A Mini Project report on

SPEECH RECOGNITION

A Dissertation Submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

In

Computer Science and Engineering

By

N. SAI DEEKSHITHA (18H51A05D4)
P . DHRUVA RAJU (18H51A05D5)
P. GOPIKA (18H51A05D6)

Under the esteemed guidance of

Mr.R.Krishna Rao

(Assistant Prof. CSE)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(AUTONOMOUS)

(NAAC Accredited with A+" Grade & NBA Accredited) (Approved by AICTE, Permanently Affiliated to JNTU Hyderabad)

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD-501401

2020-2021

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the mini project - 2 report entitled **" SPEECH RECOGNITION"** being submitted by **N.SAI DEEKSHITHA** ***(18H51A05D4)*** **,P.DHRUVARAJU** ***(18H51A05D5),*** **P.GOPIKA** ***(18H51A05D6),*** in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr. R Krishna Rao
Assistant Prof. CSE
Dept. of CSE
Project Guide

Dr. K Vijaya Kumar
Professor and HOD
Dept. of CSE

Submitted viva voice for Examination held on

External Examiner

ACKNOWLEDGEMENT

With great pleasure I want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

I am grateful to thank **Mr.R.Krishna Rao** ,Assistant. Prof. CSE, Dept. of Computer Science and Engineering for his valuable suggestions and guidance during the execution of this project work.

I would like to thank **Dr. K. Vijaya Kumar**, Head of the Department of ComputerScience and Engineering, for his moral support throughout the period of my study inCMRCET.

I am highly indebted to Major **Dr. V.A. Narayana**, Principal CMRCET forgiving permission to carry out this project in a successful and fruitful way.

I would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering for their co-operation.

Finally, I express my sincere thanks to **Mr. Ch. Gopal Reddy**, Secretary, CMRGroup of Institutions, for his continuous care. I sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

N SAIDEEKSHITHA (18H51A05D4)
P.DHRUVA RAJU (18H51A05D5)
P.GOPIKA (18H51A05D6)

DECLARATION

We hereby declare that results embodied in this Report of Mini Projectn2 on **"SPEECH RECOGNITION"** are from work carried out by using partial fulfillment of the requirements for the award of B.Tech degree. We have not submitted this report to any other university/institute for the award of any other degree.

SIGNATURE

N.SAIDEEKSHITHA (18H51A05D4)
P.DHRUVA RAJU (18H51A05D5)
P.GOPIKA (18H51A05D6)

ABSTRACT

PROBLEM STATEMENT:

Technology that enables human speech to be converted automatically into the text.

ABSTRACT:

Speech is the most common means of communication and the majority of the population in the world relies on speech to communicate with one another. Speech recognitionsystem basically translates spoken languages into text. There are various real-life examples of speech recognition systems. Suppose we are building a model and instead of a written approach we want our system to respond to speech, it becomes fairly difficult and requires a lot of data to be processed. A speech recognition system overcomes this barrier by translating speech to text. In this, we will go through the speech recognition module in python.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER-1: INTRODUCTION

INTRODUCTION:

Speech recognition is the process by which a computer (or other type of machine) identifies spoken words. Basically, it means talking to your computer, AND having itcorrectly recognize what you are saying.

The following definitions are the basics needed for understanding speech recognition technology.

Utterance:

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences.

Speaker Dependance:

Speaker dependent systems are designed around a specific speaker. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy.

Vocabularies:

Vocabularies (or dictionaries) are lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. They can be as long as a sentence or two. Smaller vocabularies can have as few as 1 or 2 recognized utterances (e.g."Wake Up"), while very large vocabularies can have a hundred thousand or more!

Accuract:

The ability of a recognizer can be examined by measuring its accuracy - or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more! The acceptable accuracy of a system really depends on the application.

Training:

Some speech recognizers have the ability to adapt to a speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy.

Training can also be used by speakers that have difficulty speaking, or pronouncing certain words. As long as the speaker can consistently repeat an utterance, ASR systems with training should be able to adapt.

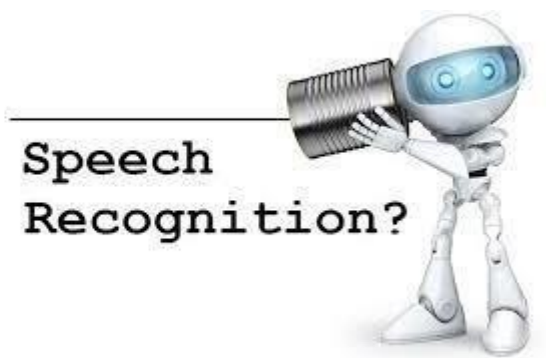

Fig 1: Speech Recognition

Today Voice recognition software allows us to tell our devices what to do by just talking to them. Now havingto use a keyboard, a mouse, or a screen dramatically changes the way we experience technology. We saw therise of voice recognition technology on our phones. Due to the many uses of voice recognition software making our lives easier, in just a few years, we brought it into our homes. Today, businesses in a wide arrayof sectors are tapping into it to make our lives better. We can now use voice recognition-based software to make purchases, check the weather, send emails, search for information on the internet, and define new waysto interact with machines.

Speech recognition is the capability of an electronic device to understand spoken words. A microphone records a person's voice and the hardware converts the signal from analog sound waves to digital audio. The audio data is then processed by software, which interprets the sound as individual words. A common type of speech recognition is "speech-to-text" or "dictation" software, such as Dragon Naturally Speaking, which outputs text as you speak..

OBJECTIVE:

Most of the tech giants have been very invested in the development of voice assistants in the last decade. This is how Google Assistant, Microsoft's Cortana, and Apple's Siri, have become household names. According to Microsoft's 2019 Voice Report, 69% of their respondents have used a digital assistant. Most of them have interacted with them on their phones (72%).

LITERATURE STUDY:

As indicated by Alapetite et al. [3], a change initiative concerning speech recognition deals with implementation of novel technology for which user satisfaction and objective functionality are strongly linked. This means that users who do not get involved with such software might not acquire the necessary skills to fully benefit from a speech recognition solution. Hence, user experience was brought into central focus during the preliminary considerations when planning implementation in our department. In Kotter's change management framework [6], it is argued that implementation of change is much facilitated by creating "short-term wins" for users. In our context, "short-term wins" are understood as the immediate experience of a versatile, adaptive software, which can be used productively (from the user perspective) after a relatively short period of time. When implementing a novel technology maximisation of the initial performance has generally been identified as beneficial [7]. In our context, this translates into a high a priori speech recognition rate, so that physicians will perceive the software to be productive. A characteristic of our department is the very specific purpose of the department with a focus on elite sports and a close link to sports sciences. For a high a priori speech recognition rate, creation of an adapted dictionary for our specific purpose in addition to the general dictionary of the discipline (internal medicine) appeared necessary to account for the specific vocabulary used for documentation..

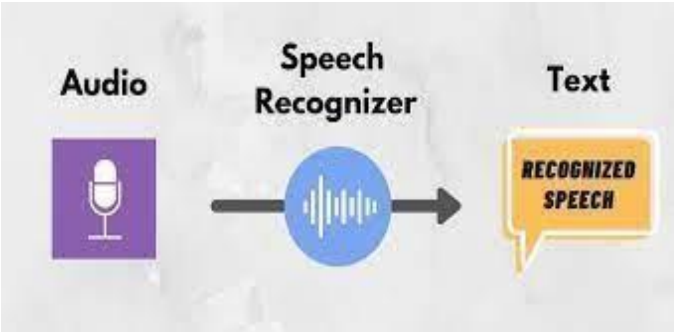

Fig 2:speech Recognition

EXISTING SOLUTIONS:

Speech recognition technology has grown leaps and bounds in the early 21st century and has literally come home to roost.

Look around you. There could be a handful of devices at your disposal at this very moment.

Let's look at a few of the leading options.

Apple's Siri:

Apple's Siri emerged as the first popular voice assistant after its debut in 2011. Since then, it has been integrated on all iPhones, iPads, the Apple Watch, the HomePod, Mac computers, and Apple TV.

Siri is even used as the key user interface in Apple's CarPlay infotainment system, as well as the wireless AirPod earbuds, and the HomePod Mini.

Siri is with you everywhere you go; on the road, in your home, and for some, literally on your body. This gave Apple a huge advantage in terms of early adoption.

Naturally, being the earliest quite often means receiving most of the flack for functionality that might not work as expected.

Although Apple had a big head start with Siri, many users expressed frustration at its seeming inability to properly understand and interpret voice commands.

If you asked Siri to send a text message or make a call on your behalf, it could easily do so. However, when it came to interacting with third-party apps, Siri was a little less robust compared to its competitors.

But today, an iPhone user can say, "Hey Siri, I'd like a ride to the airport" or "Hey Siri, order me a car," and Siri will open whatever ride service app you have on your phone and book the trip.

Focusing on the system's ability to handle follow-up questions, language translation, and revamping Siri's voice to something more human-esque is helping to iron out the voice assistant's user experience.

As of 2021, Apple hovers over its competitors in terms of availability by country and thus inSiri's understanding of foreign accents. Siri is available in more than 30 countries and 21 languages – and, in some cases, several different dialects.

Amazon Alexa:

Amazon announced Alexa and the Echo to the world in 2014, kicking off the age of the smart speaker.

Alexa is now housed inside the Echo, the Echo Show (a voice-controlled tablet), the Echo Spot (a voice-controlled alarm clock), and the Echo Buds headphones (Amazon's version of Apple's AirPods).

In contrast to Apple, Amazon has always believed the voice assistant with the most "skills", (its term for voice apps on its Echo assistant devices) "will gain a loyal following, even if it sometimes makes mistakes and takes more effort to use".

Although some users pegged Alexa's word recognition rate as being a shade behind other voice platforms, the good news is that Alexa adapts to your voice over time, offsetting any issues it may have with your particular accent or dialect.

Speaking of skills, Amazon's Alexa Skills Kit (ASK) is perhaps what has propelled Alexa forward as a bonafide platform. ASK allows third-party developers to create apps and tap into the power of Alexa without ever needing native support.

Alexa was ahead of the curve with its integration with smart home devices such as cameras, door locks, entertainment systems, lighting, and thermostats.

Ultimately, giving users absolute control of their home whether they're cozying up on their couch or on-the-go. With Amazon's Smart Home Skill API, you can enable customers to controltheir connected devices from tens of millions of Alexa-enabled endpoints.

When you ask Siri to add something to your shopping list, she adds it to your shopping list – without buying it for you. Alexa however goes a step further.

If you ask Alexa to re-order your garbage bags, she'll just go through Amazon and order them. In fact, you can order millions of products off Amazon without ever lifting a finger; a natural and unique ability that Alexa has over its competitors.

Google Assistant:

How many of us have said or heard "let me Google that for you"? Almost everyone, it seems. It only makes sense then, that Google Assistant prevails when it comes to answering (and understanding) all questions its users may have.

From asking for a phrase to be translated into another language, to converting the number of sticks of butter in one cup, Google Assistant not only answers correctly, but also gives some additional context and cites a source website for the information.

Given that it's backed by Google's powerful search technology, perhaps it's an unsurprising caveat.

Though Amazon's Alexa was released (through the introduction of Echo) two years earlier than Google Home, Google has made great strides in catching up with Alexa in a very short time. Google Home was released in late 2016, and within a year, had already established itself as the most meaningful opponent to Alexa.

In 2017, Google boasted a 95% word accuracy rate for U.S. English, the highest out of all the voice-assistants currently out there. This translates to a 4.9%-word error rate – making Google the first of the

Fig: Existing solutions

PROPOSED SYSYTEM:

ASR systems can be grouped into one of four types: feature based, segmental, template based, or statistical. In a feature-based system, specific features that are known to be associated with phonetic properties are extracted from the speech signal. A feature-based ASR system extracts a large number of these features and combines the evidence over time to arrive at hypothesized phonemes, which can be further combined to arrive at hypothesized words. Although intuitive and in many ways well motivated, feature-based approaches suffer from a “multiplication of errors” effect: Small errors in extracting each of the features accumulate, and even low error rates in feature extraction can result in high errors at the word level.

A segment-based ASR system operates in three steps: First, the speech is analyzed to determine the likely locations of phoneme boundaries; second, the segments that result from this analysis are classified based on features taken from throughout the segment; and third, word-level results are obtained from the hypothesized phoneme segments. Segment-based systems have demonstrated performance that can be competitive with other types of systems, such as HMMs, especially on difficult tasks such as alphabet recognition. One of the difficulties in implementing a segment-based ASR system is in reliably determining the locations of phoneme boundaries; methods that work well for certain types of phonetic boundaries may not work well for other types, and in order to avoid “missing” a phoneme, many more segments are hypothesized than are likely to exist in the signal.

A template-based ASR system represents each phoneme or word using individual templates, or sets of time frames of features from the speech signal that are obtained from representative examples of each word or phoneme. The input speech is then matched with each of the templates, with time differences between the input and the template normalized using the dynamic programming algorithm called dynamic time warping. The template with the best fit tothe input speech is considered the recognized word. One of the problems with the template- based approach is that it is unable to efficiently account for the wide variety of ways in which words are pronounced. A single speaker may utter the same word in a number of different ways, with variation in rhythm, relative loud-ness, or articulation effort. Selection of the most appropriate utterance from the training set can become a difficult issue. A speaker-dependent, word-level, template-based system may perform well in some cases, but it will be restricted to that speaker and that vocabulary; the variety of phonemes in different contexts prohibits the template-based approach from being applied to more general-purpose phoneme-level recognition. For speaker-independent ASR, the variety within and between speakers becomes insurmountable with template-based systems.

Finally, the most commonly used type of ASR system is now the statistical ASR system. In this case, the speech is divided into short, equally spaced frames (with a typical length of 10 msec), and the likelihood of a subphonetic or subword unit at each frame is computed based on statistics such as the average and standard deviation of the features for each unit found in the training set. These likelihoods are combined to arrive at phoneme-level or word-level results, typically by assuming mathematical independence between likelihoods at different frames in the signal. Statistical ASR systems, which generally use the HMM framework, are able to account for more variation than the template-based approach, do not require prior segmentation of the signal into phonetic regions as in the segment-based approach, and do not depend on accurate feature extraction required for the feature-based approach. However, the statistical approach has its own set of weaknesses, including a conflict between sufficiently complex classification units and a sufficient amount of training data, as well as certain overly simplified mathematical assumptions.

CHAPTER-2: DESIGN COMPONENTS

The basic principle of voice recognition involves the fact that speech or words spoken by any human being cause vibrations in air, known as sound waves. These continuous or analog waves are digitized and processed and then decoded to appropriate words and then appropriate sentences.

voice recognition

Components of a Speech Recognition System

So what does a basic Speech Recognition System consists of?Components

of a Speech Recognition System

A speech capturing Device: It consists of a microphone, which converts the sound wave signals to electrical signals and an Analog to Digital Converter which samples and digitizes the analog signals to obtain the discrete data that the computer can understand.
A Digital Signal Module or a Processor: It performs processing on the raw speech signal like frequency domain conversion, restoring only the required information etc.
Preprocessed signal storage: The preprocessed speech is stored in the memory to carry out further task of speech recognition.
Reference Speech patterns: The computer or the system consists of predefined speech patterns or templates already stored in the memory, to be used as the reference for matching.
Pattern matching algorithm: The unknown speech signal is compared with the reference speech pattern to determine the actual words or the pattern of words.
Working of the System

Now let us see how the whole system actually works.Working

of the System

A speech can be seen as an acoustic waveform, i.e. signal carrying message information. A normal human being with the limited rate of motion of his/her articulators (speech organs) can produce speech at a average rate of 10 sounds per second. The average information rate is about 50-60 bits/second. It means actually only 50 bits/second of information is required in the speech signal. This acoustic waveform is converted to analog electrical signals by the microphone. The Analog to Digital converter converts this analog signal to digital samples by taking precise measurements of the wave at discrete intervals.

The digitized signal consists of a stream of periodic signals sampled at 16000 times per second and is not suitable to carry out actual speech recognition process as the pattern cannot be easily located. To extract the actual information, the signal in time domain is converted to signal in frequency domain. This is done by the Digital Signal Processor using FFT technique. In the digital signal, the component after every 1/100th of a second is analyzed and the frequency spectrum for each such component is computed. In other words the digitized signal is segmented into small parts of frequency amplitudes

CHAPTER 3: IMPLEMENTATION OF SPEECH RECOGNITION

Speech recognition is one from the fastest growing technologies. It has a wide range of application and has been effectively deployed in mobile and embedded devices, providing functionalities including personal assistants such as Siri, Ok! Google and Cortana. Speech recognition applications include voice user interfaces such as voice dialing e.g. (call home), call routing, search, simple data entry, speech-to-text processing. The aim of this project is to design and implement a speech recognition based android application, purposed for use during emergency situations. It will allow a user to interact with it through voice commands, using preset keywords to activate it, and to navigate and seek assistance, for example, voice commands to enable calling an operator or send anautomatic text to a listed emergency contact.

WORKING:

Speech recognition software works by breaking down the audio of a speech recording into individual sounds, analyzing each sound, using algorithms to find the most probable word fit in that language, and transcribing those sounds into text.

Speech recognition software uses natural language processing (NLP) and deep learning neural networks. "NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way," according to the Algorithma blog. This means that the software breaks the speech down into bits it can interpret, converts it into a digital format, and analyzes the pieces of content.

From there, the software makes determinations based on programming and speech patterns, making hypotheses about what the user is actually saying. After determining what the users mostlikely said, the software transcribes the conversation into text.

This all sounds simple enough, but the advances in technology mean these multiple, intricate processes are happening at lightning speed. Machines can actually transcribe human speech moreaccurately, correctly, and quickly than humans can.

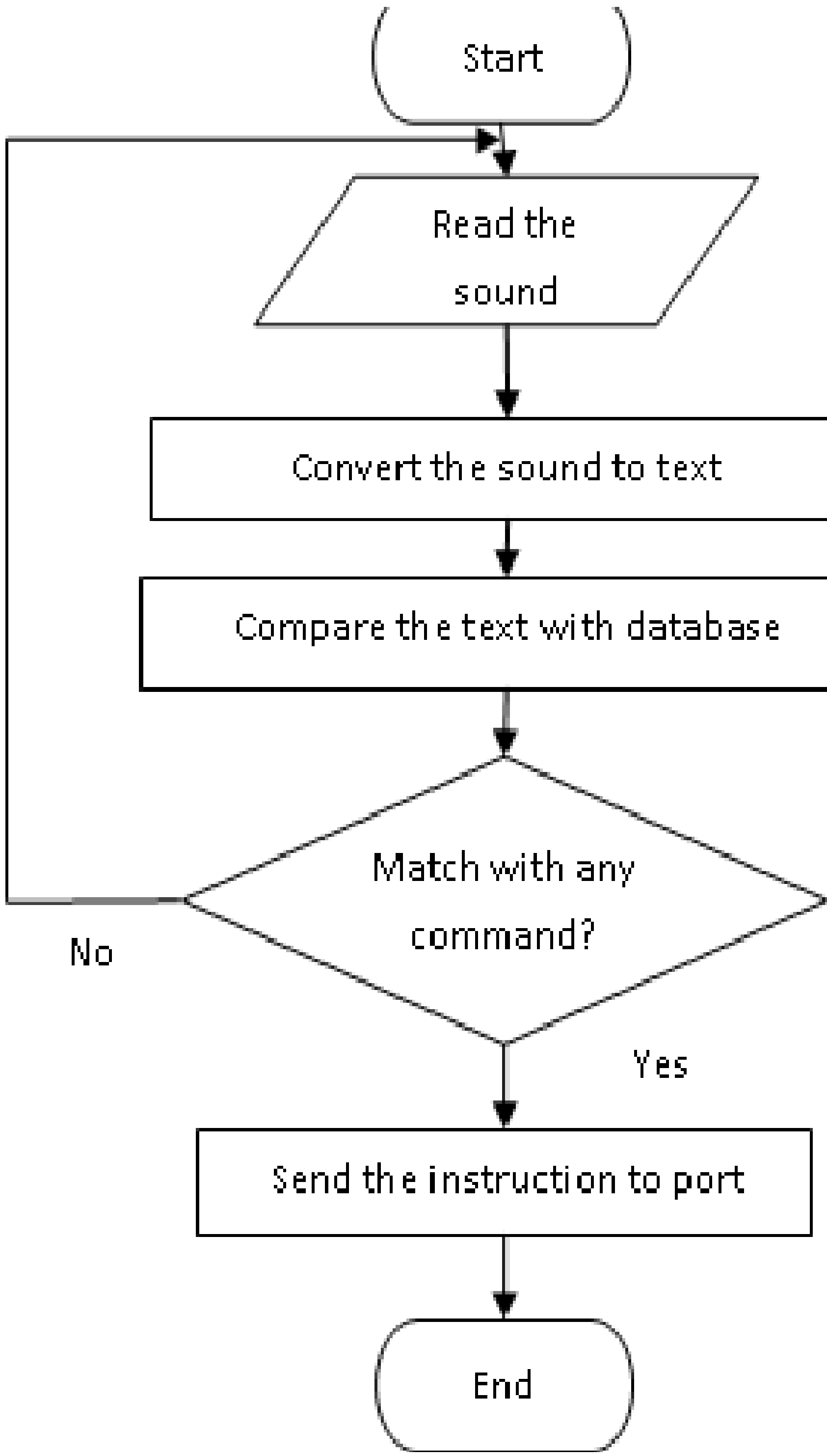

Fig : Flow Chart

CHAPTER 4:

RESULT:

Speech recognition technology allows computers to take spoken audio, interpret it and generate text from it. But how do computers understand human speech? The short answer is…the wonder of signal processing. Speech is simply a series of sound waves created by our vocal chords when they cause air to vibrate around them. These soundwaves are recorded by a microphone, and then converted into an electrical signal. The signal is then processed using advanced signal processing technologies, isolating syllables and words. Over time, the computer can learn to understand speech from experience, thanks to incredible recent advances in artificial intelligence and machine learning.But signal processing is what makes it all possible.

So, what are the benefits of speech recognition technology? Why, exactly, do we need computers to understand our speech when typing is usually faster (and quieter)? Speech is a natural interface for many programs that don’t run on computers, which are becoming more common. Here are some important ways in which speech recognition technology plays a vital role in people’s lives.

Talking to Robots: You might not think that speaking with robots is a common activity. But robots are increasingly being employed in roles once performed by humans, including in conversation and interface. For example, firms are already exploring using robots and software to perform initial job interviews. As interviews must be conversational, it’s essential that the robot can interpret what the interviewee is saying. That requires speech recognition technology.

Controlling Digital Devices: Digital personal assistants like Alexa and Google Home obviously require verbal communication between humans and computers. They are also great examples ofhow computers use machine learning to better understand your speech over time through experience. But in order to do so, speech recognition technology, enabled by signal processing, iskey.

Aiding the Visually- and Hearing-Impaired: There are many people with visual impairments who rely on screen readers and text-to-speech dictation systems. And converting audio into text can be a critical communication tool for the hearing-impaired.

Enabling Hands Free Technology: When your eyes and hands are busy, such as when you’re driving, speech is incredibly useful. Being able to communicate with Apple’s Siri or Google Maps to take you where you need to go reduces your chances of getting lost and removes the need to pull over and navigate a phone or read a map.

Why Speech Recognition Technology is a Growth Skillset: Speech recognition technology is already a part of our everyday lives, but for now is still limited to relatively simple commands. As the technology advances, researchers will be able to create more intelligent systems that understand conversational speech (remember the robot job interviewers?). One day, you will be able to talk to your computer the way you would talk to any human, and it will be able to transmit reasoned responses back to you. All this will be made possible by signal processing technologies. The number of specialists needed in this field are growing, and many companies are looking for talented people who want to be a part of it. Processing, interpreting and understanding a speech signal is the key to many powerful new technologies and methods of communication. Given current trends, speech recognition technology will be a fast-growing (and world-changing) subset of signal processing for years to come.

```
[search edureka: search youtube]
speak now
search your query
python

Process finished with exit code 0
```

CONCLUSION:

Speech synthesis has been developed steadily over the last decades and it has been incorporated into several new applications. For most applications, the intelligibility and comprehensibility of synthetic speech have reached the acceptable level. However, in prosodic, text preprocessing, and pronunciation fields there is still much work and improvements to be done to achieve more natural sounding speech. Natural speech has so many dynamic changes that perfect naturalness may be impossible to achieve. However, since the markets of speech synthesis related applications are increasing steadily, the interest for giving more efforts and funds into this research area is also increasing. Present speech synthesis systems are so complicated that one researcher can not handle the entire system. With good modularity it is possible to divide the system into several individual modules whose developing process can be done separately if the communication between the modules is made carefully.

The three basic methods used in speech synthesis have been introduced in Chapter 5. The most commonly used techniques in present systems are based on formant and concatenative synthesis. The latter one is becoming more and more popular since the methods to minimize the problems with the discontinuity effects in concatenation points are becoming more effective. The concatenative method provides more natural and individual sounding speech, but the quality with some consonants may vary considerably and the controlling of pitch and duration may be in some cases difficult, especially with longer units. However, with for example diphone methods, such as PSOLA may be used. Some other efforts for controlling of pitch and duration have been made by forexample Galanes et al. (1995). They proposed an interpolation/decimation method for resampling the speech signals. With concatenation methods the collecting and labeling of speech samples have usually been difficult and very time-consuming. Currently most of this work can be done automatically by using for example speech-recognition.

REFERENCES:

The National Center for Voice and Speech provides some basic information on preserving "Vocal Health" on their WWW site: http://www.shc.uiowa.edu/hygiene/home.html
Voice Users Mailing List: detail in Q1.4.html of the FAQ.
Typing Injury FAQ: http://www.cs.princeton.edu:80/~dwallach/tifaq/ has a range of information on Typing Injuries, avoiding them, alternatives and more.
Typing Injuries Page: http://alumni.caltech.edu/~dank/typing-archive.html has links to dozens of useful resources.
Voice Problems -- Prevention and Correction: advice on preserving your voice with specific hints for using speech recognition. ftp://ftp.csua.berkeley.edu/pub/typing-injury/voice-problems
" Talking to a PC May Be Hazard To Your Throat", by Julie Chao in the Wall Street Journal.
" Talking to Computers Has its Hazards", by Gordon Arnaut in The Globe and Mail
On the WWW

Survey of the State of the Art in Human Language Technology: Report edited by Ronald A. Cole et.al. with a section on Spoken Input Technologies. http://www.cse.ogi.edu/CSLU/HLTsurvey/ch1node2.html
Technology: General and Introductory

Some general introduction books on speech recognition technology:

Fundamentals of Speech Recognition; Lawrence Rabiner & Biing-Hwang Juang Englewood CliffsNJ: PTR Prentice Hall (Signal Processing Series), c1993, ISBN 0-13-015157-2
Speech recognition by machine; W.A. Ainsworth London: Peregrinus for the Institution of Electrical Engineers, c1988
Speech synthesis and recognition; J.N. Holmes Wokingham: Van Nostrand Reinhold, c1988
Speech Communication: Human and Machine, Douglas O'Shaughnessy; Addison Wesley series in Electrical Engineering: Digital Signal Processing, 1987.
Electronic speech recognition: techniques, technology and applications, edited by Geoff Bristow, London: Collins, 1986
Readings in Speech Recognition; edited by Alex Waibel & Kai-Fu Lee. San Mateo: Morgan Kaufmann, c1990