Available online at www.elixirpublishers.com (Elixir International Journal)

Image Processing

Elixir Image Processing 90 (2016) 37722-37725

ISSN: 2229-712X

Speech to Text Converter

Vandana Sawant, Serena Saldanha, Supriya Patil, Sweta Rajagiri and Ruchika Rokade
Department of Electronics and Telecommunication Department, Mumbai University,SIES GST,Sector 5, Nerul,India.

ARTICLE INFO

Article history:
Received: 23 April 2015;
Received in revised form:
8 January 2016;
Accepted: 14 January 2016;

Keywords
SAPI,
ASR,
DTW,
Microsoft Visual Studio,
C sharp.

ABSTRACT
People with disability such as visual impaired and also elderly for whom it's very hard to identify the screen text and area where the keyboard and mouse may not be an appropriate means of communication between systems, it would be a helpful to use voices to navigate and control the computer systems. Microsoft has designed an interface called SAPI (Speech Application Programming Interface) which supports dynamic speech input and output, and is integrated in our current operating systems. In this paper we have described a model which is developed for conversion of multilingual audio into multilingual editable text for continuous speech in offline mode. Automatic Speech Recognition, (ASR) is used that works with Dynamic Time Wrapping (DTW) algorithm. This text will be transmitted and displayed on computer or LCD. Software used is Microsoft Visual Studio. Coding language used is c sharp.

Introduction

Speech is the most natural form of interaction and communication between humans while symbols and texts are the most common form of transaction in computer systems. Hence interest for speech oriented human-computer interaction regarding conversion between text and speech is increasing gradually. Text-to-Speech (TTS) translation known as speech synthesis is straight forward and easier than Speech-to-Text (STT) conversion. The pronunciation, motion and manner of words are the aspects of voice biometrics and give difficulty in recognizing the speech and converting it to text.

In Speech synthesis technology, with a help of microphone, computer is able to capture the words spoken by a human. These words are later on recognized by speech recognizer and then the system outputs the recognized words. All words uttered by a human are recognized by a speech synthesis engine but practically speech synthesis engine's performance is influenced by number of factors like noisy environment, multiple users and vocabularies.

In computer science, speech recognition (SR) is the translation of spoken words into text which is also known as Automatic Speech Recognition, (ASR) or Speech to Text system (STT). For more accurate transcription, the system analyzes the person's voice and then the system will use it to fine tune the recognition of that person's speech. [2]

There are three types of speech recognition systems:-
Speaker independent - System that do not use training.
Speaker dependent - Systems that use training.
Speaker adaptive - Speaker adaptive systems are now emerging. These systems usually begin with a speaker independent model and adjust these models more closely to each individual during a brief period of training. [6]

This paper introduces Windows application design an attempt has been made to develop multilingual speech recognition for languages like Hindi, English, Marathi and French, then later converting them into text.

Literature Survey

Dynamic Time Wrapping [6]

DTW is a method which will allow a computer to find an optimal match between two given sequences (e.g. time series) with obvious restrictions. To determine a measure of their similarity independent of definite nonlinear variations in the time dimension, the sequences are "warped" in the time dimension non-linearly. This method is often used in Hidden Markov Models

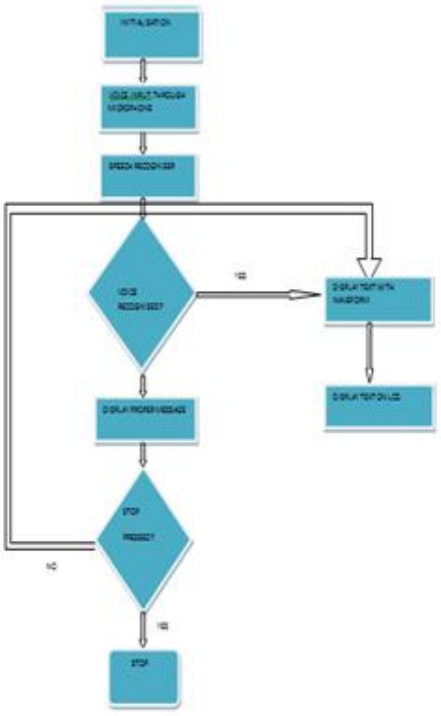


Fig 1. A flow chart of recognition and display on LCD

In the flow chart the process of speech recognition and its conversion is explained. The input to speech recognizer is given through microphone which reduces noise. If the spoken voice

Tele:
E-mail addresses: ruchikarokade@gmail.com

matches with a word stored in dictionary then the word gets displayed on the screen with the wavefile. In this project we are working with asynchronous mode of speech detection, so the process of speech recognition will not stop until it is ended manually.

Working

Select language

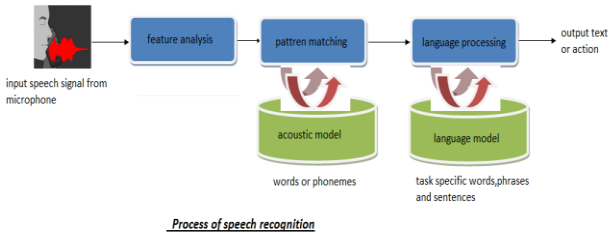
Fig 2. UI for selecting language

In this, one of the four languages is selected after which .resx file that is the dictionary which contains list of words we are going to display; called grammar is loaded into speech recognition engine. After that the engine sets to a culture (language) i.e. it prepares itself to recognize a particular language. UI controls get updated according to the selected language.

Speech Recognition

Fig 3. The UI form after recognition

MENU is at the top and controls are provided at right. "START"-The process of speech recognition will start and voice input will be stored in a stream form in a variable. "STOP"-The process will end after we click on the STOP button, and speech recognition engine will stop. The spoken word will be displayed on the text window, and the waveform will be drawn in the waveviewer.



Process of speech recognition

Basically, the microphone will convert the voice to an analog signal. This will be processed by the sound card present in the computer, which will take the signal to the digital stage. The words we speak are transformed into digital forms of the basic speech elements (phonemes).

Later, spoken word is compared to the digital "dictionary" which is already present in computer memory. When it finds an optimal match based on the digital form it will display that word on the screen. This is the basic process which is followed by all speech recognition system software. [2]

Waveviewer

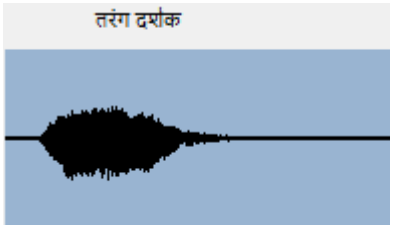


Fig 4. Figure showing waveviewer

When working with the sound files the very common requirement is to display and analyze the sound wave within the form element. Open source NAudio dll will used to load sound files. Based on the extracted sound samples custom drawings will be performed. After the sound file is loaded we need to calculate bytesPerSample variable that depends on BitsPerSample value and the number of channels. [3]

NAudio is an open source audio library for .NET, which supports audio recording, playback and sample manipulation as well as reading and writing various audio file formats.

Creating Database

If user wishes to improvise the database user can add key and value in the form. The condition here is that the current language which is selected the value should be in the same language, after this a path to resource file of that language is given, this is done in event called updateresourcefile().

Name	Value
AProjectBy	सुप्रिया पाटील , रुचिका रोकडे ,अर्चना राजगिरी आणि सेरेना साल्ढान्हा यांच्या सौजन्याने
Cancel	रद्द करा
Controls	नियंत्रणे
Copy	कॉपी
DateAndTime	तारीख आणि वेळ
Edit	संपादन
Exit	बाहेर पडा
File	फाइल
Font	फॉण्ट निवडा
Help	मदत
Key	कुंजी
Language	भाषा
NoLogFound	नोंदी सापडल्या नाहीत
Ok	ठीक
Print	प्रिंट करा
Save	जतन
SaveAs	जतन करा
SelectAll	सर्व निवडा
Sentence	वाक्य
Setting	सेटिंग
ShowLog	लॉग दाखवा
SpeechToText	भाषण ते मजकूर
Start	सुरवात

Fig 5. Database of marathi language

Setting

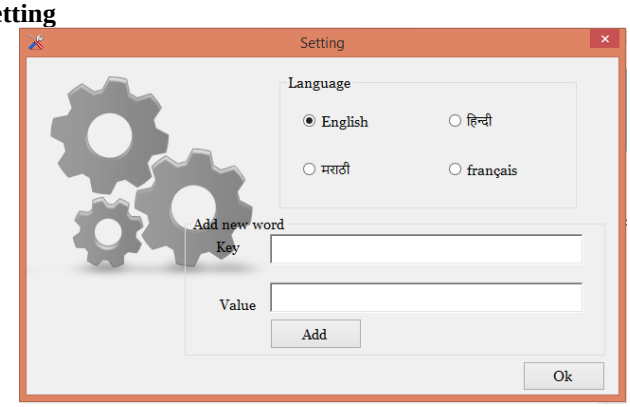


Fig 6. UI of setting form

A special provision is given for changing the language in between the process without closing that application, which is done in SETTING. The advantage of setting is to add a word in the dictionary. The process of changing the language goes as follows-

If any of the languages from Marathi, Hindi, English and French is chosen other than the current language an event called Languagechanged() is invoked so accordingly the 2nd UI again will be modified to the selected language. And culture will be provided to speech engine accordingly.

Log

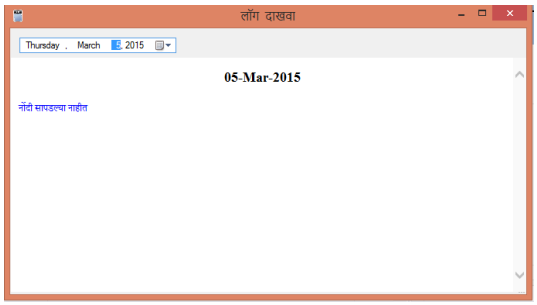


Fig 7. UI of log form

When the user views previous displayed speech, the entire log gets displayed in the selected language. The log will first receive culture (language) from language resource file. The user needs to select the date and then entire log of that date will be displayed. The log will be saved according to the system time using "DateTimePicker.Value" property which is available in the C# libraries. "DateTimePicker.Value" property gets or sets the date/time value assigned to the control.

The speech content is stored in the String variable and each sentence in the speech is retrieved as a substring and whole speech is stored in the memory in the text format sentence by sentence with the file name set according to the system date.

Applications

Prospects of SR are very high in high-performance fighter aircraft, battle management, healthcare, telephony, air traffic controllers, and other real life application domains. Devices like mobile phones use Speech-to-Text (STT) conversion and Speech Recognition (SR) in many of its applications. These applications are like writing text messages by speech input, e-mail documentation, mobile games commands, music player songs selection etc. [5]

Technical Challenges and Future Research

The key factor in designing such system is the target audience, for example, physically handicapped people should be able to wear a headset and have their hands and eyes free in order to operate the system.

There are several scenarios where speech recognition is either being researched, developed, delivered or seriously discussed are computer and video games, wearable computers, precision surgery, domestic applications etc.

There are several challenges that system will have to deal with in the upcoming future. First, improve the overall robustness of the system for facilitating implementation in real life applications involving telephones, mobile phones and computer systems. Second, the system must reject irrelevant speech that does not contain valid words or commands. And finally, the voice systems must be viable on low-cost processors. This will enable the technology to be applied in almost any product. [1]

This can also be integrated with Mobile Phones. Using Dictionary as the database the speech processing unit can also be embedded into presently available mobile phones. The dictionary present in the mobile phones can serve as the database for the speech processing unit. This could be proved as an efficient and excellent translator as the number of words in the dictionary is usually inexhaustible. [4]

Test Cases

Test Case	Case Description	Action	Expected Result	Actual Result	Pass/ Fail
Working of microphone for speech recognition	User will first speak out some words and check whether spoken words are recorded correctly.	User speak some words for given time period	Whatever user speaks should be recorded correctly.	User voice is recorded correctly	Pass
Proper machine speed	Speech recognition speed depends upon speed and better functioning of processor.	Check for users system configuration	Configuration should be as per requirement of Speech Recognition System	Few machines have proper configuration and few one may not have.	Fail
Error rate increase as the vocabulary size grows	User will provide unique single isolated word for recording/recognition	User provides input in continuous speech.	User should provide single isolated words only.	System takes more time and ultimately provide incorrect output	Pass
Reducing Computation and Increasing accuracy	Figuring out what words were spoken, this should be an easy task.	User provides input in different accents.	User should provide input in standard accent.	System take long time to recognize and accuracy decrease	Fail
Working in noisy environment	User will speak out in noisy environment	User speaks	Will not work properly	Not able to recognize any input from user	Pass
Change in voice due to user illness	User having changed voice due to physical illness like cold, and trying to use system	User speaks and expecting output.	Character should be recognized.	Character recognition failed.	Pass
Terminating speech synthesis system.	.user wants to terminate application as per his/her choice.	User click on close button on the system	Speech recognition system should turn off	Speech recognition system gets terminated.	Pass

Results

	ENGLISH	MARATHI	HINDI	FRENCH
One lettered word	75%	75%	75%	70%
Two lettered word	70%	71%	71%	70%
Three lettered word	85%	80%	80%	65%
Four lettered word	80%	80%	80%	70%
Five lettered word	90%	75%	75%	75%
Six lettered word	90%	80%	80%	70%
More than six lettered word	92%	85%	85%	75%
Words starting with same sounds	90%	67.5%	67.5%	--
Words ending with same sounds	75%	95%	95%	--

Fig 8. Results

Conclusion

This paper proposes a scheme for multilingual speech to text for continuous speech in offline mode. We are working with four languages, and database for each language can be modified and improvised by the user according to the application.

References

[1] Kamlesh Sharma, Dr. T.V. Prasad, Dr. S. V. A. V. Prasad, "Hindi Speech Enabled Windows Application Using Microsoft SAPI," International Journal Of Computer Engineering & Technology (IJCET), Volume 4, Issue 2, ISSN : 0976 – 6375(online), 0976 – 6367(print), March – April 2013, pp. 425-436.

[2] Parwinder Pal Singh, Er. Bhupinder Singh, "Speech Recognition as Emerging Revolutionary Technology," International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 2, Issue 10, ISSN : 2277 128X, October 2012, pp. 410-413.

[3] dotNet Geek. (2013, January 11). Creating custom sound wave analysis control [Online]. Available: http://www.dotnet-geek.co.uk.

[4] Umeaz Kheradia, Abha Kondwilkar, "Speech To Speech Language Translator," International Journal of Scientific and Research Publications (IJSRP), Volume 2, Issue 12, ISSN: 2250-3153, December 2012.

[5] Nishant Allawadi, Parteek Kumar, "Speech to Text," International Organization of Scientific Research (IOSR), ISSN 0970-647X, Volume 36, Issue 2, May 2012.

[6] Shally Gujral, Monika Tuteja, Baljit Kaur, "Various Issues in Computerized Speech Recognition Systems," International Journal of Engineering Research and General Science, ISSN 2091-2730, Volume 2, Issue 4, June-July 2014

[7] http://www.msdn.com.