A REVIEW ON HUMAN COMPUTER INTERACTION USING
VOICE RECOGNITION TECHNOLOGY
Priyanka Purohit 1, Aditi Remje 2, Darshan Shah 3, Prof. Anuja Ajay 4, Prof. Manisha Tiwari 5

Computer Engineering Department
SVKM’s NMIMS Mukesh Patel School of Technology Management and Engineering,
Mumbai, India

ABSTRACT:
The Speech is most important & central mode of communication among human beings. Speech recognition technology is among the fast-growing engineering technologies. The interactive voice applications are becoming more popular nowadays in the market. If voice recognition system is being used, there's no need of using mouse, keyboard or any other input devices, because it allows us to input voice commands to any application. Voice processing can be of two forms, voice identification and voice verification. Nearly 20% people of the world are suffering from various disabilities; many of them are visually impaired or unable to use their hands effectively. The voice recognition system in those particular cases provides an important help to them, so that they can use voice input to share information with people. A speech recognition technology can be used in emergency situation to secure cities. For the emergency detection in general CCTVs (closed circuit television) environment of our daily life, the monitoring by only images through CCTVs information occurs some problems especially in emergency state. Therefore voice recognition can be used for detecting emergency state dynamically through CCTVs as well as resolving some problem. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition.

Keywords—Speech recognition, Human computer interaction, Smart City.

[1] INTRODUCTION
Speech is the most natural and efficient form of exchanging information among humans. Since 1960s computer scientists have been researching ways and means to make computers able
to record, interpret and understand human speech. In computer-aided learning, speech recognition is the translation of spoken words into text. It is also known as automatic speech recognition, ASR, computer speech recognition, speech to text, or just STT.

Speech recognition is the command of human speech to identify and understand and react accordingly. It is based on the voice as the method for identification, it allows the machine to automatically identify and understand human spoken language through speech signal processing and pattern recognition. The speech recognition technology allows the machine to convert the voice signal into the suitable text or command by identifying, analyzing and understanding. Speech recognition technology is gradually becoming a key technology in the computer information processing technology, with the rapid development of computer hardware, software and information technology.

A speaker-dependent system necessitates the user to record an example of the word, sentence or phrase before its being recognized by the system i.e. the user trains the system. Some speaker-dependent systems require only that the user record a subset of system lexicon to make the entire word stock recognizable. A speaker-independent system does not require any such recording prior to system use. It has been developed to operate for any speaker of a particular type. A speaker adaptive system is developed to adapt its operation to the features of new speaker.

Over the years the Speech Recognition Systems have come a long way the process has ensured its presence due to the well-established need of voice operated systems. However, there is a lot to be proficient. Most of research done so far is attributed to the fact that speech is a very subjective phenomenon. The general known problems are Speaker Variation, Background Noise and Continuous Character of Speech. Perhaps the most evident source of performance degradation in speech recognition is Noise. Noise can be classified as either environmental i.e. traffic, rain, other people talking or speaker included i.e. coughing, sneezing, swallowing, breathing, chewing, etc.

[2] SPEECH RECOGNITION TECHNIQUES

The main objective of speech recognition is for a machine to be able to "hear understand" and "act upon" the information provided through voice input. The aim of automatic speaker recognition is to analyze, extract, characterize and recognize information about the speaker identity. The speaker recognition system works in three stages:

1. Analysis
2. Feature extraction
3. Modeling

2.1 Speech analysis technique

Speaker identity can be shown by different type of information that is present in speech data. This incorporates speaker specific information due to vocal tract, excitation source and behavior feature. The details about the behavior feature are also embedded in signal and that can
be used for speaker recognition. This stage deals with suitable frame size for segmenting speech signal for further analysis and extracting.

2.2. Feature Extraction Technique

The speech feature extraction in the process of placing in groups or classes is about decreasing the dimensionality of the input vector while maintaining the discriminating power of the signal. From basic formation of speaker identification and verification system, we know that the number of training and test vector needed for the classification problem grows with the dimension of the given input; therefore we need feature extraction of speech signal.

2.3. Modeling

The aim of modeling technique is to create speaker models using speaker specific feature vector. This technique is divided into two parts; speaker recognition and speaker identification. The speaker identification technique identifies by itself, who is speaking on basis of individual information integrated in speech signal.

[3] SPEECH RECOGNITION TECHNOLOGIES

3.1 SAPI

The speech Application Programming Interface is an API developed by Microsoft to allow the use of speech recognition and speech synthesis within windows applications. Until now, a number of versions of the API have been released, which have shipped either as part of a speech SDK, or as part of the Windows OS itself. Applications that use SAPI include Microsoft Office, Microsoft Agent and Microsoft Speech Server.

SAPI has seven main components:

Voice Command: It is a high-level interface that provides command and control speech recognition for applications. It allows a developer to create a Voice Command based menu that contains voice commands, such as "new file" or "send mail to someone@anywhere.net" that a user speaks into a microphone or other audio device. The user can control the computer without needing a keyboard or mouse.

Voice Dictation: In any application, that support speech recognition, voice dictation allows the user to dictate to that application. An invisible or virtual edit box is used to receive the text that the user dictates and displays the text in an application window. Voice Dictation allows text formatting such as capitalization, translation of punctuation words into punctuation symbols, built-in glossary entries, and correction of the last word spoken or a selected word. Applications which uses this component of SAPI, arrange speech by topics that use different language styles. Topics include e-mail speech, formal writing, or programming speech. Voice Dictation keeps the details for each topic on your hard drive.
Voice Text: It converts text into speech that is played on computer speakers or sent over a telephone line. The played speech has several different modes each with a different voice.

Voice Telephony: Voice Telephony uses telephony controls that are similar to Windows controls. Windows controls incorporate buttons, list boxes, sliders and other objects that can be manipulated by a mouse or keyboard. Telephony controls are codes that identify spoken responses such as Yes or No, your phone number, the date, and the time. Telephony controls create a conversation between the user and the computer. For example, a user calls a vendor to order an item. Then the user answers several questions by speaking into the telephone receiver. The telephony controls then recognize these responses and send them to the application that processes these responses. Telephony controls are also used for handling error conditions and variations of answers such as “January 4th” or tomorrow.

Direct Speech Recognition: This is a low-level interface which is similar to Voice Command. The main difference is Direct Speech Recognition speaks directly to the speech engine. This gives more control and speed to the application.

Direct Text To Speech: This is a low-level interface similar to Voice Text that also speaks directly to the speech engine.

Audio Objects: An Audio Object tells the speech engine where to get its audio.

3.2 Text-To-Speech (TTS)

Text-to-speech is used to create a spoken sound version of the text in a computer document, such as a help file or a Web page. TTS can enable the reading of information on computer display for the visually challenged person, or may simply be used to expand the reading of a text message. TTS is often used with voice recognition programs. There are numerous TTS products available, including Read Please 2000, Proverb Speech Unit, and Next Up Technology’s TextAloud. Lucent, Elan, and AT&T each have products called "Text-to-Speech.” Synthesizing artificial human speech from text, commonly known as text-to-speech (TTS), is an essential component in many applications such as speech-enabled devices, navigation systems, and accessibility for the visually-impaired. Fundamentally, it allows human-technology interaction without requiring visual interfaces. Modern TTS systems are based on complex, multistage processing, which rely on hand-engineered features and heuristics. Due to this complexity, developing new TTS systems can be very labour intensive and difficult.

3.3 Speech-To-Text (STT)

Speech-to-text software is a type of software that takes audio content and put it into written words in a word processor or other display destination. This type of speech recognition software is extremely valuable to anyone who needs to generate a lot of written content without a lot of manual typing. It is also useful for people with disabilities that make it difficult for them to use a keyboard. All speech-to-text systems rely on at least two models: an acoustic model and a language model. In addition large vocabulary systems use a pronunciation model. It is necessary to understand that there is no such thing as a universal speech recognizer. To get the
best transcription quality, all of these models can be specialized for a given language, dialect, application domain, type of speech, and communication channel. Similar to any other pattern recognition technology, speech recognition cannot be error free. The speech transcript accuracy is highly dependent on the speaker, the style of speech and the environmental conditions. Speech recognition is a process more hard than what people commonly think, even for a human being. Humans are used for understanding the speech, not to put it in written form, and only speech that is well formulated can be transcribed without ambiguity. From user's frame of mind, a speech-to-text system can be classified based on its use: command and control, dialog system, text dictation, audio document transcription, etc. Each use has particular requirements in terms of latency, memory constraints, vocabulary size, and adaptive features.

[4] ARCHITECTURE

Fig 1: Schematic structure of the basic steps of a speech recognition system

The proposed architecture (Fig 1) consists of five main components which are designed to work together. Modules Language model” and “Dialogue Manager” can be used and integrated into other independent systems. Module “Preprocessing of the incoming signal includes filters to clear the noise from the received signals or due to imperfections in the sensors. After this the cleared analogue signal is being converted into a digital one by means of an analogue/digital converter. The Speech recognition module is performed by the formation of many characteristic features of object recognition by input digital data flow. Then a description of classes and data necessary to perform the classification (de-coding) in the Classification” block is included. This block is connected to the Speech Database, through which training, self-organization and recognizing words. After this the speech is converted into text.
Understanding natural language recognizes the importance of words and phrases in the context they are used. If the word is not recognized, then it returns to “Classification” block and is recorded in the database as associated with keyword depending on its context. Interpretation and response to those carried out in the next module of the “Automatic generation of speech. The tree of conclusions is determined in the Speech interpreter block. Then, the information which must be returned as a response to the user of the system is determined in the Response generation block. The Acoustic model block generates the phonetic representation of the speech signal response of the system. In developing the software, it will be selected and modified some existing methods of acoustic modelling, which best will match the characteristics of the Bulgarian language used by children. The “Dialogue manager module controls Automatic speech generation” operations.

The most appropriate response based on information submitted to the block appears here. Thus, creating a prerequisite for conducting a real dialogue and retain its history. “Dialog manager” provides solution to two specific problems: (1) providing a logical overall structure for communication that extends beyond the single turn, (2) correctly manage mixed initiative communication, allowing users to guide interaction as per their goals while allowing the system to guide interaction towards successful completion.

Organization of Speech database will be similar to the one used before. Separate words and short sentences covering all the features of the phoneme in Bulgarian will be kept. Recordings of the words will be made by a desktop microphone to eliminate the need of expensive and specialized equipment. Data set will be divided into sets for training, development and evaluation. The record of the words will be made by English children in different age groups from 2 - 7 years of age and without speech defects. Each child will record 80-100 words, as their number will be determined later. The data base formed in this way will be used for training, testing and evaluation.

[5] APPLICATIONS

Various applications of speech recognition domain have been discussed in the following table: -

<table>
<thead>
<tr>
<th>Problem domain</th>
<th>Application</th>
<th>Pattern Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech/Telephone/Communication</td>
<td>Telephone directory enquiry without operator assistance</td>
<td>Spoken Words</td>
</tr>
</tbody>
</table>
Secure Smart City | Use of speech recognition to make the smart city more secure by monitoring the appropriate footage via CCTV to distinguish amongst the emergency and non-emergency situations to take the necessary decisions. | Spoken words monitored via CCTV
---|---|---
Education Sector | Teaching students of foreign languages to pronounce vocabulary correctly. Teaching overseas students to pronounce English correctly. | Spoken Words
Outside education sector | Computer and video games, Gambling, Precision surgery. | Spoken words
Domestic Sector | Oven, refrigerators, dishwashers and washing machines. | Spoken words.
Military Sector | High performance fighter aircraft, Helicopters, Battle management, Training air traffic controller. | Spoken words.
Artificial Intelligence Sector | Robotics | Spoken words.
Medical sector | Health care, Medical Transcriptions | Spoken words.
General | Automated transcription, Telematics, Air traffic control, Multimodal interacting, court reporting, Grocery shops | Spoken words.
Physically Handicapped | Useful to people with limited mobility in their arms and hands. | Spoken words.

Table 1: Applications of speech recognition

[6] FUTURE WORK

Accuracy will become better and better. Dictation speech recognition will gradually become accepted. Powerful use will be made of “intelligent systems” which will attempt to guess what the speaker meant to say, rather than what was actually said, as people often miss-
speak and make unintentional mistakes. Microphone and sound systems will be designed to adapt more quickly to changing background noise levels, different environments, with better recognition of extraneous material to be discarded. Multiple languages can be included according to region, can also be used to translate non-English languages into English and can perform functions to add or delete files and folders.

[7] CONCLUSION

Speech is the primary, and the most convenient means of communication between people. Speech recognition is not just for the disabled anymore. Properly implementing this technology on a wide scale basis can increase productivity; reduce costs involved in the prevention of repetitive stress related injuries and gives computer access to the disabled individual. We’re at a turning point where voice and natural language understanding are suddenly at the forefront. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Also the use of speech recognition technology in smart city. It is for the recognition of emergency or non-emergency sound to detect the emergency situation in smart city. The biggest advantage of the architecture is its universality. This is because the Dialog Manager and Language Model are not dependent on the task being executed. Since humans do a daily activity of speech recognition, it is one of the most consolidating areas of machine intelligence. Speech recognition has attracted scientists as an important discipline and has created a technological impact on society and is expected to flourish further in this area of human machine interaction.

REFERENCES


Priyanka Purohit, Aditi Remje, Darshan Shah, Prof. Anuja Ajay and Prof. Manisha Tiwari