People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Chadli Bendjedid University
Faculty of Science and Technology
Department of Computer Science

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة الشاذلي بن جديد
كلية العلوم و التكنولوجيا
قسم الإعلام الآلي

جامعة الشاذلي بن جديد
UNIVERSITE CHADLI BENDJEDID

THESIS

Presented by

KEZIHI SOUFYANE

Submitted in partial fulfilment of the requirements for a Master's degree

Field: Computer Science

Specialty: Intelligent Computer Systems

Theme

Design and Implementation of Voice Recognition System

Sustained on: 29/06 / 2022

In front of the Jury composed of:

Quality	Name and Surname	Rank	University
President	Mrs.GASMI I	MCA	Chadli Bendjedid El-Tarf
Supervisor	**Mrs.MAKHLOUF A**	**MCB**	**Chadli Bendjedid El-Tarf**
Examiner	Mr.BENMACHIHCE A	MCA	Chadli Bendjedid El-Tarf

University Year: 2021/2022

Acknowledgment

First, I want to thank ALLAH for guiding me on the bright way of knowledge.

I also want to express my sincere and deep gratitude to all who contributed by near or far in the realization of this work, especially Dr.MAKHLOUF AMINA for her advice and valuable guidance.

I want to thank every person that helped me sustained and encouraged me for the realization of this modest work.

Thank you all

Dedication

I dedicate this modest work

For my dear father, who brought me a lot in this life, my greatest gratitude is for you, I hope this thesis can reflect the image of your efforts, dad.

For my light and my patience in life, my lovable mother, I thank you for your moral support, for all your sacrifices for me and all family members, may Allah protect you and keep you for us mom.

Also, my two dear brothers my sweetheart

sister, and to all my close friends

May ALLAH preserve you a happy life full of love and success

Summary

For the last many years, a gigantic measure of research has been done on the utilization of machine learning for speech processing applications, particularly speech recognition. In any case, in a couple of years, researchers have zeroed in on using deep learning for discourse-related applications. This new area of machine learning has yielded much better outcomes when contrasted with others in an assortment of utilizations including speech and accordingly, turned into an exceptionally appealing area of exploration while deep learning initially emerged as another area of machine learning, for speech applications.

In this research paper, we present a speech recognition system using a convolutional neural network that is able to recognize words, we note that our work is 79% accurate.

Keywords:

Natural Language Processing, Machine Learning, Deep Learning, Speech Recognition, Convolutional Neural Network.

Abstract

Au cours des dernières années, de nombreuses recherches ont été menées sur l'utilisation de l'apprentissage automatique pour les applications de traitement de la parole, en particulier la reconnaissance vocale. Quoi qu'il en soit, en quelques années, les chercheurs se sont concentrés sur l'utilisation de l'apprentissage en profondeur pour des applications liées au discours. Ce nouveau domaine de l'apprentissage automatique a donné de bien meilleurs résultats lorsqu'il est comparé à d'autres dans un assortiment d'utilisations, y compris la parole et, par conséquent, s'est transformé en un domaine d'exploration exceptionnellement attrayant, tandis que l'apprentissage en profondeur est initialement apparu comme un autre domaine de l'apprentissage automatique, pour les applications vocales.

Dans ce document de recherche, nous présentons un système de reconnaissance de la parole utilisant un réseau de neurones convolutifs capable de reconnaître des mots, nous notons que notre travail est précis à 79%.

Mots Clés : Traitement du langage naturel, apprentissage automatique, apprentissage en profondeur, reconnaissance vocale, réseau de neurones convolutifs.

ملّخص

على مدى السنوات العديدة الماضية ، تم إجراء مقياس هائل للبحث حول استخدام التعلم الآلي لتطبيقات معالجة الكلام ، وخاصة التعرف على الكلام. على أي حال ، في غضون عامين ، ركز الباحثون على استخدام التعلم العميق للتطبيقات المتعلقة بالخطاب. لقد أسفر هذا المجال الجديد من التعلم الآلي عن نتائج أفضل بكثير عند مقارنته مع الآخرين في مجموعة متنوعة من الاستخدامات بما في ذلك الكلام ، وبالتالي ، تحول إلى منطقة .استكشاف جذابة بشكل استثنائي بينما ظهر التعلم العميق في البداية كمجال آخر للتعلم الآلي ، لتطبيقات الكلام في ورقة البحث هذه ، نقدم نظام التعرف على الكلام باستخدام شبكة عصبية تلافيفية قادرة على التعرف على 79٪. الكلمات ، ونلاحظ أن عملنا دقيق بنسبة

الكلمات المفتاحية:
.معالجة اللغة الطبيعية ، التعلم الآلي ، التعلم العميق ، التعرف على الكلام ، الشبكة العصبية التلافيفية

Table of Contents

Acknowledgment II

Dedication III

Table of contents VII

List of figures IX

List of tables X

List of acronyms XI

General Introduction 2

Chapter 1: State of the Art 5

1. Introduction 5

2. Section 2 5

A. History 5

B. Definitions 6

3. Used Systems 8

4. Voice Recognition Process 10

5 Speech Recognition System 11

5.1 The Speech Signal 12

5.2 Speech Recognition Approaches 12

5.3 The Involved Process Of Speech Recognition 16

5.4 The Most Popular Feature Extraction Methods 19

5.5 The Traditional Probability Mapping And Selection Method 19

5.6 The Mathematical Fundamental Equation Of Statical Speech Recognition 20

5.7 The Speed of A Speech Recognition System Measured 21

5.8 The Word Error Rate 21

6 Speech Recognition Based On Deep Learning & Neural Network 24

6.1 Artificial Neural Network 24

6.2 Deep Neural Network 25

6.3 Multilayer Perceptron 26

6.4 Recurrent Neural Network 27
6.5 Convolutional Neural Network 28
7. Application Domains.....31
8. Advantages And Disadvantages 33
9. Challenges And Future Approaches.....34
10. Conclusion.....35
Chapter 2: Conceptual Study.....37
1. Introduction 37
2. System Design.....37
3. Presenting Dataset 39
4.Model Architecture 42
4.1 CNN Model 42
4.2 Model Configuration And Number Of Parameters 43
5.Conclusion.....45
Chapter 3: Implementation and obtained results 47
1. Introduction 47
2.Representation Of The Development Tools 47
2.1. Physical Environment 47
2.2. Software And Libraries Used In The Implementation 48
3 Discussion And Comparison Of The Obtained Results 50
3.1 Discussion Of The Obtained Results.....50
3.2 Comparison Of The Obtained Results.....54
4 Representing Our System Interface.....56
5 Test 58
6 Conclusion.....60
General Conclusion And Perspectives 62
References 63
Summary ..... .

List of figures

Figure 1. Voice recognition process 10
Figure 2. Overview of a speech recognition system 11
Figure 3. Acoustic-phonetic speech recognition system 13
Figure 4. Block diagram of patter recognition speech recognizer 14
Figure 5. Steps involved in speech recognition 16
Figure 6. Chart is showing the reduction of SR over time, attributed to deep learning 22
Figure 7. Illustration of a possible neural network 25
Figure 8. Multilayer perceptron 26
Figure 9. Architecture of CNN properties 28
Figure 10. Architecture of CNN layers 29
Figure 11 CNN activation function 30
Figure 12 Application Domains 31
Figure 13 Our system architecture 38
Figure 14 CNN model architecture 42
Figure 15 The configuration of our model 43
Figure 16 Chart of accuracy (training) and validation accuracy (test) [15 epochs] 50
Figure 17 Chart of loss(training) and validation loss(test) [15 epochs] 50
Figure 18 Chart of accuracy (training) and validation accuracy (test) [35 epochs] 51
Figure 19 Chart of loss(training) and validation loss(test) [35 epochs] 19
Figure 20 Chart of accuracy (training) and validation accuracy (test) [55 epochs] 52
Figure 21 Chart of loss(training) and validation loss (test) [55 epochs] 52
Figure 22 Chart of accuracy (training) and validation accuracy (test) [85 epochs] 53
Figure 23 Chart of loss(training) and validation loss (test) [85 epochs] 53
Figure 24 Home Page 56
Figure 25 From Home Page To Test Page 57
Figure 26 Test Page 57
Figure 27 Input Phase 58
Figure 28 Output Phase 58
Figures ‘29-30-31-32-33’: Outputs “Cat -Seven -Sheila -Bird -Happy” 59

List of tables

Table 1. The SR used systems from the beginning till nowadays 8
Table 2. Illustrates the properties of CNN layers 30
Table 3. The used words and their sample number for model training 40
Table 4. Desktop Hardware 47
Table 5. Ilustre the results using 15 epochs 50
Table 6. Ilustre the results using 35 epochs 51
Table 7. Ilustre the results using 55 epochs 52
Table 8. Ilustre the results using 85 epochs 53
Table 9. The obtained results by training our model with different numbers of epochs 54
Table 10. Comparison between the obtained results of ANN and CNN models 54

List of acronyms

NLP	Natural Language Processing;
AI	Artificial Intelligence
ANN	Artificial Neural Network
DNN	Deep Neural Network
RNN	Recurrent Neural Network
MLP	Multilayer Perceptron
CNN	Convolutional Neural Network
VC	Voice Recognition
SR	Speech Recognition
IVR	Interactive Voice Response
IBM	International Business Machines
PC	Personal Computer
HMM	Hidden Markov Model
DARFA	Defense Advanced Research Projects Agency
SUR	Speech Understanding Research
VRCP	Voice Recognition Call Processing
LPC	Linear Predictive Coding
DFT	Discrete Fourier Transform
ADC	Analog Digital Converter
RASTA-PLP	Relative Spectral Transform-Perceptual Linear Prediction
LPCC	Linear Predictive Cepstral Coefficient
MFCC	Mel Frequency Cepstral Coefficient
IFT	Inverse Fourier Transform
RTF	Real Time Factor
WER	Word Error Rate
RAM	Random Access Memory
MB	Mother Board
CPU	Control Processing Unit
GPU	Graphic Processing Unit
API	Application Programming Interface
GUI	Graphical User Interface

General Introduction

General Introduction

NLP (Natural Language Processing) is the field of Artificial Intelligence that concentrates on the cooperation among computers and human languages, specifically how to program computers to process and dissect a lot of natural language Data. NLP works intimately with Speech/Voice Recognition and Text Recognition motors. Presently NLP and Associated AI technologies have entered the buyer domain. NLP alludes to the developing arrangement of Computer and AI-based Technologies that permit computers to learn, comprehend, and produce content in human languages. The Technology works intimately with Speech/voice Recognition and Text Recognition motors. While Text/Character Recognition and Speech/Voice Recognition permits computers to input the data, NLP permits sorting out this information.

After decades of research, speech recognition technology has advanced in both theory and practice. Speech recognition technology laid the foundation for statistical modeling of the language from the first attempt at speech analysis using a digital computer to dynamic time warping pattern matching. It has evolved with the introduction of technology. Hidden Markov Model was introduced into speech recognition, making statistical approaches ubiquitous in speech processing. This established the core technology of speech recognition and entered the era of modern speech recognition engines. Several efforts were made to improve the accuracy of speech recognition systems by modeling speech using large amounts of speech data and performing extensive evaluations of speech recognition in a variety of tasks and languages. The maturity of speech recognition technology achieved during these years has also enabled the development of practical applications for human-computer speech interaction and speech information retrieval. The great potential of such applications is researched from speech recognition, collected in a controlled environment, and strictly limited to domain-oriented content, to modeling conversational speech with all variability and language specific problems. This has created a next-generation speech recognition system that aims to reliably recognize large vocabulary and continuous speech even in acoustically unfavorable environments and various operating conditions. Therefore, the main issues today are the robustness and scalability of automated speech recognition systems, as well as integration with other speech processing applications.

This factsheet gives an outline of how you can utilize speech recognition. You can use it to educate a brilliant speaker, control a savvy home, and order phones and tablets. Furthermore, you can set updates and connect without hands with individual innovations. The main use is for the section of text without utilizing an on-screen or actual console. Communication technology is

developing rapidly. It's now much easier to type text using speech recognition, check the spelling of words, and dictate messages. Most on-screen keyboards have a microphone icon that allows the user to easily switch from typing to speaking. Speech recognition opens a world of productive possibilities for some people with disabilities who find it difficult or impossible to use a mouse or keyboard. It frees people from typing and keyboard use helps people with

disabilities and reduces the risk of repetitive strain injury due to excessive typing and mouse use. For example, a person with dyslexia may find that speech recognition can be used to write more fluently, accurately, and quickly, and is less stressful than traditional handwriting or typing. For employers, enabling speech recognition in the system

and encouraging use in the workplace can be a "reasonable adaptation." That is, to prevent discrimination and maximize the productivity of employees with disabilities. For a "reasonable adaptation." That is, to prevent discrimination and maximize the productivity of employees with disabilities.

This thesis is organized as follows:

1. We start with the first chapter which is called "the state of the art", it focused on the field of speech recognition and more detailed description on convolutional neural networks (CNNs) which is the chosen method of our project.

2. The second chapter is called "conceptual study", consist of the design of our system. We present the details of our approach, thus a comparison and discussion of the results found will be presented.

3. Chapter three and the last one, presents the implementation of our system as well as the used programming tools.

The brief ends as usual with a conclusion. We also indicate some possible perspectives for

our research activities.

Chapter one:

State of the art

Chapter 1: State of the Art

1. Introduction

Language is the main method of communication and speech, is its fundamental medium. In human to machine interface, the speech signal is changed into a simple and digital wave structure that can be perceived by machines. Speech technologies are incomprehensibly utilized and have limitless purposes. These technologies empower machines to respond accurately and dependably to human voices and offer helpful and important types of assistance.

Speech recognition is also called automatic speech recognition (ASR) and voice recognition. It recognizes the verbally expressed words and expressions and converts them to a machine-readable format. By changing over spoken sound into text, SR technology lets users control digital devices by talking as opposed to utilizing ordinary apparatuses, for example, keystrokes, buttons, consoles; and so on.

Speech recognition has a wide scope of utilization and is well spread out in contact places, IVR systems, mobile and embedded devices, dictation solutions, and assistive applications.

2. Section 2

A. History

A.1. Voice Recognition

There has been Exponential Growth in Voice Recognition innovation throughout recent many years. Tracing all the way back to 1976, PCs could see somewhat in excess of 1,000 words [1]. That complete leaped to approximately 20,000 during the 1980s [1] as IBM kept on creating Voice Recognition Technology.

The primary speaker recognition section for buyers, was sent off in 1990 by Dragon[1], called DragonDictate. In 1996, IBM introduced the very first voice recognition provision that could recognize continuous speech[1]. After the dispatch of cell phones in the final part of the 2000s, Google sent off its voice search application for the iPhone. After three years, Apple presented Siri, which is currently a conspicuous Voice Recognition Assistant.

During the former decade, a few other innovation pioneers have likewise grown more refined Voice Recognition Software, with Amazon's Echo highlighting Alexa and Microsoft's Cortana -- the two of which go about as private assistants that answer voice commands.

A.2. Speech Recognition

The primary significant period of this development started at IBM's Bell Labs [2]. In 1952, IBM announced Audrey[3], the main recorded speech recognition feature. Audrey was a completely analog system, containing a single number with a stop in between. Ten years later, IBM announced the Shoebox [1]. This is a device that can recognize 16 English words and numbers from 0 to 9. In the mid of 1970s, there was a leap forward in this innovation. This is primarily due to DARPA, the US Department of Defense's Research and Development Agency.

After five years of exploration, Harpy was born at Carnegie Mellon University [4]. A machine had the ability of 1011 words. Also, the harpy was not essentially the same as his ancestors. There might be sentences. In the mid of 1980s, the scope of speech recognition frameworks expanded to thousands of words. This was mainly achieved by the hidden Markov model [5]. Speech recognition has changed from advanced draft-based signal processing to using statistical models to predict words from ambiguous sounds.

In addition, machines became out to be extra particular in perceiving words. The Speech Recognition Group at IBM provided Tangora [6], an exploratory file framework, at some point in the 80s. Tangora turned into the match for recognizing 20000 words. At the beginning of 1990s, Speech Recognition items, for example, DragonDictate unfolded to consumers as a consequence of PCs [6]. Over the maximum latest twenty years, several techs monsters were interested in this Technology.

B. Definitions

B.1. Voice recognition

Voice recognition is the capacity of the program to distinguish individuals in view of their voiceprints. Scan the language and work by setting the desired speech fingerprint and match. The development of AI opened a wide possibility for this subfield of computer science. It allows us to interact with machines without touching them. It is growing fast, and developers will find more ways to apply it in various fields.

B.2. Speech recognition

The straightforward meaning of Speech Recognition is an innovation that empowers a PC to perceive, comprehend, and make an interpretation of human discourse into text. Speech Recognition Technology utilizes Natural Language Processing NLP and AI to decipher human discourse. Engineers utilized the term Automatic Speech Recognition, or ASR, in the mid-1990s

to stretch that Speech Recognition is machine handled. Yet, today, ASR and Speech Recognition are synonymous terms.

B.3. The difference between voice recognition and speech recognition?

Understanding the distinctions between these two is significant. The objective of Voice Recognition is to recognize the voice proprietor. Speech Recognition's objective is to recognize the words of the speaker. In the principal task, the program needs an excellent voiceprint of the speaker for examination. The program needs an enormous word reference to distinguish the speaker's words in the subsequent undertaking.

B.4 Speech Recognition types*:* Based on the sort of words, Speech Recognizing systems can perceive, the Speech Recognition system is classified into the accompanying classes:

B.4.1 Isolated Word: Isolated Word requires every expression to have calm on the two sides of the test window. At a time only single words and single expressions are acknowledged and it is having a "Listen and Non-Listen state".

B.4.2 Continuous Word: Consistent speech recognizers give the users an easier way to talk in a continuous style and naturally and simultaneously the gadget decides the substance of the speech. recognizers delivering the office of continuous speech abilities are basically challenging to make since they require an extraordinary and particular a strategy to decide the limits of the expressions.

B.4.3 Connected Word: Connected words are a lot of the same the isolated words, yet they permit separate expressions to be executed with "minimal pauses" in the middle between them.

B.4.4 Spontaneous speech: At a rudimentary level, Spontaneous Speech can be considered as a Speech that is coming out naturally and not a practiced one. An Automatic speech recognizer should have the option to deal with a wide scope of discourse highlights like the words being run together.

3. Used Systems

The advancement of speech recognition technology has gone on since the 1950s. We should investigate how this technology has advanced throughout the long term, how has our technique for utilizing speech recognition and speech-to-text capacities advanced with technology?
Across the globe, different countries created equipment that could recognize sound and speech. Toward the finish of the '60s, the technology could uphold words with four vowels and nine consonants. In the 70s, a growing vocabulary, the capacity for computers to handle normal human language could demonstrate priceless in quite a few regions in the military and public safeguard. Julie, Harpy, had the option to answer a speaker and had the ability to recognize the speaker's voice.
The capacity to recognize speakers was not by any means the only progression made during this time. Researchers began leaving the thought that speech recognition must be absolutely acoustically based. they moved to Natural Language Processing (NLP). Rather than simply utilizing sounds, researchers went to algorithms to program systems with English language rules.

Year	Authors	System technology	Deduction
1950	Bell Laboratories	AUDREY	The primary Speech Recognition Systems were centered around numbers, not words. The AUDREY System perceived spoken numbers with 97-close to 100% precision.[3]
1962	International Business Machines (IBM)	SHOEBOX	Shoebox recognized numbers and simple math terms.[1]
1970	The US Department of Defense (DARPA)	Speech Understanding Research (SUR) funded HARPY	HARPY came from SUR program and was able to understand over 1,000.[4]
1980	IBM	ASR using HMM	the HMM estimated the probability of the unknown sounds being words, rather than simply involving words and searching for sound patterns.[5]
1992	AT&T	Voice Recognition Call Processing (VRCP)	VRCP handled around 1.2 billion voice transactions every year.[7]

1997	Dragon Dictate	Naturally Speaking	Dragon allowed for natural speech to be processed without the need for pauses. [6]
/	/	/	Speech Recognition Technology had accomplished near 80% accuracy. For the greater part of the ten years, there weren't much of progressions until Google showed up. [6]
In the 2000s	Google	Google Voice Search System	At the time, the launch of Google Voice Search included 230 billion words from user searches in English.
2010	Google	Google voice search app	Google made a game-changing development that carried speech recognition technology to the front line of innovation. It aimed to lessen the problem of composing on your phone's small console and was the first of its sort to use cloud server farms. It was likewise customized to your voice and had the option to 'learn' your speech patterns for higher accuracy.
2011	Apple	Siri	Siri turned out to be immediately popular for her amazing capacity to precisely handle normal expressions, as well as her capacity to answer utilizing conversational language. Siri's prosperity carried speech recognition technology to the bleeding edge of development and innovation.
2014	Amazon	Alexa	Alexa was on the ball with its joining with smart home devices, for example, cameras, lighting, thermostats, door locks, and entertainment systems.

Table1: The SR used systems from the beginning till nowadays.

The technology to help voice applications is currently both moderately economical and strong. With the headways in computerized reasoning and the rising measures of speech data that can be effectively mined, truly conceivable voice turns into the following predominant connection point. The potential outcomes of AI and machine learning propose that voice integration has a splendid future ahead, and that voice integration technology may before long be nearer to the core of the cutting edge working environment.

4. Voice Recognition Process

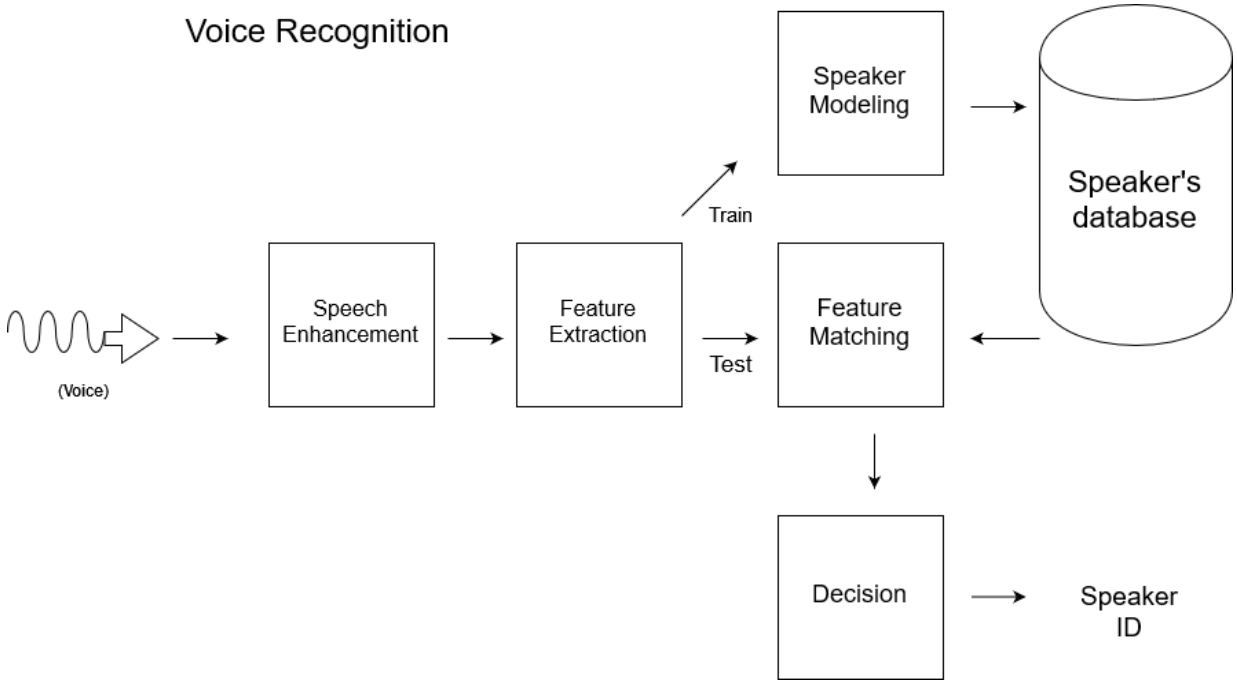

Figure 1: Voice recognition process.[8]

Voice Recognition relies upon a recorded layout of a user's voice, called "Template Matching." A program should be "Trained" to Recognize a user's voice.

In the first place, the program will show a printed word or expression that the user talks and rehashes multiple times into the system's microphone to train the voice recognition software.

Then, the program processes a statistical average of several samples of a similar word or expression.

At long last, the program stores the average sample as a template in its data structure.

5. Speech Recognition System

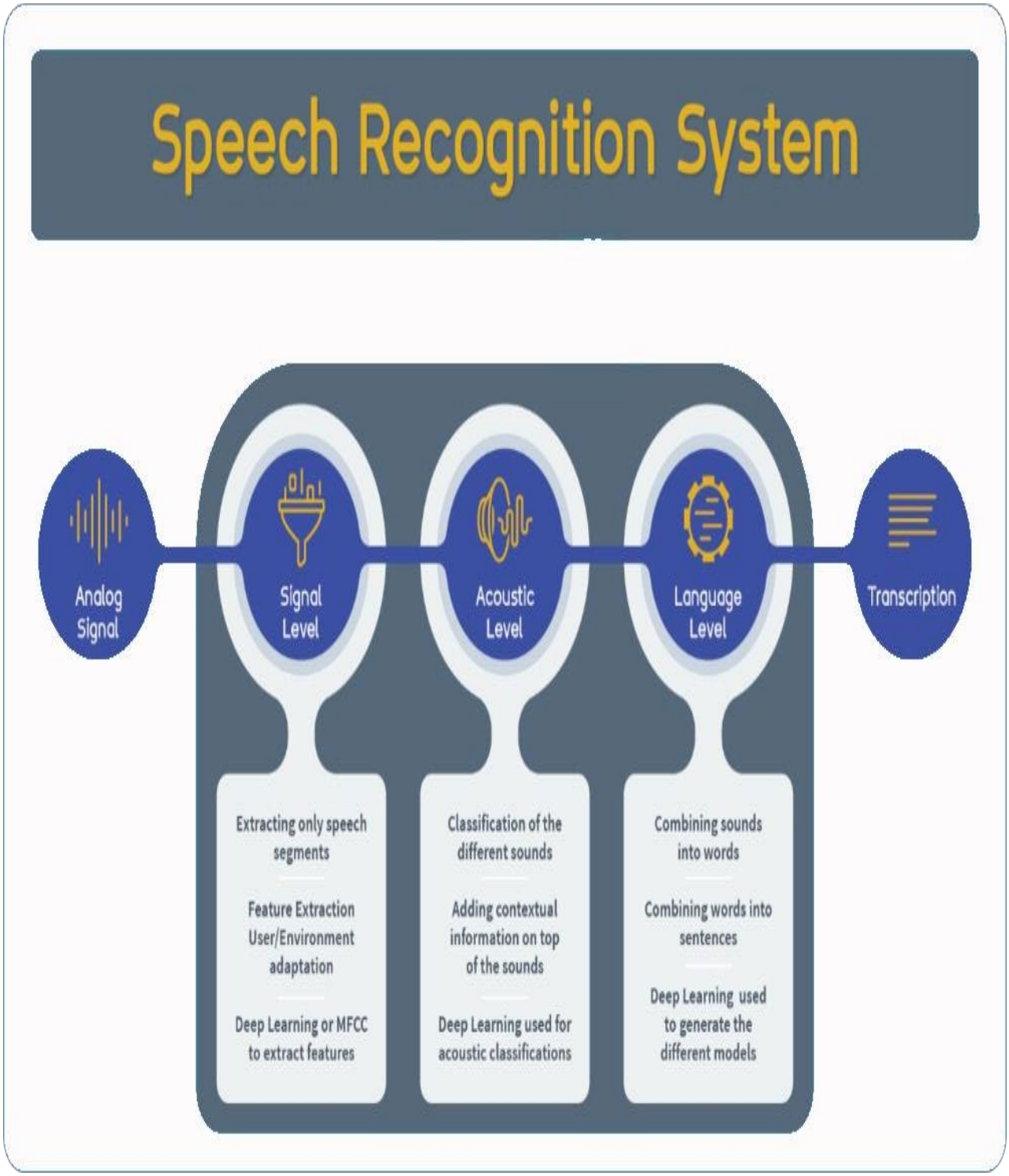

Figure 2: Overview of a speech recognition system.[9]

5.1 The Speech Signal

Speech signal includes signal features and frequency domain signal features which are mainly used for segmenting speech signals. Three important characteristics of the speech signal are short-time zero-crossing, energy and auto-correlation. The short-time energy and the short-time zero-crossing rates are important properties for detecting the endpoint of a speech signal analysis. Especially these two properties are used in voiced and unvoiced segmentation and classification.[10]

Other characteristics of speech signals are pitch, stress, power spectral density, vowel duration, rhythm and intonation patterns. These characteristics are mainly involved in speech segmentation which is recognizable and meaningful.[11]

Basic speech signal characteristics such as pitch and intonation which is identified by producing the speech. The speech is produced by air pressure come from lungs through vocal cords via vocal track. When vocal cord does not vibrate unvoiced sound is produced. Voiced sounds are produced when vocal card vibrate correctly.[10]

Pitch frequency is an important parameter of speech processing. Vocal cord produced voiced and unvoiced sounds based on it vibration. A vibration sounds are delivered with glottal pulse, it has fundamental frequency and harmonics. The fundamental frequency of glottal pulse is called pitch. Basically, pitch is also known as frequency of sound. Sound can be characterized based on pitch value, loudness and quality. It is compared like high or low in musical sounds. The pitch is just ear response of frequency. Human can hear range of sound between 3Hz to 3,000 Hz. Intonation speeches is basically a matter of vibrating in the pitch level of the voice [12]

5.2 Speech recognition approaches

There are several approaches to speech recognition, broadly speaking, there are three approaches to speech recognition namely [13]:

1. The acoustic-phonetic approach.
2. The pattern recognition approach.
3. The artificial intelligence approach.

5.2.1 Acoustic-phonetic approach to speech recognition

There are three steps in the acoustic-phonetic approach [13]:

1. Speech analysis.
2. Feature detection.
3. Segmentation and labeling.

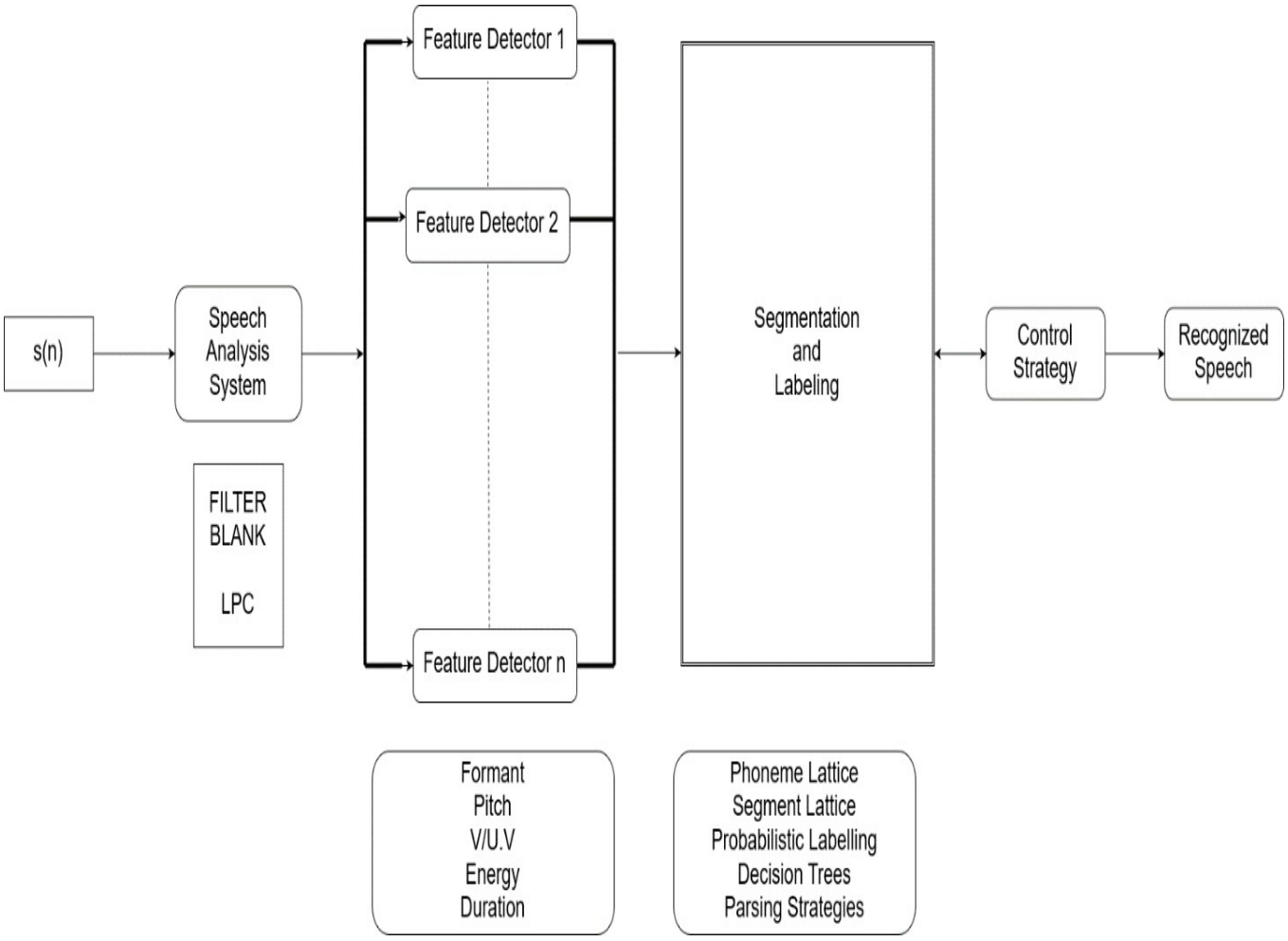

Figure 3: acoustic-phonetic speech recognition system.

5.2.1.1 Speech analysis

A step common to all approaches in speech recognition, provides an appropriate (spectral) representation of the characteristics of the time-varying speech signal. Most common technique of spectral analysis are the class of filter bank methods and the class of linear predictive coding (LPC) methods. [13]

5.2.1.2 Feature detection

converts the spectral measurement to a set of features that describes the broad acoustic properties of the different phonetic units. Features proposed for recognition are nasality, friction, formant locations, voiced and unvoiced classification. It usually consists of a set of detectors that operate in parallel and use appropriate processing and logic to make the decision as to the presence or absence, or value, of a feature. [13]

5.2.1.3 Segmentation and labeling

Here the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech. From phoneme lattice characterization of the speech, a lexical access procedure determined the best matching word and sequence of words. [13]

5.2.2 Pattern-recognition approach to speech recognition

The pattern recognition paradigm has four steps [13]:

1. Feature measurement.

2. Pattern training.

3. Pattern classification.

4. Decision logic.

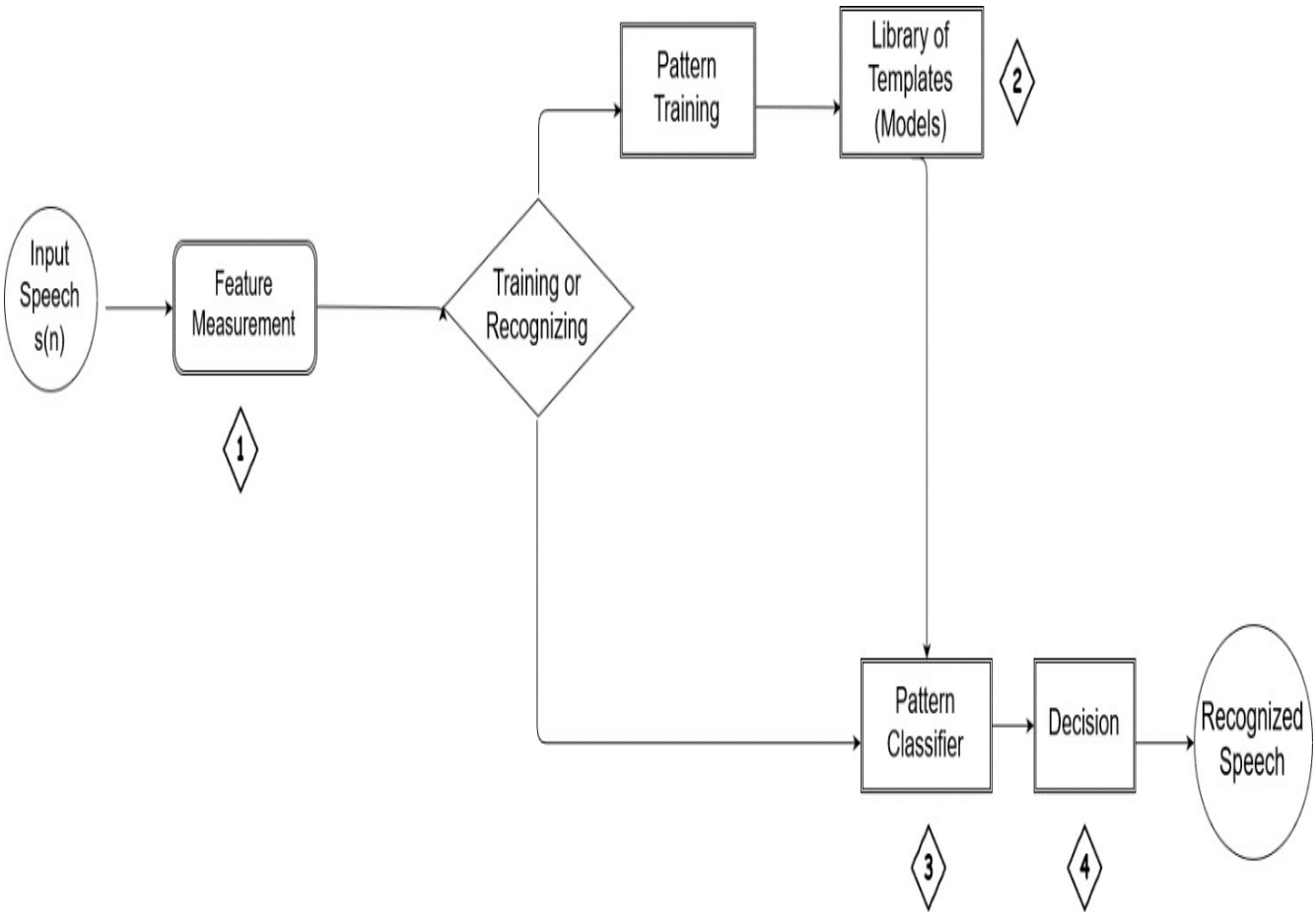

Figure 4: block diagram of patter recognition speech recognizer.

5.2.2.1 Feature measurement

In which a sequence of measurement is made on the input signal to define the "test-pattern". For speech signal, the feature measurement is usually output of some type of spectral analysis technique such as a filter bank analyzer, a linear predictive coding analysis, or discrete Fourier transform (DFT) analysis. [13]

5.2.2.2 Pattern training

In which one or more test patterns corresponding to speech sound of the same class are used to create a pattern representative of the features of that class. The resulting pattern generally called a reference pattern, can be an exemplar or a template, derived from some type of averaging technique. [13]

5.2.2.3 Pattern classification

In which the unknown test pattern is compared with each (sound) class reference pattern and a measure of similarity (distance between test pattern and each reference pattern is computed).To compare speech patterns (which consist of sequence of spectral vectors), we require both the spectral "distance" between two well-defined spectral vectors and global time alignment procedure that compensates for different rates of speaking (time scale) of the two patterns. [13]

5.2.2.4 Decision logic

In which reference pattern similarity scores are used to decide which reference pattern best matches the unknown test patterns. [13]

5.2.3 Artificial intelligence approaches to speech recognition

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Thus for example, the AI approach to segmentation and labeling would be to augment the generally used acoustic knowledge with phonetic knowledge, lexical knowledge, syntactic knowledge, semantic knowledge, and even pragmatic knowledge. [13]

5.3 The involved process of speech recognition

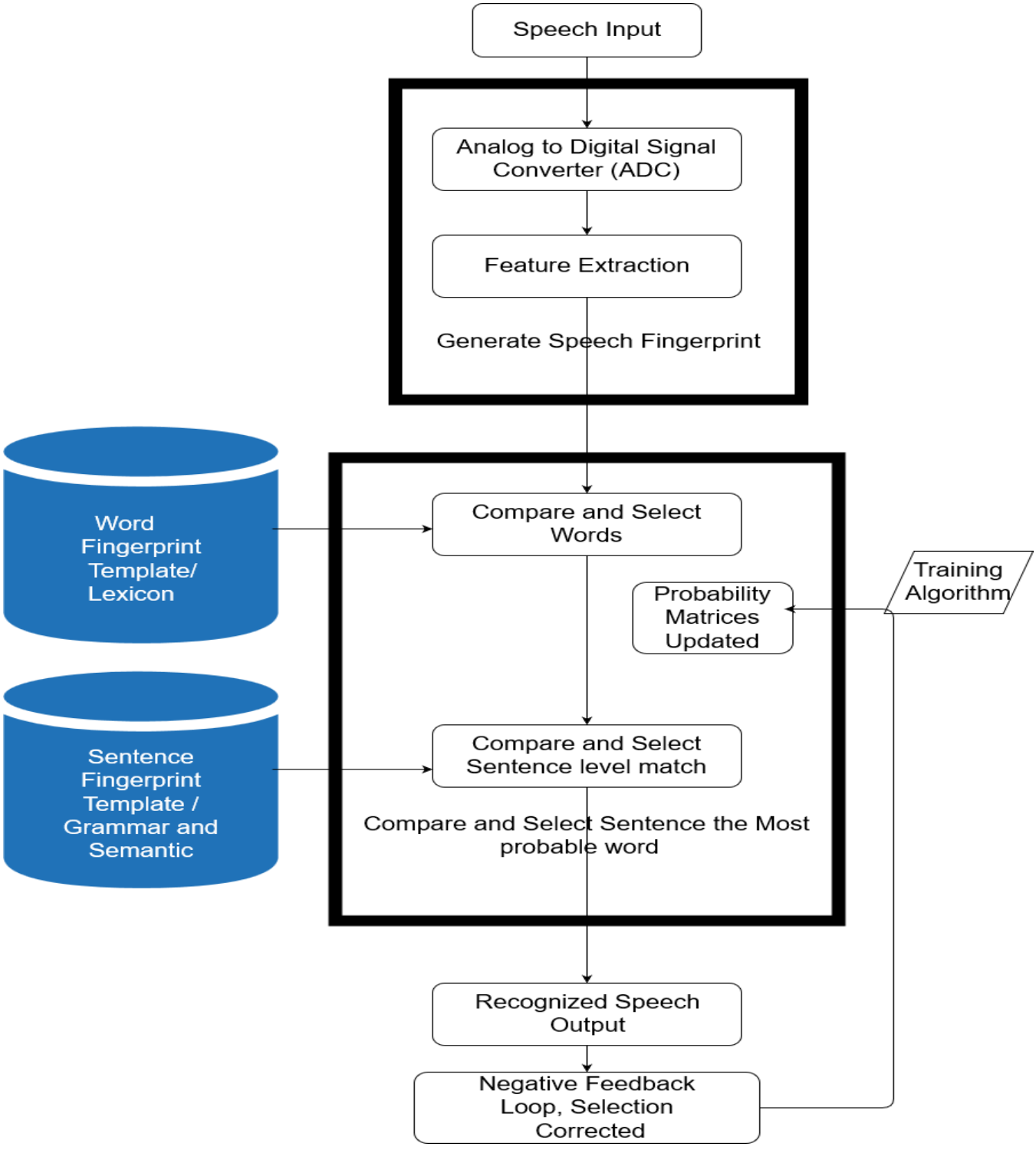

Figure 5: Steps involved in Speech Recognition.[14]

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information.

For the process of the speech recognition, we can identify the following main steps:

Speech acquisition (Analog-to-Digital Conversion).

Pre-processing.

Feature Extraction.

Modeling.

Recognition.

5.3.1 Speech acquisition (Analog-to-Digital Conversion):

Speech is defined as the ability to express one's thoughts and feelings by articulating sounds. Initially, the speech of a person is received in the form of a waveform. Also, there are numerous tools and software available which record the speech delivered by humans. The phonic environment and the equipment device used to have a significant impact on the speech generated. There is a possibility of having background or room reverberation blended with the speech, but this is completely undesirable.[16]

Speech is usually recorded/available in analog format. Standard sampling techniques/devices are available to convert analog speech to digital using techniques of sampling and quantization. Digital speech is usually a one-dimension vector of speech samples, each of which is an integer. [15]

5.3.2 Pre-processing:

The solution to the problem described above is "Pre-Processing". It plays an influential role in canceling out the trivial sources of variation. The speech pre-processing typically includes reverberation canceling echo cancellation, windowing, noise filtering, and smoothing all of which conclusively improve the accuracy of speech recognition.[16]

Recorded speech usually comes with background noise and long sequences of silence. Speech pre-processing involves the identification and removal of silence frames and signal processing techniques to reduce/eliminate noise. After pre-processing, the speech is broken down into frames of 20ms each for further steps of feature extraction. [15]

5.3.3 Feature Extraction:

It is the process of converting speech frames into feature vector which indicates which phoneme/syllable is being spoken. [15]

Each and every person has different speech and different intonation. This is due to the different characteristics ingrained in their utterance. There should be a probability of identifying speech from the theoretical waveform, at least theoretically. As a result of an enormous variation in speech, there is an imminent need to reduce the variations by performing some feature extraction. The ensuing section depicts some of the feature extraction technologies which are extremely used nowadays.[16]

5.3.4 Modeling:

Based on a language model/probability model, the sequence of phonemes/features are converted into the word being spoken. [15]

The basic objective of modeling technique is to generate speaker models using speaker specific feature vector. The speaker modeling technique divided as: speaker recognition and speaker identification. It automatically identifies who is speaking on the basis of individual information integrated in speech signal. The speaker recognition is also divided into two parts that mean speaker dependent and speaker independent. In Speaker independent approach of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the expected message. Besides this, in speaker dependent, recognizing machine should extract speaker characteristics in the acoustic signal.[16]

The main aim of speaker identification is to compare a speech signal from an unknown speaker to a database of known speakers. The system can recognize the speaker, which has been trained with a number of speakers. Speaker recognition can also be divided into two methods, text-dependent, text-independent methods. In the text-dependent method, the speaker says keywords or sentences having the same text for both training and recognition trials, whereas the text-independent does not dependent on a specific text being spoken [16] new. Following are approaches to speech recognition.[16]

5.3.5 Recognition

Training and testing are the phases for speech recognition to be completed. The training phase is like object identification. It very well may be repeated ordinarily for better recognition which improves the performance while testing. The testing phase incorporates the comparison between the pattern scored while training and spoken words at testing time. The degree of closeness in these two phases counts for improving the system performance. In any case, fluctuation experienced

during recognition impacts a great deal on the constant recognition rate.[16]

5.4 The most popular feature extraction methods:

5.4.1 Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP):

PLP is a way of warping spectra to minimize differences between speakers while preserving important speech information.

RASTA applies a band-pass filter to the energy in each frequency subband, in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel e.g., from a telephone line. [17]

5.4.2 Linear Predictive Cepstral Coefficients (LPCCs):

Cepstrum is the result of taking the Inverse Fourier Transform (IFT) of the logarithm of the estimated spectrum of a signal. The power cepstrum is used in the analysis of human speech. [17]

5.4.3 Mel Frequency Cepstral Coefficients (MFCCs):

These are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between LPCC & mel-frequency cepstrum is that in the MFCC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping allows for a better representation of sound. [17]

5.5 The traditional probability mapping and selection method:

A Hidden Markov Model is a type of graphical model often used to model temporal data. Hidden Markov Models (HMMs) assume that the data observed is not the actual state of the model but is instead generated by the underlying hidden (the H in HMM) states. While this would normally make inference difficult, the Markov Property (the first M in HMM) of HMMs makes inference efficient. The hidden Markov model can be represented as the simplest dynamic Bayesian network. The mathematics behind the HMM was developed by L. E. Baum and coworkers. Because of their flexibility and computational efficiency, Hidden Markov Models have found a wide application in many different fields like speech recognition, handwriting recognition, and speech synthesis. [18]

5.6 The mathematical fundamental equation of statistical speech recognition:

Acoustic models include the representation of knowledge about acoustics, phonetics, microphone and environmental variability, gender, and dialect differences among speakers, etc. [19]

Language models refer to a system's knowledge of what constitutes a possible word, what words are likely to co-occur, and in what sequence. [19]

The semantics and functions related to an operation, a user may wish to perform may also be necessary for the language model. Many uncertainties exist in these areas, associated with speaker characteristics, speech style and rate, the recognition of basic speech segments, possible words, likely words, unknown words, grammatical variation, noise interference, nonnative accents, and the confidence scoring of results. [19]

A successful speech-recognition system must contend with all of these uncertainties. The acoustic uncertainty of the different accents and speaking styles of individual speakers are compounded by the lexical and grammatical complexity and variations of spoken language, which are all represented in the language model. [19]

The division of acoustic modeling and language modeling discussed above can be succinctly described by the fundamental equation of statistical speech recognition[19]:

$$\widehat{W} = \arg max_w P(W/A) = \arg max_w \frac{P(W)P(A/W)}{P(A)} \quad (1)$$

For the given acoustic observation or feature vector sequence X = X1X2 . . . Xn, the goal of speech recognition is to find out the corresponding word sequence $\hat{W} = W_1, W_2, \ldots, W_m$ that has the maximum posterior probability P(W|X) as expressed with Equation (1).

Since the maximization of Equation (1) is carried out with the observation X fixed, the above maximization is equivalent to the maximization of the numerator in the equation (2) [19]:

$$\widehat{W} = \arg max_w P(W)P\left(\frac{X}{W}\right) \quad (2)$$

where P(W) and P(X|W) constitute the probabilistic quantities computed by the language modeling and acoustic modeling components, respectively, of speech recognition systems. [19]

5.7 The speed of a speech recognition system measured:

$$\mathrm{RTF} = \frac{\mathrm{Time}\left(\mathrm{Decode\ (a)}\right)}{\mathrm{Length(a)}} \tag{3}$$

Real-Time Factor (see eq.3) is a very natural measure of a speech decoding speed that expresses how much the recognizer decodes slower than the user speaks. The latency measures the time between the end of the user's speech and the time when a decoder returns the hypothesis (equation 3), which is the most important speed measure for SR [20].

Real-time Factor (RTF): the ratio of the speech recognition response time to the utterance duration. Usually, both mean RTF (average over all utterances), and 90th percentile RTF is examined in efficiency analysis [21].

5.8 The Word Error Rate (WER):

Word error rate is a common metric of the performance of speech recognition or machine translation system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level [22].

This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Word error rate can then be computed as [23]:

$$WER = \frac{S+D+I}{N} \tag{4}$$

Where:

• S is the number of substitutions,

• D is the number of the deletions,

• I is the number of the insertions,

N is the number of words in the reference.

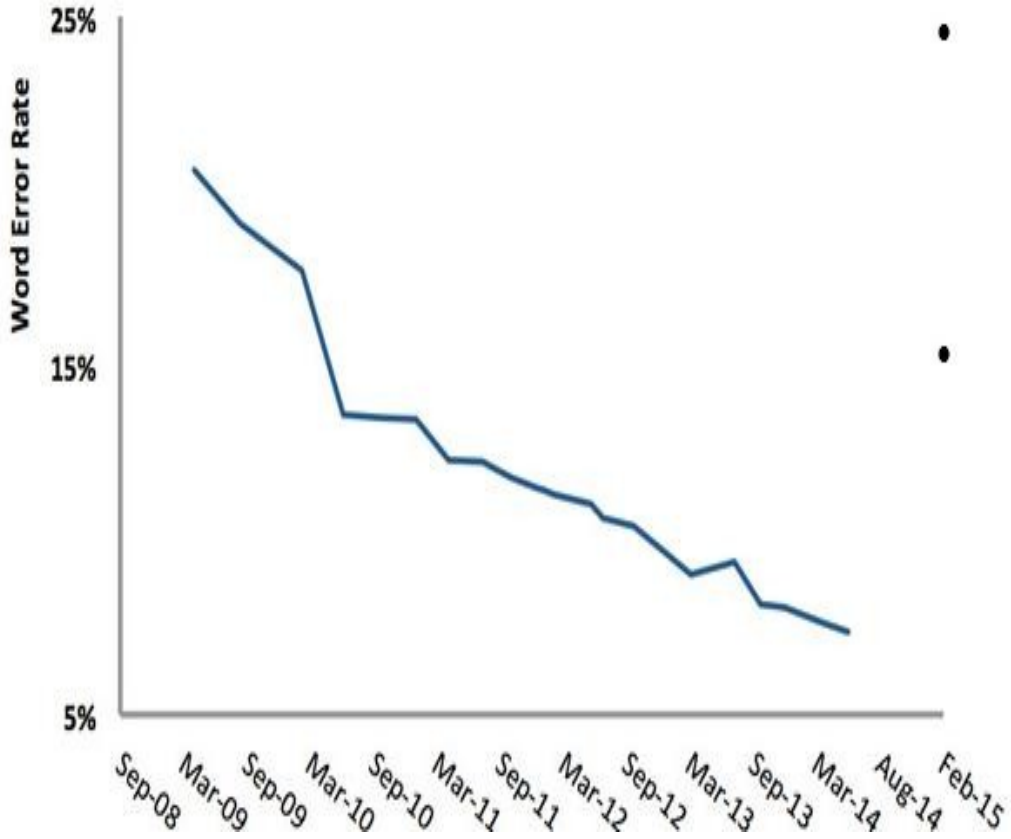

Figure 6: Chart is showing the reduction of SR error rate over time, attributed to deep learning.[17]

This graph, from a blog by speech specialist Nuance's research director Nils Lenke, shows how WERs declined sharply from around 2010 as neural networks were successfully incorporated into SR systems.[24]

The current state of the art for WER in voice search and virtual assistants is now in single figures.[24]

1. Continuous server SR word error rate (WER) reduction ~ 18% per year:

Combination of Algorithms, data, and computing. [24]

2. Deep learning (DNNs) is driving recent performance improvements in speech recognition and meaning extraction. [24]

several fundamental characteristics, and it is divided into two types of features such as time-domain speech

6. Speech recognition based on deep learning & neural network

Deep Learning is a subset of Machine Learning where the models, for the most part, Artificial Neural Networks, utilize different layers to make a more extravagant portrayal of the Data. Conventional AI calculations like Support Vector Machines, require hand-made highlights to arrive at ideal outcomes, while Deep Learning calculations are equipped for making their elements from additional crude Data, in light of what is figured out, and how to be important during preparation.

Deep Learning allows computational models that are made out of various handling layers to learn representations of Data with numerous Levels of abstraction. These techniques have emphatically worked on the best in class in Speech Recognition, Visual Object Recognition, Object Detection, and numerous different areas like medication revelation and genomics. Deep Learning tracks down capricious plans in gigantic instructive assortments by using the Backpropagation Algorithm to show how a machine must alter inside limits are used to enlist the depiction in each layer from the depiction in the previous ones.

6.1 Artificial neural networks (ANN)

ANN is a classification approach used to obtain higher accuracy by finding the sentiments of sarcastic sentences. ANN has shown outstanding capabilities for modeling difficult word composition in one sentence. A sarcastic text is a sequence of words or combination of words. ANN is used to store these text signals into its temporary storage [25] (Figure 7).

ANN is a computational algorithm that is used to solve the difficult problem in a similar fashion as a brain does. The inputs are provided to the input layer, and weight is added individually to the inputs in the hidden layer. The weight is added or removed as per the bias function. Theweighted values along with the input are provided to the activation function. There are different types of activation functions.[25]

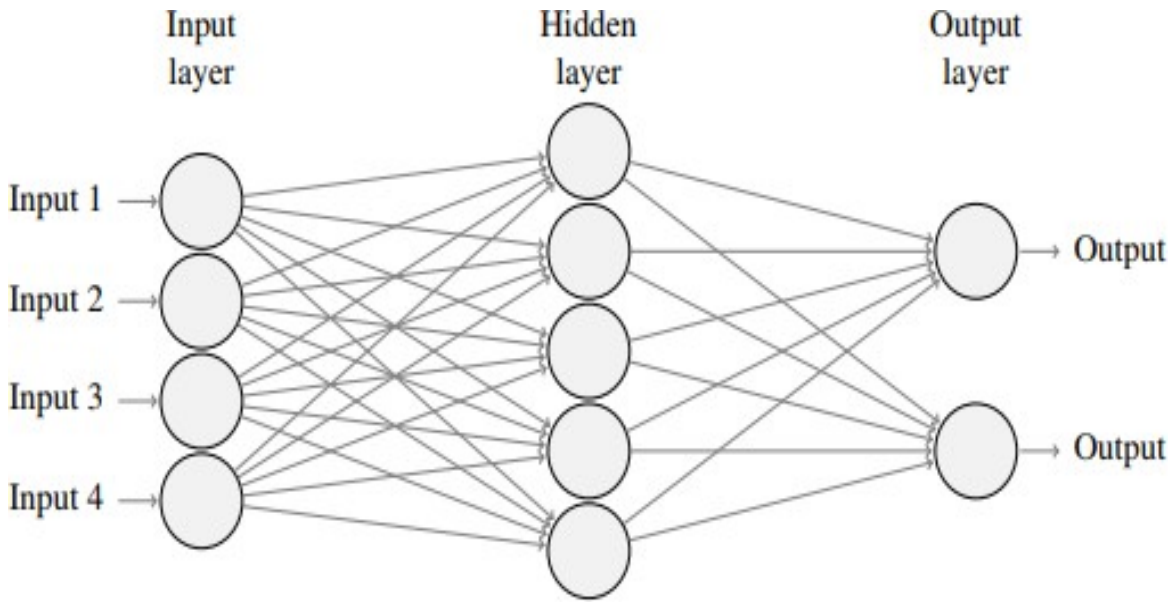

Figure 7: Illustration of a possible neural network.[26]

Artificial neural networks (ANN) are, as the name suggests, propelled by the modern usefulness of the human mind where neurons process data in parallel. ANN comprises of a layer of information hubs, at that point one shrouded layer of hubs and lastly a layer of yield hubs (Figure 7). Deep neural networks (DNN) adds more concealed layers to that. Most SR frameworks utilize HMMs to manage worldly assortment and GMMs to decide how well each HMM state fits a casing of the acoustic info, i.e. the likelihood, however DNNs has as of late been demonstrated to beat GMMs on an assortment of benchmarks and are presently utilized as a part of some path by numerous real business SR frameworks, e.g. Xbox, Skype Interpreter, Google Now, Windows Cortana, Apple Siri, Amazon Alexa and so on.

6.2 Deep Neural Network (DNN)

Deep neural networks (DNNs) have recently demonstrated impressive performance in complex machine learning tasks such as image classification or speech recognition. [27] Deep neural network models have become a powerful tool of machine learning and artificial intelligence. DNN is an ANN with multiple layers between the input and output layers. DNN models were originally inspired by neurobiology. On a high level, a biological neuron receives multiple signals through the synapses contacting its dendrites and sends a single stream of action potentials out through its axon. [27]

The complexity of multiple inputs is reduced by categorizing its input patterns. Inspired by this intuition, artificial neural network models are composed of units that combine multiple inputs and produce a single output. [27]

The following parts are the popularly used and they represent deep neural networks type

1. Multi-Layer Perceptron (MLP)
2. Recurrent Neural Networks (RNN)
3. Convolutional Neural Networks (CNN)

6.3 Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) is a class of a feedforward artificial neural network (ANN). MLPs models are the most basic deep neural network, which is composed of a series of fully connected layers. Today, MLP machine learning methods can be used to overcome the requirement of high computing power required by modern deep learning architectures. [27] Each new layer is a set of nonlinear functions of a weighted sum of all outputs (fully connected) from the prior one. [27]

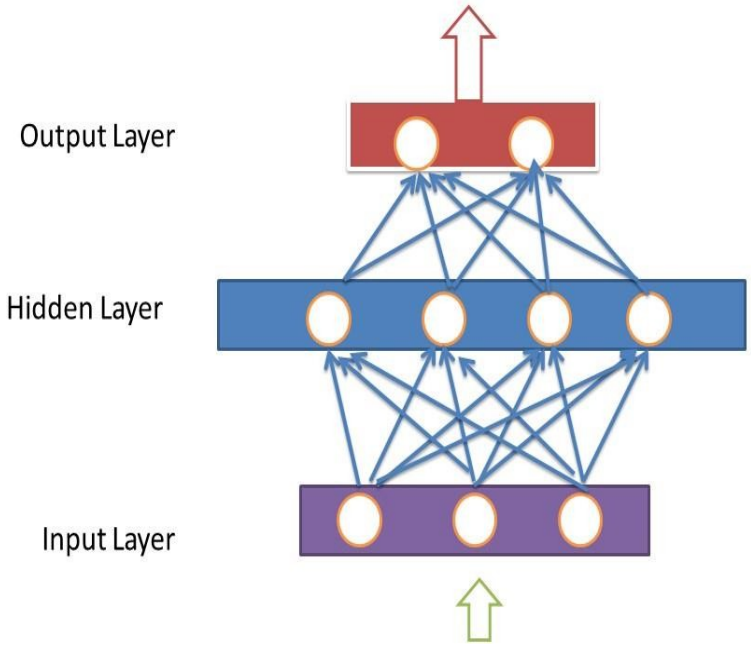

Figure 8: Multilayer perceptron.[19]

Backpropagation:

The backpropagation algorithm is one of the most important tools of an artificial neural network, it is specifically the part that deals with the training of the network, where it actually learns. During this process, the network updates the weights of all the edges to make it perform the correct output given a specific input.[28]

The backpropagation algorithm takes care of observing the network output, comparing it with the expected output, and slightly modifying the network weights so that the output approaches the expected output. To carry out this process, the difference between the network output and the expected output is calculated. The function that takes care of calculating this error is called the **loss function**. Being it an approximation, the loss function uses its derivative, and must therefore be differentiable by definition in order to work with the backpropagation algorithm.[28]

6.4 Recurrent Neural Networks (RNN)

A recurrent neural network (RNN) is another class of artificial neural networks that uses sequential data feeding. RNNs have been developed to address the time-series problem of sequential input data. [27]

The input of RNN consists of the current input and the previous samples. Therefore, the connections between nodes form a directed graph along a temporal sequence. Furthermore, each neuron in an RNN owns an internal memory that keeps the information of the computation from the previous samples. [27]

RNN models are widely used in NLP due to the superiority of processing the data with an input length that is not fixed. The task of the AI here is to build a system that can comprehend natural language spoken by humans, e.g., natural language modeling, word embedding, and machine translation. [27]

In RNNs, each subsequent layer is a collection of nonlinear functions of weighted sums of outputs and the previous state. Thus, the basic unit of RNN is called “cell”, and each cell consists of layers and a series of cells that enables the sequential processing of recurrent neural network models. [27]

6.5 Convolutional Neural Networks (CNN)

CNNs are the well-known variations of Deep Learning that are broadly embraced in SR systems. CNNs have numerous alluring advancements, Weight Sharing, Convolutional Filters, and Pooling. Thusly, CNNs have accomplished a noteworthy execution in SR.

Convolutional networks were the beginnings of Hubel and Wiesel who found that a single network architecture could reduce complexity in the feedback neural network when studying neurons used for local sensitivity and orientation selection in the cerebral cortex of cats. [29]

CNN is often used with image processing that requires a two-dimensional matrix containing features and may be three-dimensional, the pixel values are in the horizontal and vertical coordinate indicators. CNN is a neural network model. Its architecture has three main ideas, as explained (figure 9). Each one of them has the susceptibility to improve speech recognition performance (figure10). [30]

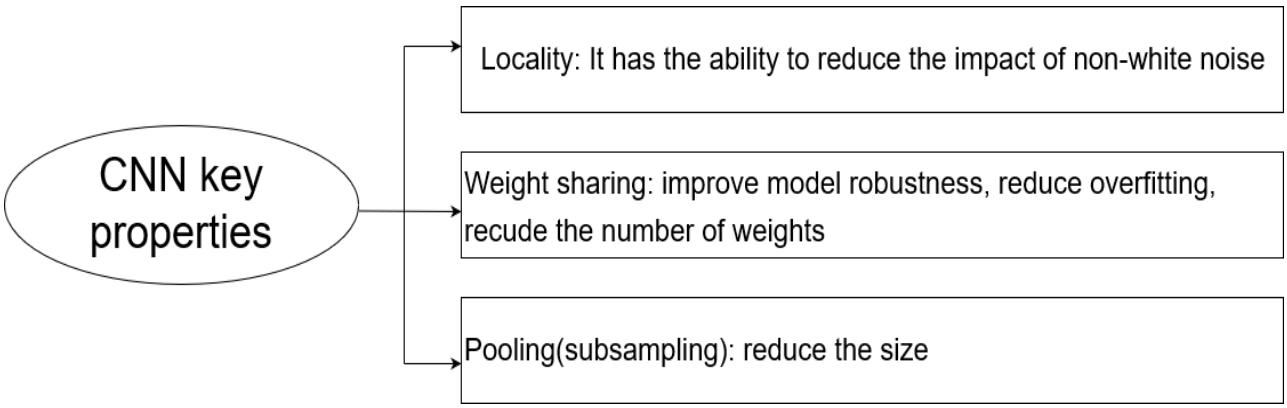

Figure 9: Architecture of CNN properties. [29]

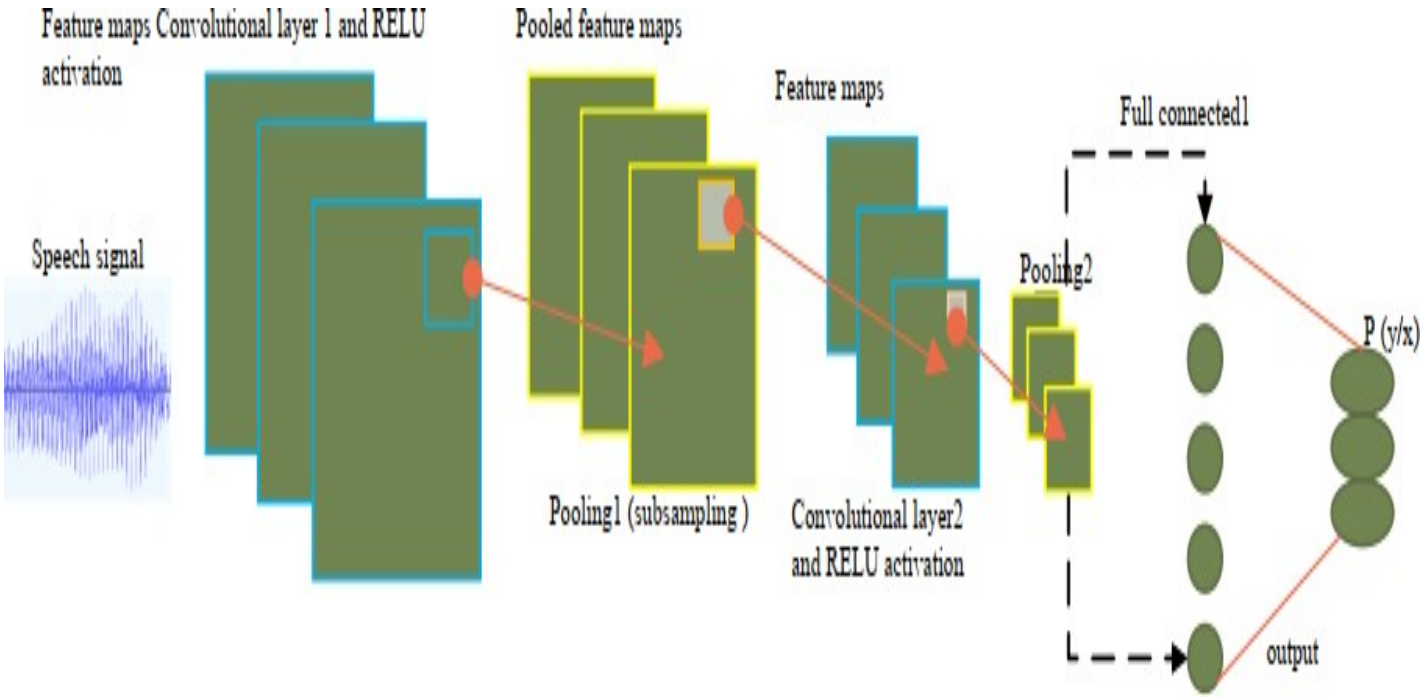

Figure 10: Architecture of CNN layers. [29]

CNN has a filter that shifts over the image to produce a feature map at convolution layers, through this window or filter, the weights of the network can identify the different features of the incoming image. [31].

The activation function decides if a particular feature is present at a particular location in the image. Usually uses a lot of filters over the image to find the necessary features [31]. CNN is often called the local network because the individual units computed in a specific location of the window depend on the local area that the window is currently looking at. [29]

Convolutional architecture is coordinated by three main layers arranged in the forward feed structure. The convolutional layer for feature extraction, sub-sampling layers, the aggregation (pooling) layer, reduce the dimensions of the input data and the output which a fully connected layer for final classes prediction [32]. linear filter and a nonlinear activation function, One of the most important elements [33]. In a convolutional layer, each plane is connected to one or more feature maps of the preceding layer [34] .an activation function is applied on to the result obtain the plane's output. The plane output is a 2-D matrix called a feature map; this name arises because each convolution output indicates the presence of a visual feature at a given pixel location [34]. A convolution layer produces one or more feature maps. Each feature map is then connected to exactly one plane in the next sub-sampling (pooling) layer.[33]

6.5.1 Activation function and how it works [34]

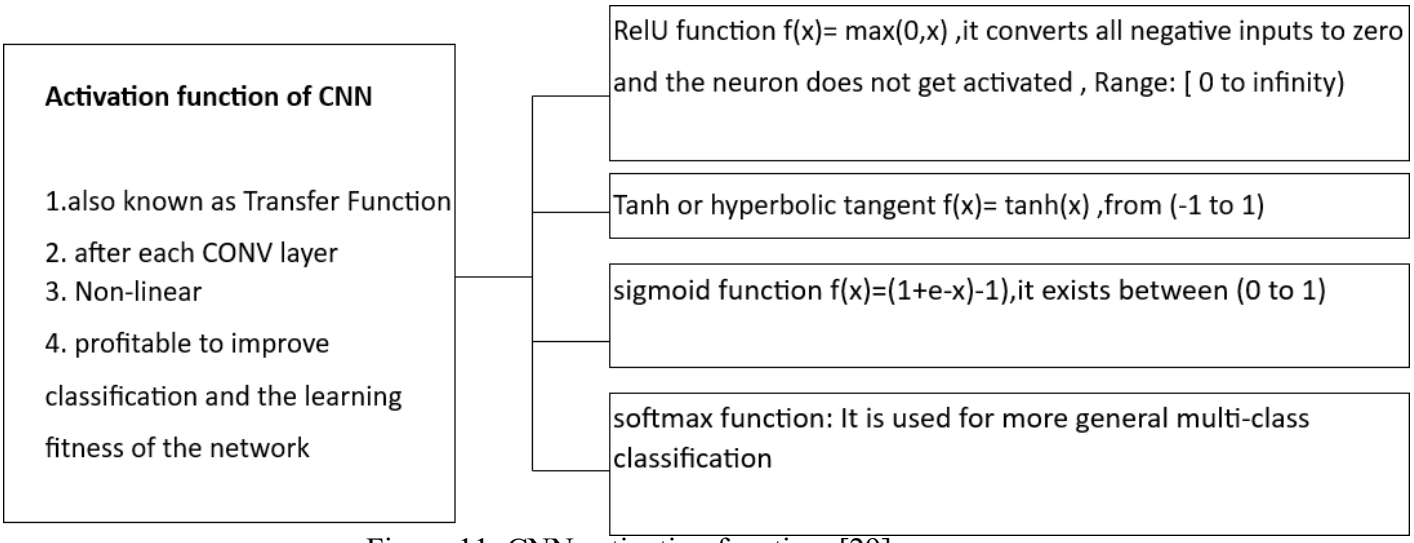

Figure 11: CNN activation function. [29]

6.6.2 Properties of CNN [31,32,33]

0

Convolution layer	**Pooling layer**	**Fully connected layers**
Filters are included to find features of an image	Reduce dimensionality	Aggregate information from final feature
The filter consists of small kernels (number of kernels)	Maximum or average area is extracted	General final classification
One bias per filters		
For every value of feature map must apply activation function	Sliding window approach	Parameters full connected, (number of nodes, activation function; usually changes depending on role of layers. RELU used for aggregating information, and SOFTMAX for producing final multi-classification)
parameters of CONV layers (size of kernels, activation function, stride, padding and regularization type and value)	Parameters of pooling, (stride and size of window).	

Table 2: illustrates the properties of CNN layers.

7. Application Domains

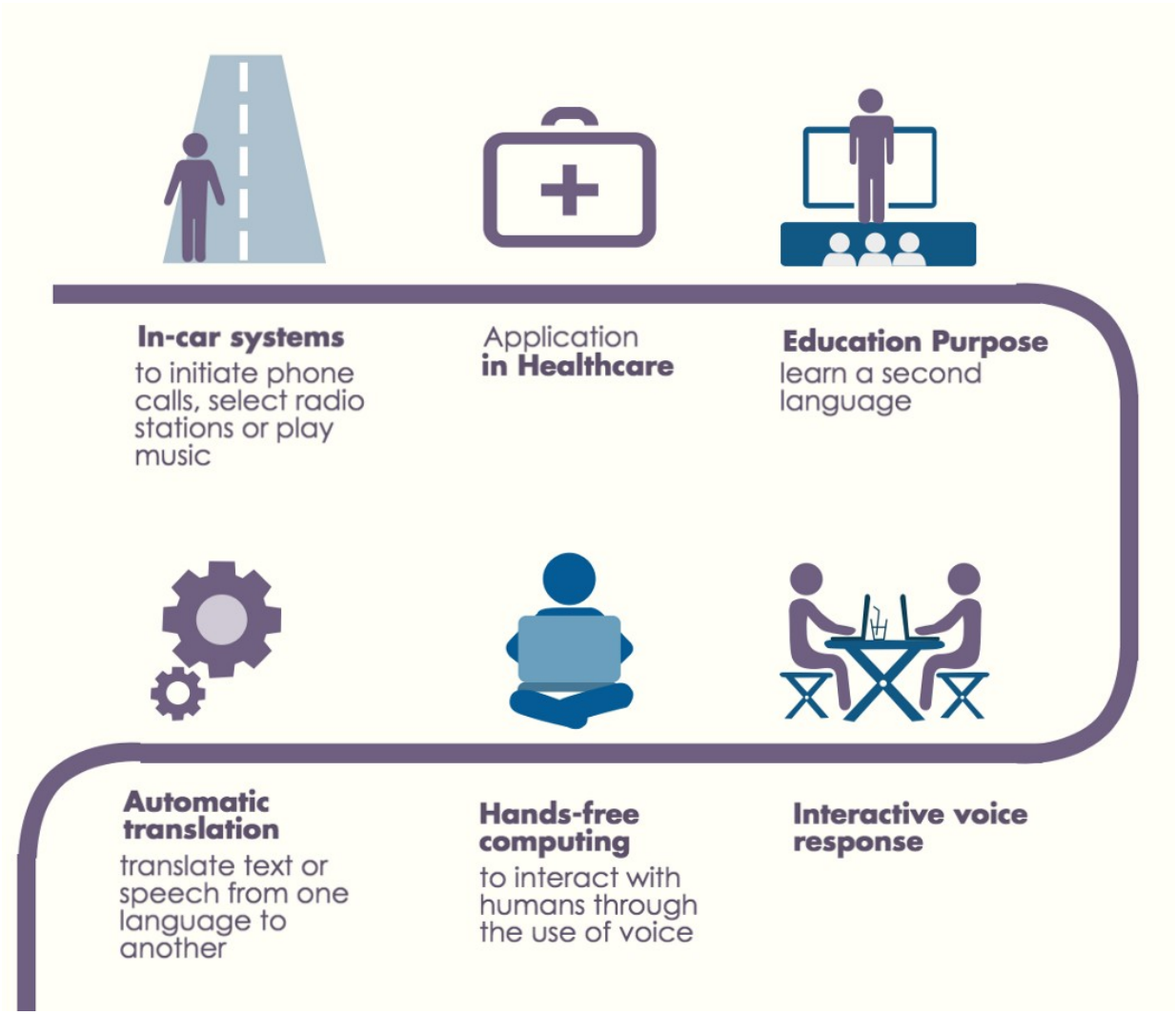

Figure 12: Application Domains.[35]

Applications of speech recognition are diverse, and we note a few [18],[36]:

- Aerospace (e.g., space exploration, spacecraft, etc.) NASA's Mars Polar Lander used speech recognition technology from Sensory, Inc. in the Mars Microphone on the Lander.
- Automatic subtitling with speech recognition.
- Automatic translation.
- Court reporting (Realtime Speech Writing).
- eDiscovery (Legal discovery).
- Education (assisting in learning a second language).

- Hands-free computing: Speech recognition computer user interface.
- Home automation (Amazon’s Alexa, Google Home etc.).
- Interactive voice response.
- Medical transcription.
- Mobile telephony, including mobile email.
- Multimodal interaction.
- People with disabilities.
- Pronunciation evaluation in computer-aided language learning applications.
- Robotics.
- Speech-to-text reporter (transcription of speech into text, video captioning, Court reporting).
- Telematics (e.g., vehicle Navigation Systems).
- User interface in telephony.
- Transcription (digital speech-to-text).
- Video games, with Tom Clancy's EndWar and Lifeline as working examples.
- Virtual assistant (e.g., Apple's Siri, Window’s Cortana).

8. Advantages and disadvantages

8.1 Voice recognition

Advantages

Voice recognition enables consumers to multitask by speaking directly to their Google Home, Amazon Alexa, or other voice recognition technology. By using machine learning and sophisticated algorithms, voice recognition technology can quickly turn your spoken word into written text. [37]

Disadvantages

While accuracy rates are improving, all voice recognition systems and programs make errors. Background noise can produce false input, which can be avoided by using the system in a quiet room. There is also a problem with words that sound alike, but that are spelled differently and have different meanings – for example, hear and here. This problem might someday be largely overcome using stored contextual information. However, this will require more RAM and faster processors than are currently available in personal computers. [37]

8.2 Speech recognition

Advantages [38]

- Increases the productivity of businesses.
- Automates the interaction between the businesses and customers.
- Adds an extra security level.
- Captures speech faster than a human can type.
- Helps people with disabilities.
- Helps control your home devices.
- Assists drivers with in-car ASR systems and more.

Disadvantages [38]

- Systems can't fully recognize speech if the speaker speaks quickly and not clearly.
- Large vocabularies are required to improve recognition accuracy.
- Each language requires separate training for ASR.
- Businesses can collect and use the user's voice data without their permission.
- Time and financial costs are high.
- ASR software consumes a lot of memory and requires a large amount of RAM.

9. Challenges and future approaches

While speech and voice recognition work differently, the two deeply intertwine to provide many cross-functional capabilities to improve our daily lives and present possibilities for the future.[8]

Given the high population density of the Asia-Pacific region, both China and India are seeing the largest growth rates, especially using voice, enabled functions in mobile banking. [39]

According to Google, 20% of queries on Google's mobile app and Android devices are voice searches, and the number is expected to grow exponentially. Google's voice assistant is now available on more than 400 million devices. [39]

The speaker recognition market was valued at USD 10.70 billion in 2020 and is expected to reach USD 27.155 billion by 2026, at a CAGR of 16.8% over the forecast period 2021 - 2026. Virtual assistants are driving this growth in retail, banking, and automotive sectors, as well as personal home use. [39]

The future of voice recognition is looking bright. Given its current global usage both in the home and on the move, it seems as though this technology will only get bigger over the next few years. [39]

The future of speech recognition is very promising. SR systems will recognize not only the words but also the emotions of a person. Speech recognition will be applied in the fields such as the aerospace industry, home automation, robotics, telematics, and video games. [38]

10. Conclusion

In this first chapter of state of art, we presented some foundations that our thesis and proposed approaches are based on. We talked about the VR and the SR, the difference between them. Furthermore, the improvement of each technology over the years has been mentioning the used methods.

We discussed the main components required for developing a SR system, also the process and techniques are presented. In addition, we have spoken about the most popular feature extraction methods.

Moreover, we presented SR based on deep learning and we explained neural networks various types for SR. We choose the convolutional neural network to implement in the second part ofthis thesis.

Chapter two:
Conceptual study

Chapter 2: Conceptual Study

1. Introduction

In the previous chapter, we have talked in general about speech recognition, the most used systems over the years, and discussed the SR system, and its involved process. Thus, we have presented an explanation detailed of SR based on deep learning and neural networks.

In this chapter, we will propose the design of our system starting with the various steps of the process for the speech recognition system. After that, we will give a precise of the proposed part of our study which is the recognition phase using the convolution neural network CNN.

2. System design

Our system is structured as in figure 13.
We presented our dataset in the form of folders, those folders represent our corpus, and each folder has a collection of samples.

We took the corpus and turn it out as **labels** and those samples as a form of vectors knowing that one samples represent a vector and each vector has 10.000 values which means every word became a group of vectors.
The samples had overcome vectors now. We saved them in a NumPy array, and we named them **features. (**This phase determinates the vector type and what word does it represent).

In the **normalization** phase, we extract the vectors between 0 and 1.
Besides, we divide our dataset as 80% for **training** and 20 % for **test**.

In the end, our **model** will finalize the work. (We presented the model in the next section).

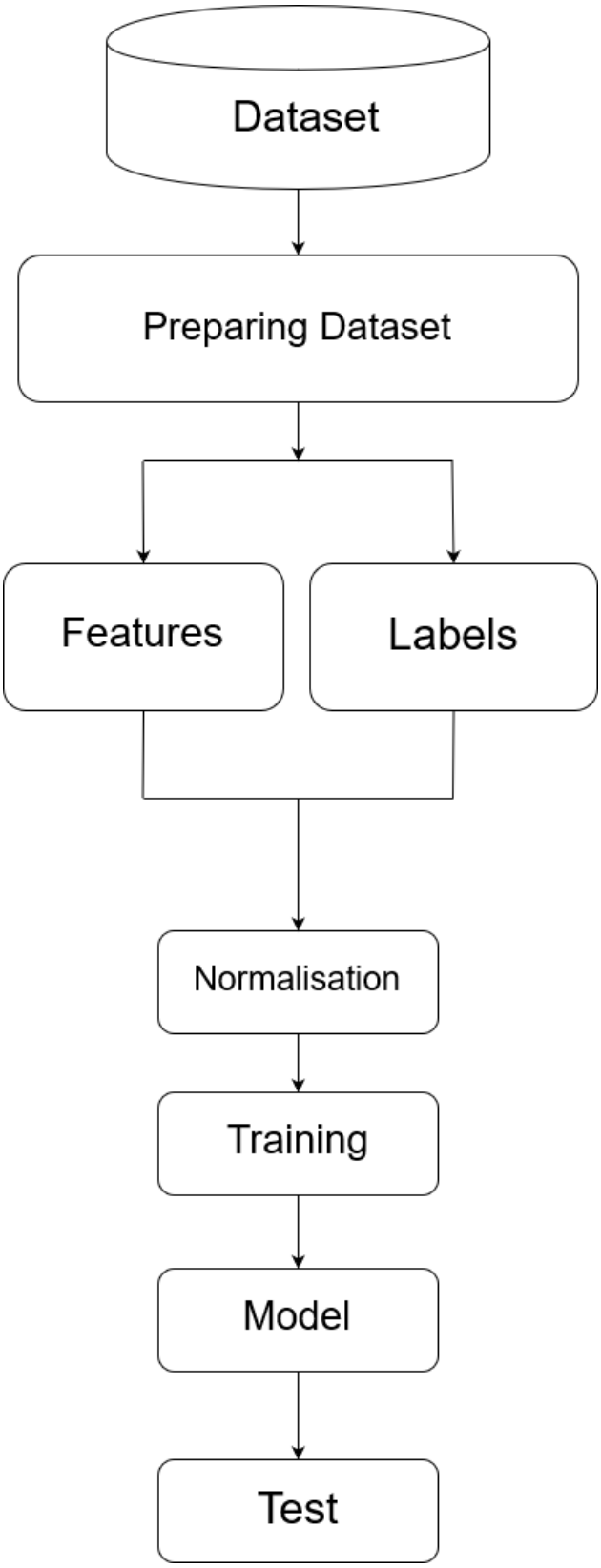

Figure 13: Our system architecture

3. Presenting Dataset

Respecting the use of the dataset “Speech Commands” in our work, credit goes to; Speech Commands Data Set v0.01_ APA-style citation: "Warden P. Speech Commands: A public dataset for single-word speech recognition, 2017. Available from “Kaggle.com”.

Audio files are part of the Speech Commands dataset and covered by the same Creative Commons BY 4.0 license.

This is version 0.01 of the dataset containing 64,727 audio files, it was released on August 3rd, 2017.

This is a set of one-second (.wav) audio files, each one containing a single spoken English word. These words are from a small set of commands and are spoken by a variety of different speakers.

The audio files are organized into folders based on the word they contain, with each directory name labeling the word that is spoken in all the contained audio files. No details were kept of any of the participant’s age, gender, or location, and random ids were assigned to every individual. The dataset is designed to help train simple machine learning models.

For some of the open-source audio collections, the audio files were collected using crowdsourcing, The objective was to accumulate instances of individuals speaking single-word commands, instead of conversational sentences, so they were prompted for individual words throughout a brief meeting.

Twenty core command words were recorded, with most speakers saying each of them five times. The core words are "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go", "Zero", "One", "Two", "Three", "Four", "Five", "Six", "Seven", "Eight", and "Nine". To help distinguish unrecognized words, there are also ten auxiliary words, which most speakers only said once, which include "Bed", "Bird", "Cat", "Dog", "Happy", "House", "Marvin", "Sheila","Tree", and "Wow".

The original audio files were collected in uncontrolled locations by people around the world. We requested that they do the recording in a closed room for privacy reasons, but didn't stipulate any quality requirements. This was by design, since we wanted examples of the sort of speech data that we're likely to encounter in consumer and robotics applications, where we don't have much control over the recording equipment or environment. The data was captured in a variety of formats, for example Ogg Vorbis encoding for the web app, and then converted to a 16-bit little-endian PCM-encoded WAVE file at a 16000 sample rate. The audio was then trimmed to a one-second length to align most utterances.

Word	Word samples number
Yes	2377
No	2375
Up	2375
Down	2359
Left	2353
Right	2367
On	2367
Off	2357
Stop	2380
Go	2372
Zero	2376
One	2370
Two	2373
Three	2356
Four	2372
Five	2357
Six	2369
Seven	2377
Eight	2352
Nine	2364
Bed	1713
Bird	1713
Cat	1733
Dog	1746

Happy	1742
House	1750
Marvin	1746
Sheila	1734
Tree	1733
Wow	1745

Table 3: The used words and their sample number for model training.

4. Model architecture

4.1 CNN model

Defining model archetecture

```
model = keras.models.Sequential()
#model.add(normalization)
model.add(keras.layers.Input(shape=(10000,1)))

model.add(keras.layers.Conv1D(8,13,activation='relu', padding='valid',strides=1) )
model.add(keras.layers.MaxPooling1D(3))
model.add(keras.layers.Dropout(0.3))
model.add(keras.layers.Conv1D(10,20,activation='relu', padding='valid',strides=1))
model.add(keras.layers.MaxPooling1D(3))

model.add(keras.layers.Conv1D(10,10,activation='relu', padding='valid',strides=1))
model.add(keras.layers.MaxPooling1D(3))
model.add(keras.layers.Dropout(0.3))

model.add(keras.layers.Conv1D(15,10,activation='relu', padding='valid',strides=1))
model.add(keras.layers.MaxPooling1D(3))

model.add(keras.layers.Flatten())
#model.add(keras.layers.Dense(256, activation='relu'))
model.add(keras.layers.Dense(units=128,activation='relu'))

model.add(keras.layers.Dense(units=64,activation='relu'))
model.add(keras.layers.Dense(units=30))
optimizer = keras.optimizers.Adam()
loss = keras.losses.SparseCategoricalCrossentropy(from_logits=True)
model.compile(optimizer=optimizer,loss=loss, metrics=['accuracy'])


model.summary()
```

Figure 14: CNN model architecture.

The figure shows the architecture of our CNN model and the used layers.
We will explain in this section the role of each type of layer:

Convolutional layer 1D: the first layer defines several filters (or also called feature detector) of height 10 (also called kernel size). Only defining one filter would allow the neural network to learn one single feature in the first layer. Following the same logic for the other convolutional layers as the first layer in order to learn higher level features.

Maxpooling: it is a layer that is often used after a CNN layer in order to reduce the complexity of the output and prevent overfitting of the data.

Dropout: this layer randomly sets input units to 0 with a frequency of rate at each step during training time, which helps prevent overfitting. the network becomes less sensitive to react to smaller variations in the data. Therefore, it should further increase our accuracy on unseen data.

Dense layer (fully connected): A linear operation in which every input is connected to every output by a weight (so there are n_inputs * n_outputs weights - which can be a lot). Generally followed by a non-linear activation function.

4.2 model configuration and number of parameters

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 9988, 8)           112

max_pooling1d (MaxPooling1D  (None, 3329, 8)           0
)

dropout (Dropout)            (None, 3329, 8)           0

conv1d_1 (Conv1D)            (None, 3310, 10)          1610

max_pooling1d_1 (MaxPooling  (None, 1103, 10)          0
1D)

conv1d_2 (Conv1D)            (None, 1094, 10)          1010

max_pooling1d_2 (MaxPooling  (None, 364, 10)           0
1D)

dropout_1 (Dropout)          (None, 364, 10)           0

conv1d_3 (Conv1D)            (None, 355, 15)           1515

max_pooling1d_3 (MaxPooling  (None, 118, 15)           0
1D)

flatten (Flatten)            (None, 1770)              0

dense (Dense)                (None, 128)               226688

dense_1 (Dense)              (None, 64)                8256

dense_2 (Dense)              (None, 30)                1950

=================================================================
Total params: 241,141
Trainable params: 241,141
Non-trainable params: 0
```

Figure 15: the configuration of our model.

In this section, we proposed a CNN based solution for speech recognition. The model that we present is composed of four convolutional layers, four Maxpooling layers, two dropout layers, and three layers of fully connected layers. the input of our model is a one-dimensional feature vector that contains 10000 features. Each convolution layer has several filters.
The ReLU activation function is applied in Each convolution Layer as well as in the first and the second Layer of the Fully connected Layers.

Maxpooling is applied after each convolutional operation to reduce the size of the feature map. A dropout layer is used after the first Maxpooling layer to prevent our model from overfitting. The dropout layer is only applied in the training phase.

The last layer of the proposed method is a fully connected layer which represents the output of our model, SoftMax activation function is applied to the output layer to detect which output is a spoken word. We used categorical cross-entropy as a loss function and an Adam optimizer to train our model.

We note that our model contains 241 parameters.

5. Conclusion

In this chapter, we have presented the design of our system displaying the used process. We demonstrate in detail, the architecture of our model CNN. We also presented the discussion and comparison of the various obtained results.

The experiment shows that the CNNs are quite promising in dealing with complex tasks such as speech recognition.

Our model reached 79.07% accuracy in the test phase.

The reliability of word recognition these days is high, but it will take time to achieve 100 percent accuracy.

Chapter 3:

Implementation

Chapter 3: Implementation and obtained results

1. Introduction

In this chapter, we will start first with a list of the chosen tools for our system development. Then, we will present the system interface and will display some screenshots of the test phase.

2. Representation of the development tools

2.1 Physical environment

In order to realize our system, we utilize this hardware:

The test phase of our system is carried out on a desktop PC which is having the following characteristics:

MB: ASUSTeK INC. H170M-E D3	
CPU: Intel (R) Core (TM) i3-6100 -Frequency: 3.7GHz	
RAM: Hyper-FuryX 8.00GB ddr3	
GPU: Nvidia Geforece GTX 750 Ti 2gb	

Table 4: Desktop Hardware

Hard disk size 1TB HDD.

In order to train our model, we used Google Colab, and to be able to import and mount our selected dataset, we used Google Drive which we saved in it.

Google Colab: Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser and is especially well suited to machine learning, data analysis, and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs [40] but it is limited to daily use.

Google Drive:

Drive is cloud-native, which eliminates the need for local files and can minimize risk to your devices. Drive can provide encrypted and secure access to your files. Files shared with you can be proactively scanned and removed when malware, spam, ransomware, or phishing is detected. [41]

2.2 Software and libraries used in the implementation

Operating system: 64-bit Windows 10.

VS Code: Visual Studio Code has a high productivity code editor which, when combined with programming language services, gives you the power of an IDE and the speed of a text editor.[42]

Python: Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed. [43]

TensorFlow: TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.[44]

Keras: Keras is an API designed for human beings, not machines. Keras follows best practices for reducing cognitive load: it offers consistent & simple APIs, it minimizes the number of user actions required for common use cases, and it provides clear & actionable error messages. It also has extensive documentation and developer guides.[45]

NumPy: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.[46]

Librosa: Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation, Speech Recognition. It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals and also do the feature extractions in it using different signal processing techniques.[47]

SKlearn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon **NumPy, SciPy** and **Matplotlib**.[48]

Cuda: (Compute Unified Device Architecture) is a parallel computing platform and programming model created by NVIDIA. With more than 20 million downloads to date, CUDA helps developers speed up their applications by harnessing the power of GPU accelerators.[49]

Cuda CNN: The NVIDIA CUDA® Deep Neural Network library (cuDNN) is a GPU accelerated library of primitives for deep neural networks. cuDNN provides highly tuned implementations for standard routines such as forward and backward convolution, pooling, normalization, and activation layers.[50]

PySimpleGUI: PySimpleGUI is a python library that wraps tkinter, Qt (pyside2), wxPython and Remi (for browser support), allowing very fast and simple-to-learn GUI programming. PySimpleGUI defaults to using tkinter, but the user can change to another supported GUI library by just changing one line.[51]

3. Discussion and comparison of the obtained results

3.1 Discussion of the obtained results:

We have trained and tested our model.
In this section, we will show the obtained results with different numbers of epochs:

For 15 epochs:

Epochs number	Training		Test	
	loss	accuracy	val_loss	val_accuracy
15	0.6140	0.8085	0.8488	0.7405

Table 5: Ilustre the results using 15 epochs.

Accuracy (15)

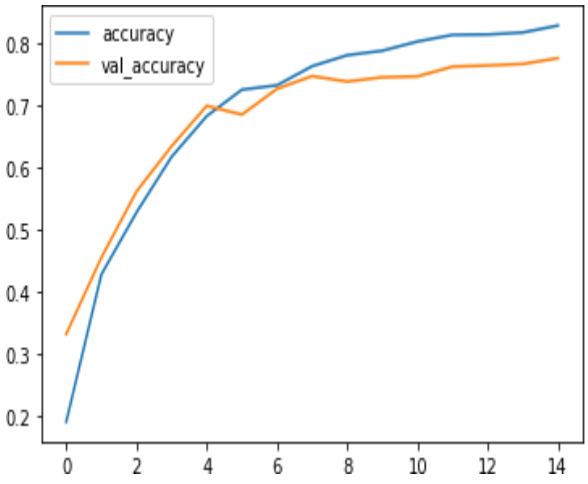

Figure 16: Chart of accuracy (training) and and validation accuracy (test).

Loss (15)

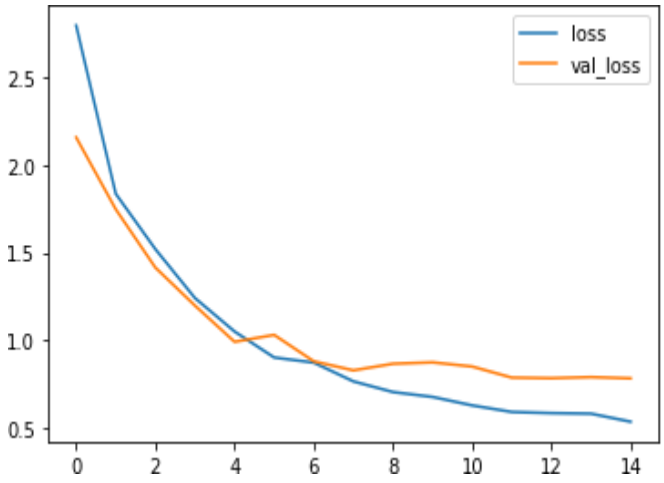

Figure 17: Chart of loss (training) and validation loss (test).

From the figures 16 and 17, above of 15 epochs.

The left side of the graph (figure 16) represents the accuracy of the training with the blue chart when the orange chart represents the validation accuracy (test)

The right side of the graph (figure17), the orange chart represents the validation loss (test) and the blue chart represents the training loss.

This the first trial, if you look at this both figures (16,17), you will notice that both training accuracy and validation accuracy increase and so are the things with training loss and validation loss which a good starting phase for just 15 epochs.

For 35 epochs:

Epochs number	Training		Test	
	loss	accuracy	val_loss	val_accuracy
35	0.3439	0.8896	0.8940	0.7649

Table 6: Ilustre the results using 35 epochs.

Accuracy (35)

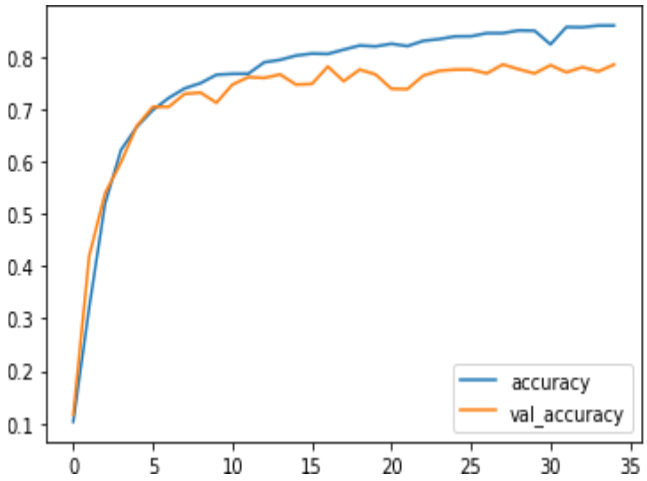

Figure 18: Chart of accuracy (training) and validation accuracy (test).

Loss (35)

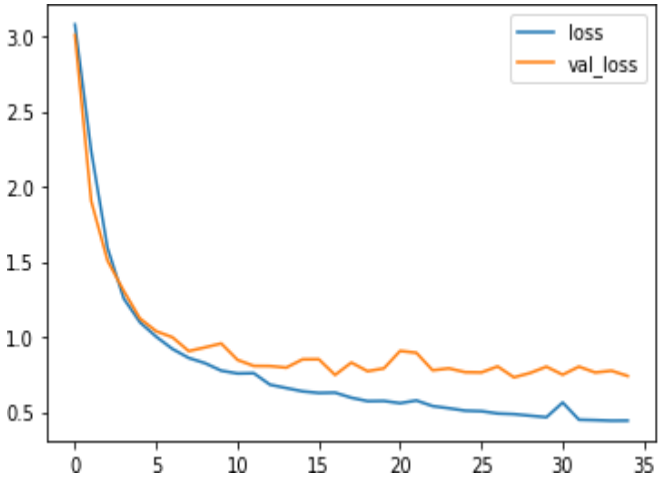

Figure 19: Chart of loss (training) and validation loss (test).

The representation of the charts and the colors remain the same as the previous phase (15 epochs).

We move on to the second one now.

After we add more number epochs which is 35, we notice that the training loss decreases however, validation loss slightly increases. In other hand, the training accuracy, and the validation accuracy increase. Figures (18, 19)

Besides, stabilization starts to slightly appears in this trial (35).

For 55 epochs:

Epochs number	Training		Test	
	loss	accuracy	val_loss	val_accuracy
55	0.2358	0.9241	0.9261	0.7842

Table 7: Ilustre the results using 55 epochs.

Accuracy (55)

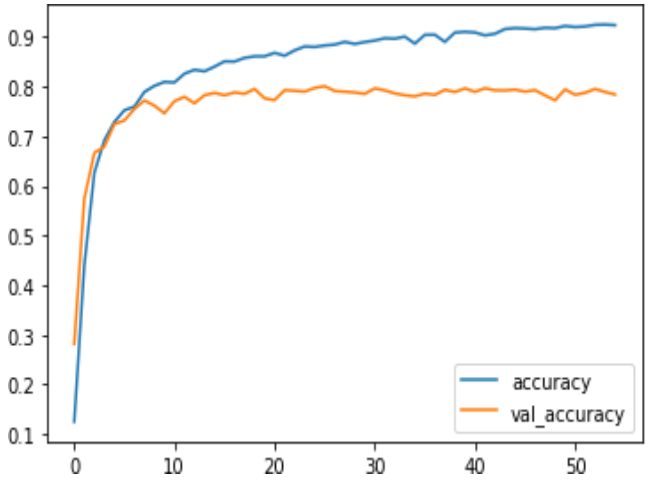

Figure 20: Chart of accuracy (training) and validation accuracy (test).

Loss (55)

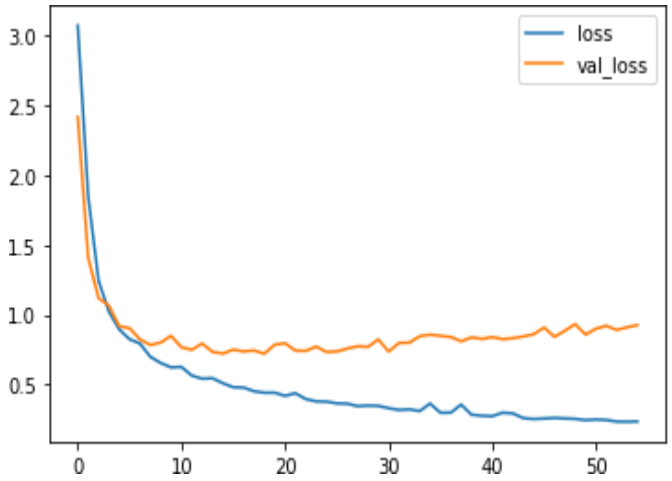

Figure 21: Chart of loss (training) and validation loss (test).

For the 55 epochs, we notice that the training loss decreases to 0.2 but, the validation loss increases a bit more. (figure 21)

The training accuracy increase in the interval {0.10}, however, it almost stabilizes in the interval of {10.55}, its increasing rise partially in .{10.40}

the validation accuracy increases in the interval {0,15} then it's quite stabilized in the interval {15,55}. (figure 20)

For 85 epochs:

Epochs number	Training		Test	
	loss	accuracy	val_loss	val_accuracy
85	0.1955	0.9384	0.9679	0.7907

Table 8: Ilustre the results using 85 epochs.

Accuracy (85)

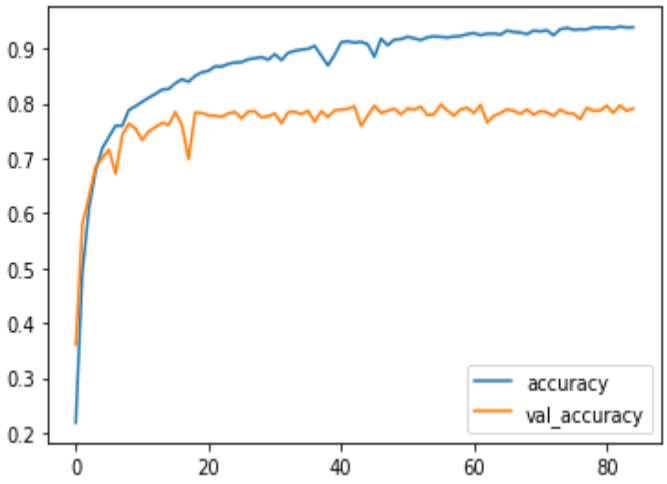

Figure 22: Chart of accuracy (training) and validation accuracy (test).

Loss (85)

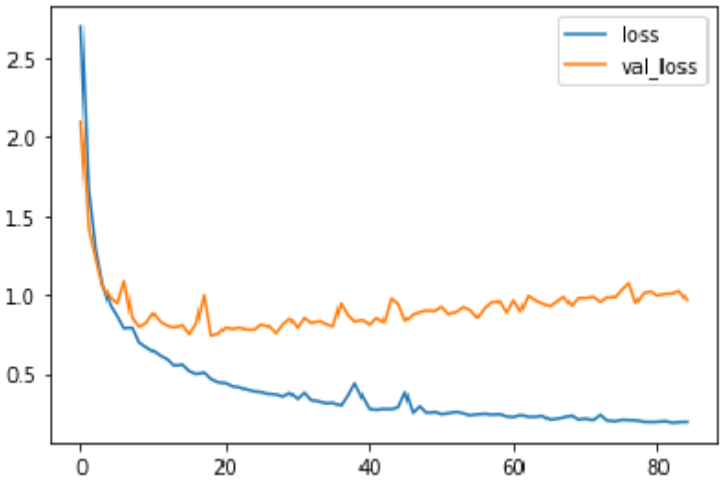

Figure 23: Chart of loss (training) and validation loss (test).

In the last trial, 85 epochs, we notice that the training loss increases more in the whole process unlike the validation loss, it started to decrease then it increased a little. (figure 23), comparing to the previous trials the val_loss is slightly higher.

For the training accuracy, it increases normally with bigger number of epochs, and similar results for the validation accuracy, (figure 22) it also increases which is totally logic and it is the goal for adding more epochs number.

3.2 Comparison of the obtained results:

Epochs number	Training		Test	
	loss	accuracy	val_loss	val_accuracy
15	0.6140	0.8085	0.8488	0.7405
35	0.3439	0.8896	0.8940	0.7649
55	0.2358	0.9241	0.9261	0.7842
85	0.1955	0.9384	0.9679	0.7907

Table 9: the obtained results by training our model with different numbers of epochs.

We notice that each time we pick more numbers of epochs, the Training accuracy decreases and so is the test validation accuracy, it decreases itself either.
For the loss of Training, it increases with more epochs number except for the validation loss of Test, it increases a bit with more epoch numbers which means our model needs more training.

Comparing of our obtained results with ANN model:

In the table below we compare our CNN model results with ANN model with the same dataset. The ANN model that we trained in order to compare

Model	Epochs number	Training		Test	
		loss	accuracy	val_loss	val_accuracy
CNN	15	0.6140	0.8085	0.8488	0.7405
	35	0.3439	0.8896	0.8940	0.7649
	55	0.2358	0.9241	0.9261	0.7842
	85	0.1955	0.9384	0.9679	0.7907
ANN	15	1.2408	0.6323	7.1135	0.067
	35	0.5440	0.8515	16.5503	0.748
	55	0.43	0.8888	25.8928	0.0733
	85	0.3330	0.9138	31.9732	0.0671

Table 10: Comparison between the obtained results of ANN and CNN models

From basics knowledge, there is no doubt that ANN would give bitter results comparing with CNN.

At any trial of epochs numbers, nevertheless there might be a logic result for the loss of training and the accuracy of training however for the test, the validation loss is unacceptable, it is too high, and it gets higher with a greater number of epochs. Furthermore, the validation accuracy is too low that much it considered neglected.

ANN is one of the simplest types of neural networks when CNN is one of the most popular neural networks.

In the summary of strength, CNN tends to be a more powerful and accurate way speech recognition when ANN is dominant for problems where datasets are limited, mentioning that our used dataset is wildly large.

4. Representing our system interface:

The figure 24 below displays the home page of our system.

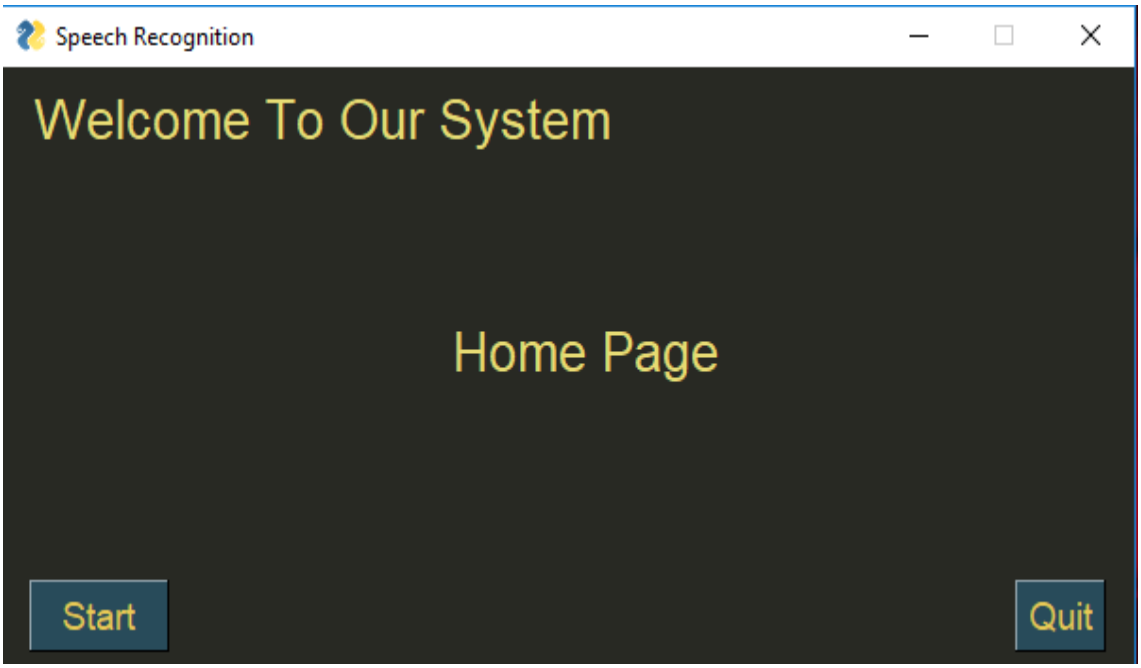

Figure 24: Home Page.

The **Quit** button is only to quit the system's interface.

After we click on the **Start** button, it takes us to the next window (figure25).

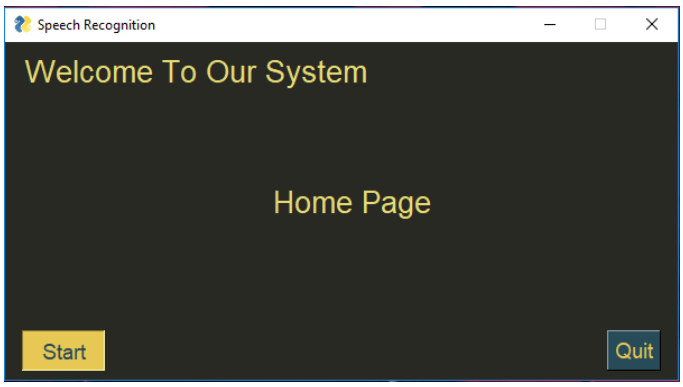

Figure 25: from home page to test page.

The (figure 26) represents the second window, the Test Page, which we will apply the Test phase here.

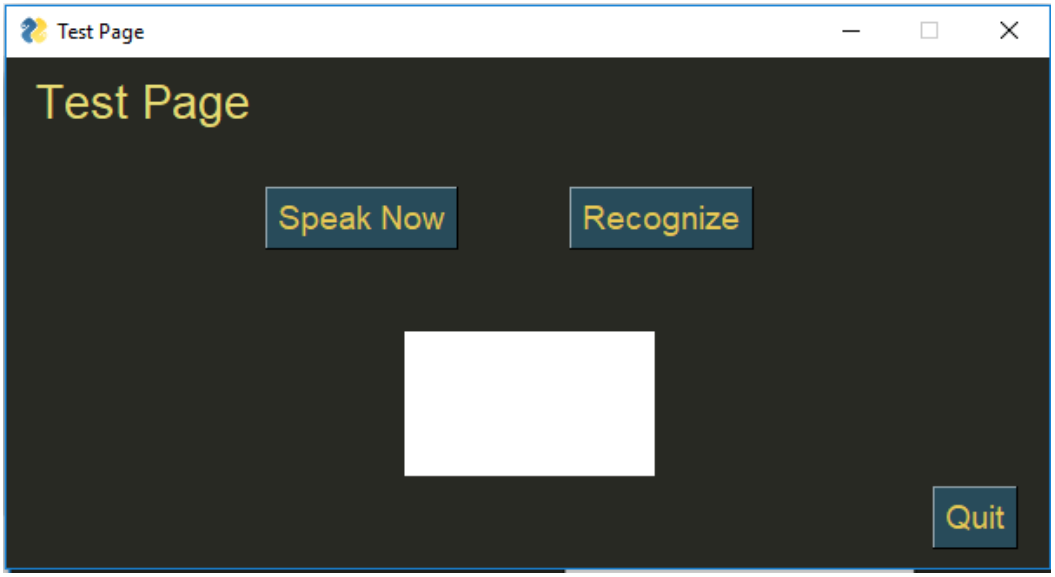

Figure 26: Test Page.

5. Test

In the second window, in order to apply the test phase, by clicking on the button Speak Now, we give our system one word from our corpus (30 words) providing that in one second.(figure 27)

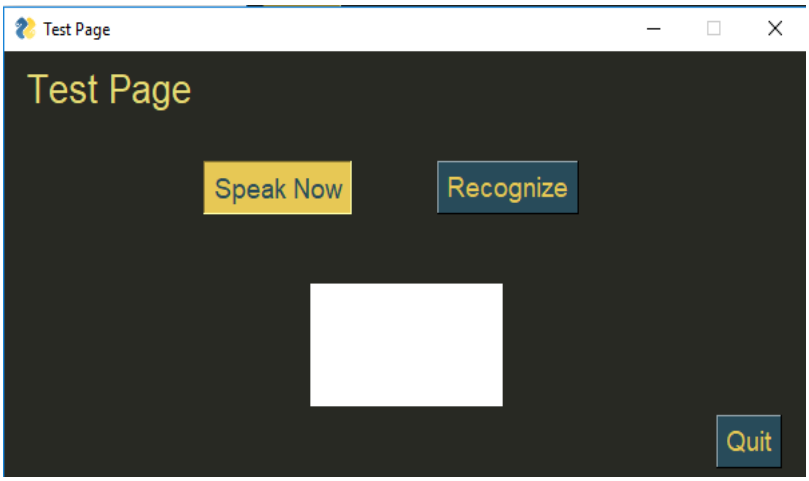

Figure 27: Input Phase.

We chose the word “UP” as the first trial to be tested as it appears in the label (figure 28).

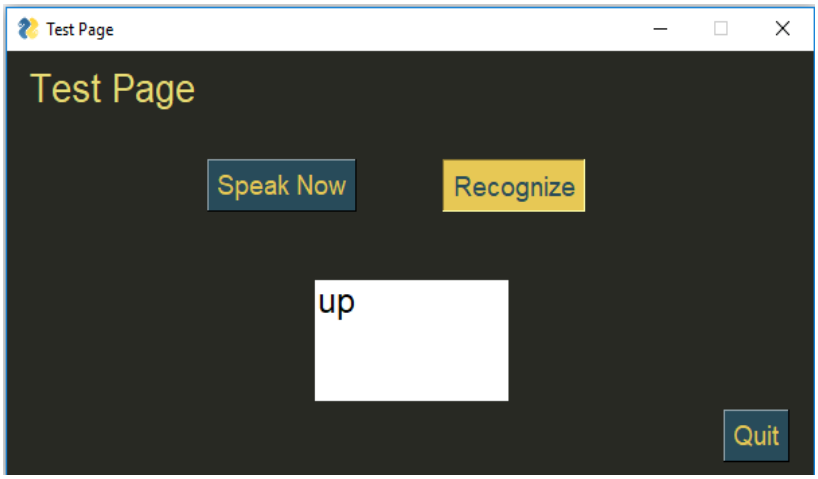

Figure 28: Output phase (text).

The next figures display (figures ;29,30,31,32,33) some various words from our corpus which have been tested multiple times.

“CAT”, “SEVEN”, “SHEILA”, “BIRD”, “HAPPY”.

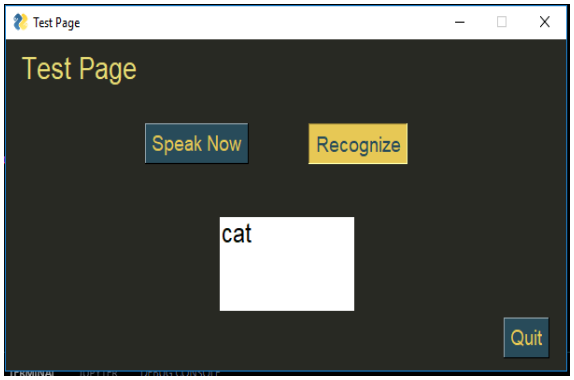

Figure 29: Output “Cat”.

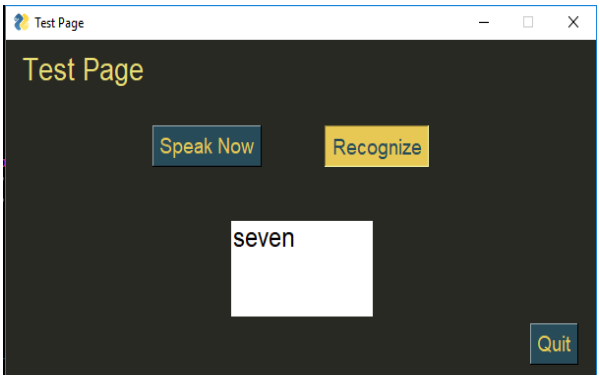

Figure 30: Output “Seven”.

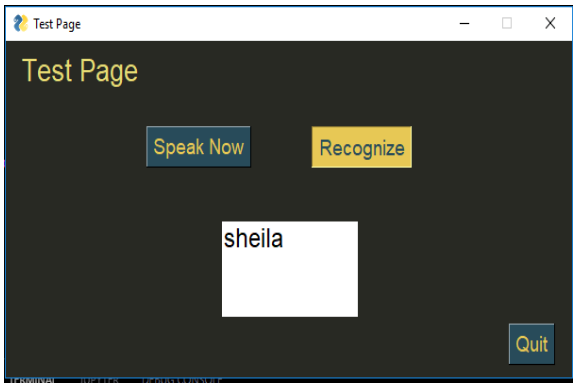

Figure 31: Output “Sheila”.

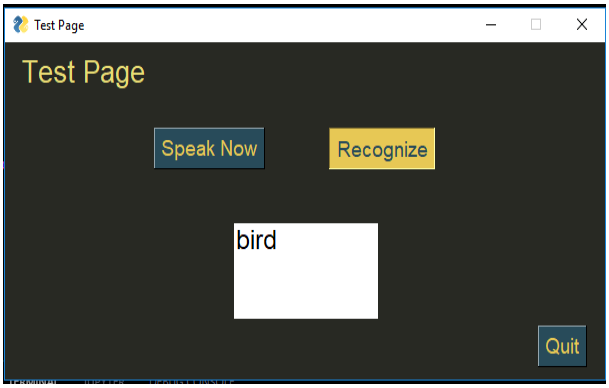

Figure 32: Output “Bird”.

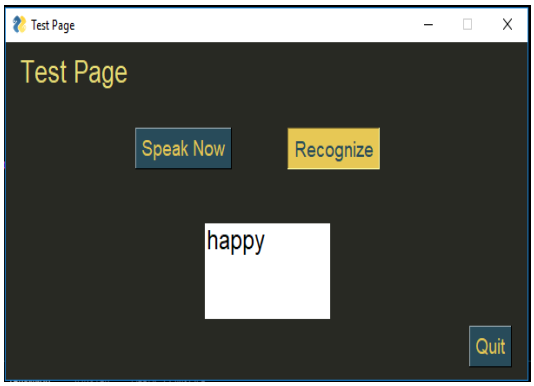

Figure 33: Output “Happy”.

6. Conclusion

In this chapter we have presented our system interface, the implementation and we demonstrate that our system allows to predict spoken words (existed in our dataset) in the test phase.

General Conclusion and Perspectives

General Conclusion and Perspectives

Until today, speaker recognizers and speech recognition in particular present a very big challenge, despite the efforts and intensive work carried out in this area, no speech recognition system is considered 100% reliable.
Speech is a fundamental way of communicating between people. Our work is about convolutional neural networks which are the current way of perceiving speech. Rather than conventional methodology, it doesn't require any insights. A Speech Recognition system must incorporate the four phases: Speech, Speech PreProcessing, Feature Extraction, Speech Classification, and Recognition; procedures as portrayed.

In this project, we have discussed the main components required for developing a speech recognition system, also the process and techniques are presented. Besides, we discussed SR based on deep learning and we explained neural networks various types for SR then we chose the convolutional neural network to implement. We have introduced the convolutional neural networks by presenting the different types of layers used for the whole process. We implemented a CNN model with an architecture that has four convolutional layers, four Maxpooling layers, two dropout layers, and three layers of fully connected layers. Subsequently, we applied for this model tests which consist in changing the number of epochs each time, and display the results at the end. The implementation was done with the python programming language and we used libraries to facilitate the task of creating our system and for the acceleration of training and finally we finished with a summary and comparison table of the tests carried out.
The comparison of the results found showed that the number of epochs is an important factor for obtaining better results. The essential of speech recognition is discussed, also its new advancement is explored.
Speech recognition utilizing convolutional neural networks is arising field now daily. Text to speech and speech to text are two applications that are helpful for impaired individuals. Paper basically centers around speech recognition of one language, which is English.
In simple terms, CNN is an important and efficient method, that perform exceptionally well.

Finally, we can see things further with the use of CNNs, we can take this system step ahead to improve our model for a better accuracy, also to make it be &able to recognize continuous words with a larger dataset.

References

[1] Pioneering Speech Recognition_ https://www.ibm.com/ibm/history/ibm100/us/en/ Accessed.05.2022

[2] J Felton, W. A., Miller, G. L., & Milner, J. M. (1984). The UNIX system: A UNIX system implementation for System/370. *AT&T Bell Laboratories technical journal*, *63*(8), 1751-1767.

[3] https://www.dictate.it/frontend/news/speech-recognition-from-audrey-to-alexa-a-brief-history/ Accessed.05.2022

[4] https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen _ Accessed.05.2022.

[5] Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*(3), 251-272.

[6] Averbuch, A., Bahl, L., Bakis, R., Brown, P., Daggett, G., Das, S., ... & Wilkens, H. (1987, April). Experiments with the TANGORA 20,000 word speech recognizer. In *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 12, pp. 701-704). IEEE.

[7] Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*, 67.

[8] "The Difference Between Speech and Voice Recognition_by Laura Tate", https://www.kardome.com/blog-posts/difference-speech-and-voice-recognition. Accessed.05.2022.

[9] Alvarez, Raziel, and Yishay Carmiel. 2017. "Kaldi now offers TensorFlow integration." Google Developers.

[10] Poornima, S. (2016). Basic characteristics of speech signal analysis. International Journal of Innovative Research and Development, 5(4), 169-173.

[11] Md.Mijanur Rahman, Md.A1-Amin Bhuiyan, "Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.3, No.11,311.

[12] http://www.physicsclassroom.com/class/sound/Lesson-2/ -Frequency_ Accessed.05.2022.

[13] Rutledge, J. C. (1995). Fundamentals of speech recognition, by lawrence rabiner and bing-hwang juang. ANNALS OF BIOMEDICAL ENGINEERING, 23, 526-526.

[14] Peterson, Casey. 2015. "A Guide to Speech Recognition Algorithms.

[15] Jurafsky, Daniel and James H. Martin. 2019. "Chapter 9: Automatic Speech Recognition." In: Speech and Language Processing, Third Edition draft.

[16] Hrugved Pawar1, Nikki Gaikwad2, Shambhavi Kulkarni3.(2020) "A Study of Techniques and Processes Involved in Speech Recognition System"._International Research Journal of

Engineering and Technology (IRJET).

[17] Alim, Sabur Ajibola and Nahrul Khair Alang Rashid. 2018. "Some Commonly Used Speech Feature Extraction Algorithms." IntechOpen.

[18] https://devopedia.org/speech-recognition,. Accessed.05.2022.

[19] Huang, X., & Deng, L. (2010). An Overview of Modern Speech Recognition. Handbook of natural language processing, 2, 339-366.

[20] Platek, Ondrej. 2014. "Automatic speech recognition using Kaldi". Institute of Formal and Applied Linguistics, Charles University in Prague.

[21] Dominguez, Javier Gonzalez. 2015. "A Real-Time End-to-End Multilingual Speech Recognition Architecture". IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL 9, NO 4.

[22] Dat Tat Tran, Fuzzy Approaches to Speech and Speaker Recognition , A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra.

[23] Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.

[24] McLellan, Charles. 2016. "How we learned to talk to computers, and how they learned to answer back". Tech Republic.

[25] Luhach, A. K., Kosa, J. A., Poonia, R. C., Gao, X. Z., & Singh, D. (Eds.). (2020). First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019. Springer.

[26] a neural network system in smartphones for automatic speech recognition _2018.

[27] https://viso.ai/deep-learning/deep-neural-network-three-popular-types.Accessed.05.2022.

[28] Giannetti, J. (2021). Deep learning methods for speech to text systems.

[29] Alsobhani, A., ALabboodi, H. M., & Mahdi, H. (2021, August). Speech Recognition using Convolution Deep Neural Networks. In Journal of Physics: Conference Series (Vol. 1973, No. 1, p. 012166). IOP Publishing.

[30] Saitoh T., Zhou Z., Zhao, G., Pietikäinen M. (2016) Concatenated frame image based cnn for visual speech recognition. In: Asian Conference on Computer Vision. p. 277-289.15- Phung, Son Lam, and Abdesselam Bouzerdoum. "Visual and Audio Signal Processing Lab University of Wollongong".

[31] Nanni L., Costa Y. M., Aguiar R. L., Mangolin, R. B., Brahnam S., Silla C. N. (2020) Ensemble of convolutional neural networks to improve animal audio classification. EURASIP Journal on Audio, Speech, and Music Processing, 1-14.

[32] Patel S. (2020) A Comprehensive Analysis of Convolutional Neural Network Models.International Journal of Advanced Science and Technology, 29(4), 771-777.

[33] Kubanek M., Bobulski J., Kulawik, J. (2019) A method of speech coding for speech recognition using a convolutional neural network. Symmetry, 11(9), 1185.

[34] Nwankpa C., Ijomah W., Gachagan, A., Marshall, S. (2018) Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378.

[35] Application of Speech Recognition Technology in Speech-Related Disabilities: An Analysis and Forecast_2016 Bian Xiaobo, Kanika Singhal, Sabrina Shilun Fang, Shankar Shyam Krishna, Vaidehi Patel, Wen-Chien (Jenny) Hsiao. Dr. Jayson Parker, Human Biology, University of Toronto.

[36] Applications of Speech Recognition_2019.

https://www.getsmarter.com/blog/market-trends/applications-of-speech-recognition. Accessed.05.2022..

[37] voice recognition (speaker recognition) By Jesse Scardina,2018.

https://www.techtarget.com/searchcustomerexperience/definition/voice-recognition-speaker-recognition. Accessed.05.2022.

[38] https://recfaces.com/articles/what-is-voice-recognition_ Accessed.05.2022.

[39] The future of voice recognition.

https://www.elasticpath.com/blog/the-future-of-voice-recognition-infographic._ Accessed.05.2022.

[40] https://research.google.com/colaboratory/faq.html_ ._ Accessed.06.2022.

[41] https://www.google.com/drive/ ._ Accessed.06.2022.

[42] https://code.visualstudio.com/docs/editor/editingevolved _ Accessed.06.2022.

[43] https://www.python.org/doc/essays/blurb/ _Accessed.06.2022.

[44] https://www.tensorflow.org/ _Accessed.06.2022.

[45] https://keras.io/ _Accessed.06.2022.

[46] https://numpy.org/doc/stable/user/whatisnumpy.html _Accessed.06.2022.

[47] https://www.geeksforgeeks.org/how-to-install-librosa-library-in-python/ _Accessed.06.2022.

[48] https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm

_Accessed.06.2022.

[49] https://blogs.nvidia.com/blog/2012/09/10/what-is-cuda-2/ _Accessed.06.2022.

[50] https://developer.nvidia.com/cudnn _Accessed.06.2022.

[51] https://wiki.python.org/moin/PySimpleGUI _Accessed.06.2022.

Summary

For the last many years, a gigantic measure of research has been done on the utilization of machine learning for speech processing applications, particularly speech recognition. In any case, in a couple of years, researchers have zeroed in on using deep learning for discourse-related applications. This new area of machine learning has yielded much better outcomes when contrasted with others in an assortment of utilizations including speech and accordingly, turned into an exceptionally appealing area of exploration while deep learning initially emerged as another area of machine learning, for speech applications.

In this research paper, we present a speech recognition system using a convolutional neural network that is able to recognize words, we note that our work is 79% accurate.

Keywords:

Natural Language Processing, Machine Learning, Deep Learning, Speech Recognition, Convolutional Neural Network.

Abstract

Au cours des dernières années, de nombreuses recherches ont été menées sur l'utilisation de l'apprentissage automatique pour les applications de traitement de la parole, en particulier la reconnaissance vocale. Quoi qu'il en soit, en quelques années, les chercheurs se sont concentrés sur l'utilisation de l'apprentissage en profondeur pour des applications liées au discours. Ce nouveau domaine de l'apprentissage automatique a donné de bien meilleurs résultats lorsqu'il est comparé à d'autres dans un assortiment d'utilisations, y compris la parole et, par conséquent, s'est transformé en un domaine d'exploration exceptionnellement attrayant, tandis que l'apprentissage en profondeur est initialement apparu comme un autre domaine de l'apprentissage automatique, pour les applications vocales.

Dans ce document de recherche, nous présentons un système de reconnaissance de la parole utilisant un réseau de neurones convolutifs capable de reconnaître des mots, nous notons que notre travail est précis à 79%.

Keywords : Traitement du langage naturel, apprentissage automatique, apprentissage en profondeur, reconnaissance vocale, réseau de neurones convolutifs.

ملخص

على مدى السنوات العديدة الماضية ، تم إجراء مقياس هائل للبحث حول استخدام التعلم الآلي لتطبيقات معالجة الكلام ، وخاصة التعرف على الكلام. على أي حال ، في غضون عامين ، ركز الباحثون على استخدام التعلم العميق للتطبيقات المتعلقة بالخطاب. لقد أسفر هذا المجال الجديد من التعلم الآلي عن نتائج أفضل بكثير عند مقارنته مع الآخرين في مجموعة متنوعة من الاستخدامات بما في ذلك الكلام ، وبالتالي ، تحول إلى منطقة استكشاف جذابة بشكل استثنائي بينما ظهر التعلم العميق في البداية كمجال آخر للتعلم الآلي ، لتطبيقات الكلام. في ورقة البحث هذه ، نقدم نظام التعرف على الكلام باستخدام شبكة عصبية تلافيفية قادرة على التعرف على الكلمات ، ونلاحظ أن عملنا دقيق بنسبة. ٪79

الكلمات المفتاحية:

معالجة اللغة الطبيعية ، التعلم الآلي ، التعلم العميق ، التعرف على الكلام ، الشبكة العصبية التلافيفية.