllabified text.
- Compute the weight for each Syllable using the following steps.
  - If the search syllable matches with next, previous and position context in speech corpus is given weight of 7.
  - Else If it matches with next and previous but the position context does not match is given weight of 6.
  - Else If it matches with next and position us, but the previous context does not match is given weight of 5.
  - Else If it matches with previous and position, but the next context does not match is given weight of 3.

- Else If it matches with previous, but the next and position context does not match is given weight of 2.
- Else If it matches with position, but the next and previous context does not match is given weight of 1.

- The syllable which gives maximum weight is selected for synthesis.
- If the next context of the syllable is space or end mark of sentence then include the silence unit depending on the unit.
- Concatenate all the units selected and generate a single wave file which has the utterance of the given text.
- Play the wave file.

E. Concatenation of Wave files

The concatenation of the wave files is done using Audio Input Stream. An audio input stream is an input stream with a specified audio format and length. The length is expressed in sample frames, not bytes. Several methods are provided for reading a certain number of bytes from the stream, or an unspecified number of bytes. The audio input stream keeps track of the last byte that was read.

The AudioSystem class includes many methods that manipulate AudioInputStream objects. These methods supports to obtain an audio input stream from an external audio file, stream, or URL write an external file from an audio input stream convert an audio input stream to a different audio format. The AudioSystem class acts as the entry point to the sampled-audio system resources. Using this class we can query and access the mixers that are installed on the system.

AudioSystem includes a number of methods for converting audio data between different formats, and for translating between audio files and streams. It also provides a method for obtaining a Line directly from the AudioSystem without dealing explicitly with mixers.

MATHEMATICAL REFERENCES

The Fourier transform is a mathematical operation that decomposes a function into its constituent frequencies, known as a frequency spectrum. For instance, the transform of a musical chord made up of pure notes is a mathematical representation of the amplitudes (and phase) of the individual notes that make it up.

There are several common conventions for defining the Fourier transform $\hat{f}$ of an integral function f: R → C (Kaiser 1994). The definition is:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \xi}\, dx,$$

For every real number ξ.

In mathematics, the discrete-time Fourier transform (DTFT) is one of the specific forms of Fourier analysis. As such, it transforms one function into another, which is called the frequency domain representation, or simply the "DTFT", of the original function (which is often a function in the time-domain). But the DTFT requires an input function that is discrete. Such inputs are often created by digitally sampling a continuous function, like a person's voice.

Given a discrete set of real or complex numbers: $x[n],\ n \in \mathbb{Z}$ (integers), the discrete-time Fourier transform (or DTFT) of $x[n]$ is usually written:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x[n]\, e^{-i\omega n}.$$

IV. SYSTEM ARCHITECTURE

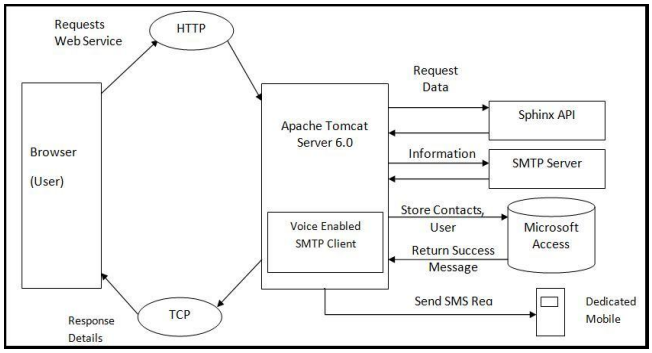

Fig. 2 Architecture Diagram

- *Browser (User):* This is the client side module used to access mail using compatible browser. User login in to his/her Email Account using HTTP/HTTPS protocols.
- *Server:* At the server side, we are using Apache Tomcat Server 6.0 to handle all the server side activities like text to speech conversion, connections, access database and send new mail direct to the registered mobile in the form of voice message, etc. Server sends new mail direct to the registered mobile in the form of voice message if the user is currently offline.
- *Voice Enabled SMTP Client:* Voice Enabled SMTP Client recognizes the voice enabled commands known as VOICE TAGS and performs appropriate action. We are using limited dictionary words for better accuracy. Inbox, compose, send, speak, read, home, logout, update user, delete are some of the voice tags.
- *Sphinx API:* Sphinx API is the standard API used for Text to Speech and Speech to Text conversion.
- *SMTP Server:* SMTP server is used for exchanging data between dedicated server and the client. SMTP is a delivery protocol only. It cannot pull messages from a remote server on demand. Other protocols, such as the Post Office Protocol (POP) and the Internet Message Access Protocol (IMAP) are specifically designed for retrieving messages and managing mail boxes. However, SMTP has a feature to initiate mail queue processing on a remote server so that the requesting system may receive any messages destined for it.
- *Microsoft Access:* For maintaining database we are using Microsoft Access. Database includes contacts, previous mails, drafts, etc.

V. FUNCTIONAL REQUIREMENTS

- System must provide facility to store and retrieve mails.
- System should maintain and link database for each user and manage their contacts.
- System should continuously monitor SMS of each & every user and enable/disable on user demand.
- System should maintain user roles in database.

A. PERFORMANCE REQUIRMENTS

High Speed:

System should process SMTP mails in parallel for various users to give quick response then system must wait for process completion.

- Time required checking user validity should be 2 sec or less.
- After mail arrival SMS alert should get within 10 sec.

B. SAFTY REQUIRMENTS

The data safety must be ensured by arranging for a secure and reliable transmission media. The source and destination information must be entered correctly to avoid any misuse or malfunctioning.

- The user must be validated on each logon to the mail account.
- Database must be backed up periodically to avoid loss of data.

C. SECURITY REQUIRMENTS

- Password protection enabled for each user.
- SMS alert is provided only to registered number.

D. SOFTWARE QUALITY ATTRIBUTES

- Availability- The software shall be available during normal operating hours (24 hours).
- Security- Access to various subsystems if any will be protected with a screen that will require a user name and password.
- Maintainability- The system is being developed in JAVA. JAVA is an object oriented language and it is easy to maintain.
- Portability- The system shall run in any Microsoft Windows environment mostly that contains JVM and Microsoft Access database.
- Reliability- Reliability should be gained by making efficient database updates and taking backup regularly.

E. DATABASE REQUIREMENTS

- Microsoft Access: For maintaining database we are using Microsoft Access. Database includes contacts, previous mails, drafts, etc.

VI. SYSTEM SCOPE

It is not possible to develop a system that makes all the requirements of the user. User requirements keep changing as the system is being used. Future enhancements that can be done to this system, this voice enabled mailing system is confined only to intranet, which can be extended and can be converted into a real time system. More number of syllables can be used based on the context and the project can be extended to use in different contexts and all Indian languages. It would be more usable if it is embedded with speech recognition system, which can be used for composing a mail.

- Mail client design: An intranet mail server that should provide the facility of composing mail, reply, forward, inbox, deleting mail etc.
- Notification for Voice Mail Service: The user will be notified about the mail he/she has received.
- SMS Alerts: User receiving Voice/Text Messages will be notified with SMS.
- Voice Tags in Mailbox: Voice tags can be enabled for the SMTP Client. User can enter voice tag and respective actions will be performed accordingly.

VII. TECHNICAL SPECIFICATIONS

A. ADVANTAGES:

- Easy and efficient UI.
- High quality mail reader, an aid for people with physical impairment.
- Multiple users can use the system at the same time i.e. flexibility.
- Fast and user friendly text to speech synthesis.
- Database provides naturalness in the speech.
- Security of data or information.

B. DISADVANTAGES:

- System fails if no internet connection.
- High performance text to speech API .i.e. Sphinx API.
- Mobile Equipment requirement for SMS alert.
- User vocal output should be clear for clear speech recognition.
- Additional hardware requirement .i.e. Microphone, Speaker set.

C. APPLICATIONS:

- Helping visually impairment learners to browse internet.

VIII. CONCLUSION

At present there is no such voice enabled mail reader which converts text to speech. We have used around fairly a large number of syllables which can be selected based on context. This mail reader makes our task more easy and efficient. VOICEBOX is a high quality mail reader which acts as an aid for people with physical impairment. This application can be installed on any PC and multiple users can use the system at any time. To incorporate more naturalness in the speech output, database which covers all possible syllables in different context should be maintained. The one more advantage of the project is to develop a fast and user friendly text to speech synthesis.

IX. REFERENCES

[1] K.V.N.Sunitha, N.Kalyani, "VMAIL-Voice Enabled Mail Reader", 2010 International Conference on Recent Trends in Information, Telecommunication and Computing

[2] Michael H. O'Malley, "Text To Speech Conversion Technology", Berkeley Speech Technologies, 1990

[3] Halimah B.Z. , Azlina A. , Behrang P. , Choo W.O, "Voice Recognition System for the Visually Impaired: Virtual Cognitive Approach", 2008

[4] Catalin Ungurean, Dragos Burileanu, "An Advanced NLP Framework for High-Quality Text-to-Speech Synthesis", 2011

[5] Abhijit V. Bapat, Lalit K. Nagalkar, "Phonetic Speech Analysis for Speech to Text Conversion", 2008 IEEE Region 10 Colloquium and the Third International Conference on Industrial and Information Systems, Kharagpur, INDIA December 8 -10, 2008.

[6] Elias Azarov, Alexander Petrovsky, Piotr Zubrycki,"Multi Voice Text To Speech Synthesis Based On The Instantaneous Parametric Voice Conversion ",

[7] S.Lemmetty, "A Review of Speech Synthesis Technology", Master Thesis,Department of Electrical and Communication Engineering, Helsinki University of Technology, Helsinki, Finland, March 1999.

[8] X. Huang, A. Acero and H.-W. Hon, "Spoken Language Processing A Guide to Theory, Algorithm and System Development", New Jersey: Prentice Hall, 2001.

[9] A.M. Zeki and N. Azizah, "A Speech Synthesizer for Malay Language", National Conference on Research and Development in Computer Science, Selangor, Malaysia, October 2001

[10] A.W. Black and N. Campbell, "Optimising Selection of Units from Speech Databases for Concatenative Synthesis", Proceeding Eurospeech '95, Madrid, Spain, September 1995, pp. 581 – 584.