Voice Based Search Engine And Web Page Reader.

[1,]Ummuhanysifa U [2,] Nizar Banu P K
[1,2,]*B.S. Abdur Rahman University Chennai*

Abstract *:This paper aims to develop a search engine which supports Man-Machine interaction purely in the form of voice. A novel Voice based Search Engine and Web-page Reader which allows the users to command and control the web browser through their voice, is introduced. The existing Search Engines get request from the user in the form of text and respond by retrieving the relevant documents from the server and displays in the form of text. Even though the existing web browsers are capable of playing audios and videos, the user has to request by typing some text in the search text box and then the user can play the interested audio/video with the help of Graphical User Interfaces (GUI). The proposed Voice based Search Engine aspires to serve the users especially the blind in browsing the Internet. The user can speak with the computer and the computer will respond to the user in the form of voice. The computer will assist the user in reading the documents as well.*

Keywords: *Web Search Engine, Web Page-Reader, Voice Recognition, Speech Synthesizer, Voice browser.*

I. INTRODUCTION

Information contained on the World Wide Web is inaccessible to many people. The web is primarily a visual medium that requires a keyboard and mouse to navigate. People, who lack motor skills to use a keyboard and mouse, find navigation troublesome. Visually impaired people have problems in accessing the web[11]**.** Those who temporarily cannot use a traditional web browser, as their eyes or hands are occupied or because they are not closer to their computer are at a minimum inconvenienced. Speech recognition and generation technologies offer a potential solution to these problems by augmenting the capabilities of a web browser.

Speech recognition accuracy can be improved in many ways, time frequency distribution [11]; HMM approach, Bayesian classification, wavelet transformation domain [2] or combination of such approaches can be used. Advances in voice recognition have made possible applications in robotics controlled by voice alone [1]. On the other hand Speech synthesis consists of three categories: Concatenation Synthesis, Articulation Synthesis, and Formant Synthesis. In [9] feature parameters for fundamental small units of speech such as syllables, phonemes or one-pitch-period speech, are stored and connected by rules. A voice browser is a web browser with at least one of the following capabilities:

- renders web pages in an audio format (speech generation)
- interprets spoken input for navigation (speech recognition)

One of the most popular search engine "Google", has introduced speech recognition in search engine, however it does not support web page reader.

Following this introduction, literature survey of relevant papers are described in section 1, proposed voice based search engine and web page reader's framework is shown and explained in section 2. Experimental results are presented in section 3. Finally, conclusion is given in section 4.

II. LITERATURE REVIEW

Voice-enabled interface with addition support for gesture based input and output approaches are for the "Social Robot Maggie**"** converting it into an aloud reader [1]. This voice recognition and synthesis can be affected by number of reasons such as the voice pitch, its speed, its volume etc. It is based on the Loquendo ETTS (Emotional Text-To-Speech) software. Robot also expresses its mood through gesture that is based on gestionary..Speech recognition accuracy can be improved by removal of noise. In [2], A Bayesian scheme is applied in a wavelet domain to separate the speech and noise components in a proposed iterative speech enhancement algorithm. This proposed method is developed in the wavelet domain to exploit the selected features in the time frequency space representation. It involves two stages: a noise estimate stage and a signal separation stage.

In [3], the Principle Component Analysis (PCA) based HMM for the visual modality of audio-visual recordings is used. PCA (Principle Component Analysis) and PDF (Probabilistic Density Analysis) are two

modalities information integrated together and obtained a Multi-Stream Hidden Markov Model (MSHMM). MSHMM method is widely used and very successful in audio visual speech recognition.

[4] Presents an approach to speech recognition using fuzzy modelling and decision making that ignores noise instead of its detection and removal. In [4], the speech spectrogram is converted into a fuzzy linguistic description and this description is used instead of precise acoustic features.In [5] Voice recognition technique combined with facial feature interaction to assist virtual artist with upper limb disabilities to create visual cut in a digital medium, preserve the individuality and authenticity of the art work.

Techniques to recover phenomena such as Sentence Boundaries, Filler words and Disfluencies referred to as structural Metadata are discussed in [6] and describe the approach that automatically adds information about the location of sentence boundaries and speech disfluencies in order to enrich speech recognition output.

Clarissa a voice enabled procedure browser [7] that is deployed on the international space station (ISS). The main components of the Clarissa system are speech recognition module a classifier for executing the open microphone accepts/reject decision, a semantic analysis and a dialog manager.

[8] Mainly focuses on expressions. To build a prosody model for each expressive state, an end pitch and a delta pitch for each syllable are predicted from a set of features gathered from the text. The expression-tagged units are then pooled with the neutral data, In a TTS system, such paralinguistic events efficiently provide cues as to the state of a transaction, and Markup specifying these events is a convenient way for a developer to achieve these types of events in the audio coming from the TTS engine.

Main features of [9] are smooth and natural sounding speech can be synthesized, the voice characteristics can be changed, it is "trainable. Limitations of the basic system is that synthesized speech is "buzz" since it is based on a vocoding technique, it has been overcome by high quality vocoder and hidden semi-Markov model based acoustic modelling.

Speech synthesis consists of three categories: Concatenation Synthesis, Articulation Synthesis, and Formant Synthesis. [10] mainly focuses on formant synthesis, array of phoneme of syllable with formants frequency is given as input, frequency of given input is processed, on collaborated with Thai-Tonal-Accent Rules convert given formants frequency format to wave format, so that audio output via soundcard.

III. VOICE SEARCH ENGINE AND WEB PAGE READER

As shown in Fig. 1, the proposed system receives voice through microphone as input; accuracy of the voice recognition can be improved by training the computer. As much the user trains computer the recognition is accurate, removal of noise from the speech also improves accuracy level [2]. Voice recognition engine converts the given voice input in the text format. The interpreted text displayed on search box. The relevant links will be retrieved from server database, displayed on the home page of the proposed search engine. The links displayed readout by the speech synthesis. Basic limitations of speech synthesis can be overcome by high quality vocoder [9].

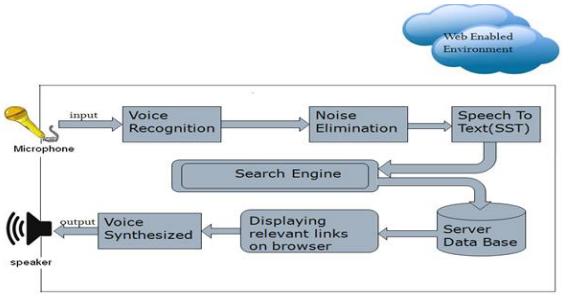

Fig. 1 Proposed Frame Work

A. Voice Recognition

Voice recognition is different from speech recognition. Voice recognition recognizes the voice of a particular person. The moment of recognizing a phrase is an event, which happens when the recognizer detects an input. An event handler is assigned to this event so there is access to the property of the event, such as the text

of the result of the recognition. It is important to notice that, there is no access to all of the properties of a recognition event. By that, it is meant; this event always returns a phrase, regardless of it being what was actually said by the user. For example, consider a user saying "life" and the system detects "like". That is because these two words have almost the same pronunciation. In such case, the system carries on the actions as the value being "like", which not something that the user inputs. The whole process of detection of words, that is, the process of translating spoken input (voice patterns) to text-words is handled by Speech Application Programming Interface (SAPI), where voice patterns are matched and from those phrases are created.

B. Noise Elimination

Noise elimination techniques are required to improve the accuracy of speech recognition. A framework of wavelet-based techniques to harness the automatic speech recognition performance in the presence of background noise is presented in [2]. [3] Proposed the integration of the audio and visual information would help to improve the recognition accuracy even at a low acoustic Signal-To Noise Ratio. PCA and PDF modalities integrated and obtained MSHMM.

C. Speech Synthesis

Speech synthesis refers to a computer's ability to produce sound that resembles human speech. Although they can't imitate the full spectrum of human cadences and intonations, speech synthesis systems can read text files and output them in a very intelligible. Many systems even allow the user to choose the type of voice [4] - for example, male or female, reading volume and speed can be controlled, emotions can be expressed [1],[8]. Speech synthesis systems are particularly valuable for seeing-impaired individuals.

IV. EXPERIMENTAL RESULTS

In the proposed system, for the prototype, language model is designed with the same items but different approaches are loaded into the recognizer. Four sample words such as India, machine, magnetic, magic are given as input, each words uttered 20 times Table 1 shows the comparison which gives the number of correct recognitions, in-correct recognitions and the number of times the system recognized nothing.

From Table 1 we concluded Table 2 which illustrates the error rate. Error rate is concluded from the following formula:

$$\left(\frac{\text{no.errors}}{\text{no.items}}\right) * 100 \qquad (1)$$

In Eq. 1, number of errors is the sum of unrecognized and incorrect recognition. We assume both as error since it is not the correct recognition.

The Eq. 1 gives a percentage of error rates on every approach. In Table 2 we compared the error rates of all four combinations and found the one with lower error rate. In this way the most accurate combination which results in lowest error rate is found. Then from the error rate the accuracy percentage is concluded, simply by subtracting error rate from 100 is shown in Table 3.

TABLE 1
COMPARISON OF COMBINATIONS

AC	*Recall/TP*	*FP*	*TN*	*FN*
0.84	*0.8*	*0*	*1*	*0.2*
0.92	*0.9*	*0*	*1*	*0.1*
1	*1*	*0*	*1*	*0*
0.92	*0.9*	*0*	*1*	*0.2*

TABLE 2
ERROR RATE OF COMBINATIONS

Combination No.	**Error Rate (%)**
C-1	20

C-2	10
C-3	0
C-4	10

TABLE 3
ACCURACY IN PERCENTAGE

Combination No.	Accuracy (%)
C-1	80
C-2	90
C-3	100
C-4	90

From the comparison of the four possible combinations, it is clear that only C-3 results in highest accuracy. Note that in Table 3 C-1, C-2 and C-4 only have 20%, 10% and 20% error rate respectively, which are quite satisfying whereas C-3 contain 0% error rate that is really appreciable. Fig. 2 shows graphical representation of Table1. Each sample word uttered 20 times, yields resultant graph as shown in Fig. 2

Table 4 shows accuracy, *TP, FP, TN, and FN* of 25 instances of each one of four sample items *India, machine, magnetic, magic*. The accuracy determined using equation *AC* may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases.

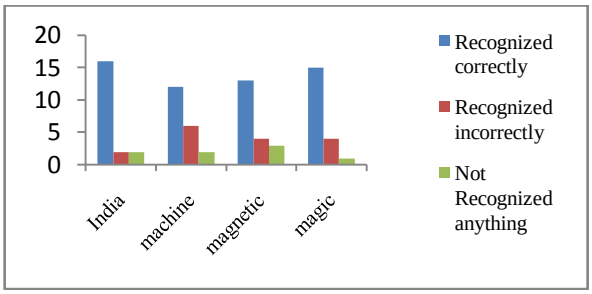

Fig. 2 Comparison of Combinations

TABLE 4
ACCURACY MEASURES

Combination No.	Sample words	No. of time recognized correctly	No. of times Recognized incorrectly	No. of times not Recognized
C-1	India	16	4	0
C-2	Machine	18	2	0
C-3	Magnetic	20	0	0
C-4	Magic	18	2	0

TABLE 5
PERFORMANCE ANALYSIS

Precision	***Sensitivity***	***Specificity***
1	*0.8*	*1*
1	*0.9*	*1*
1	*1*	*1*
1	*0.9*	*1*

Table 5 shows precision, sensitivity, specificity of 25 instances of four sample words as mentioned above.

A calm place leads to higher accuracy; certain hardware can also help in increasing the overall accuracy of the system. In our experience first we used a normal (non-noise-cancelling) microphone and later on we checked with a noise-cancelling microphone. There is a noticeable difference in recognition performance when a noise-cancelling microphone is used. On the other hand noise cancelling techniques such as Bayesian algorithms [2], time frequency distribution [11], fuzzy logics [4], etc can be implemented to improve the accuracy dramatically in the presence of noise.

V. CONCLUSION

A new search engine enabled with voice recognizing and voice synthesis mechanisms is proposed. From the Experiments and observations it is understood that the accuracy of the voice recognition can be further improved when we add additional hardware and software. Other techniques mentioned in section 2 can be applied in designing voice search engine to increase its performance and accuracy.

REFERENCES

[1] Arnaud Ramey, Javier F. Gorostiza, Miguel A. Salichs, "A Social Robot as an Aloud Reader: Putting together Recognition and Synthesis of Voice and Gestures for HRI Experimentation" International conference on Human-Robot Interaction, March 2012.

[2] Yu shao, chip-hong chang, "Bayesian Separation with Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition" Journal on systems, man, and cybernetics, volume 41, issue 2, March 2011.

[3] Zhanyu Ma and Arne Leijon, "A Probabilistic Principal Component Analysis Based Hidden Markov Model for Audio-Visual Speech Recognition", Conference on computing & processing, Digital Object Identifier: 10.1109/ACSSC.2008.5074819 - 2008 .

[4] Ramin Halavati, Saeed Bagheri shouraki and Saman Harati Zadeh, "Recognition of human speech phonemes using a novel fuzzy approach" ,Journal appliedto soft computing, Volume 7, Issue 3, June 2007

[5] Dharani Perera, "Voice recognition technology for visual artists with disabilities in their upper limbs", conference on Computer-Human Interaction**,** OZCHI '05, November 2005

[6] Yang Liu, Elizabeth Shriberg, Andreas Stolcke , Dustin Hillard, Mari Ostendorf, Mary Harper, "Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies" Journals and magazines, Volume 14, issue 5, September 2006

[7] Manny Rayner, Beth Ann Hockey, Nikos CVhatzichrisafis, Kim Farrell, Jean Michel Renders, "A Voice Enabled Procedure Browser for the International Space Station", Association for Computational Linguistics, June 2005

[8] Heiga Zen and Tomoki Toda " An Overview of Nitech HMM-based Speech Synthesis System for Blizzard challenge", Blizzard Workshop-2005

[9] Saiyan Saiyod, Sakchai Thipchaksurat, and Somsak Mitatha "Thai Speech Synthesis for Text-to-Speech based on Formant Synthesis Technique" , source www.ecti-thailand.org-, April 2012

[10] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to <ahem/> expressive speech synthesis", IBM T. J Watson Research Center Yorktown Heights, NY 10598, 2004

[11] Alexandros Potamianos, Member, IEEE, and Petros Maragos, Fellow, IEEE "Time-Frequency Distributions for Automatic Speech Recognition" , IEEE Transactions on Speech and Audio Processing, vol. 9, no. 3, March 2001