A

Seminar Report

on

**Voice Recognition**

**Bachelor of Technology**

**Information Technology Engineering**

by

**Reenu (LIT-993-2k16)**

Under supervision of

**Mrs. Parul Gupta**



Department of Computer Engineering

YMCA University of Science & Technology

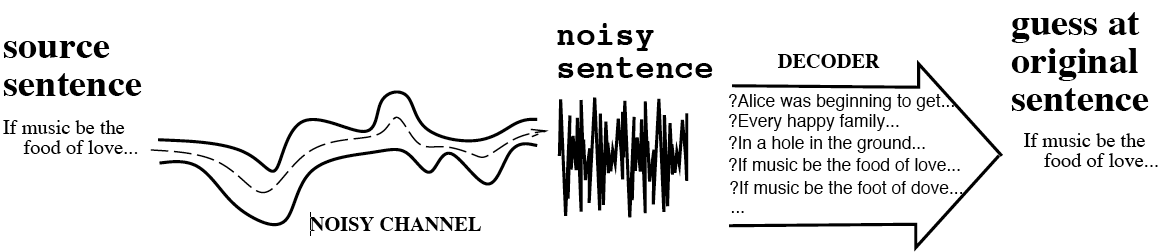
Faridabad-121006

Dec 2019

**Voice Recognition**

An introduction

*Speech recognition basically means talking to a computer, having it recognize what we are saying*. This process fundamentally functions as a pipeline that converts PCM (Pulse Code Modulation) digital audio from a sound card into recognized speech. Speech recognition technology has evolved for more than 40 years, spurred on by advances in signal processing, algorithms, architectures, and hardware. During that time it has gone from a laboratory curiosity to an art, and eventually to a full-fledged technology that is practiced and understood by a wide range of engineers, scientists, linguists, psychologists, and systems designers. Over those 4 decades, the technology of speech recognition has evolved, leading to a steady stream of increasingly more difficult asks which have been tackled and solved.



The figure shows a block diagram of a typical integrated continuous speech recognition system. Interestingly enough, this generic block diagram can be made to work on virtually any speech recognition task that has been devised in the past 40 years, i.e. isolated word recognition, connected word recognition, continuous speech recognition, etc. The feature analysis module provides the acoustic feature vectors used to characterize the spectral properties of the time-varying speech signal. The word level acoustic match module evaluates the similarity between the input feature vector sequence (corresponding to a portion of the input speech) and a set of acoustic word models for all words in the recognition task vocabulary to determine which words were most likely spoken.

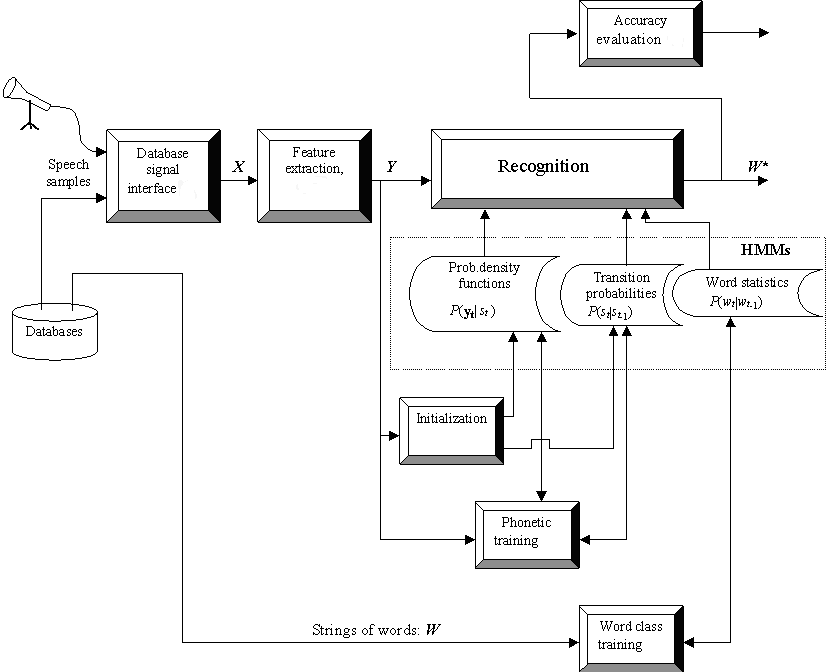
The sentence-level match module uses a language model (i.e., a model of syntax and semantics) to determine the most likely sequence of words. Syntactic and semantic rules can be specified, either manually, based on task constraints, or with statistical models such as word and class N-gram probabilities. Search and recognition decisions are made by 502 considering all likely word sequences and choosing the one with the best matching score as the recognized sentence. Speech recognition software is an ever-increasing part of our lives, though thankfully not like the sci-fi movies of the 90’s led us to believe.

### How does speech recognition work?

There are two main parts to speech recognition. First, the computer needs to “hear” what is being said. Think of a language you are completely unfamiliar with. You can hear what is being said, but since the sounds don’t match any existing patterns in your brain, it’s just noise; you can’t even tell what language it is. The same thing goes for computers. If the incoming audio pattern matches what it knows through the data it has collected to date, it will be able to successfully convert the audio to text. Otherwise, sorry. Need more data.

Part two is making sense of what is being heard. As a Turkish speaker, I can almost perfectly convert Italian audio to text. But the text I end up with is meaningless as my brain has no mapping between the Italian words and their meanings. Again, the same thing goes for a computer. Somebody needs to tell the computer the difference between “on” and “off”, and how it should behave under different circumstances. Otherwise, you’ll receive that familiar response - “Sorry, I am not sure what to do with that”.

As you can imagine, the task becomes increasingly more complex when a device or software is geared towards multiple different markets around the world. Have you ever experienced an accent where you barely understood what the person was saying, even though both of you were fluent in that language? A few more meetings with the same person and you understand just fine. Your brain keeps collecting and processing the data about the phonetic differences and makes the necessary connections. The situation is a little more complicated for computers, however, as there are infinite possibilities that Alexa or Google Home needs to address (imagine the difference in English accents over the US households). To top it off, there are even more complexities when you bring in mixed language use - but we won’t get into that today.



### So, how do the machines understand us?

Think about how a child learns a language. From day one of the child’s life, they are hearing language used all around them. Parents speak to the child knowing that they can’t answer yet. But, even though the child doesn’t respond, they are absorbing all kinds of verbal cues; intonation, inflection, and pronunciation. This is called input. Their brain is forming patterns and connections based on how their parents use language.

Though it may seem as humans are hardwired to listen and understand, we have actually been training our entire lives to develop this so-called natural ability. It takes five or six years for a child to be able to have a full conversation, and then we spend the next 15 years in school collecting more data and increasing our vocabulary. By the time we reach adulthood, we can mentally change these “phonemes” into words and then into meaning, almost instantly.

Speech recognition technology works in essentially the same way. Whereas humans have refined our process, we are still figuring out the best practices for computers. We have to train them as our parents and teachers trained us. And that training involves a lot of innovative thinking, manpower, research and data. Lots and lots of data.

I

n

&

m

b

&

r

i

s

e

v

&

n

th

r

E

n

I

n

z

E

r

o

t

ü

s

e

v

&

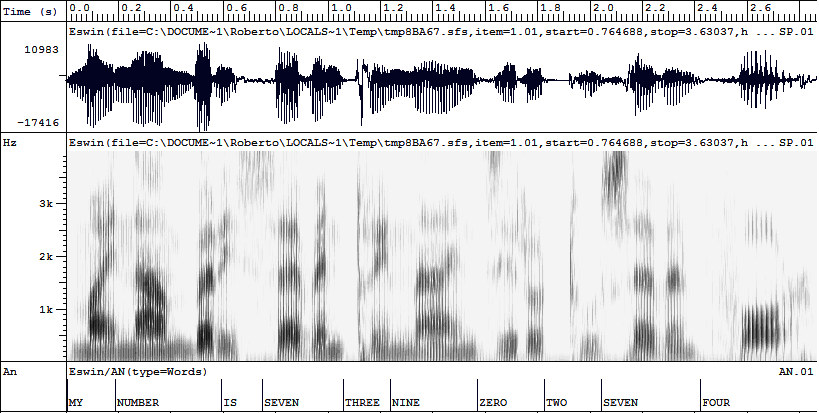
n

f

O

r

11



NP

NP

VP

**MY**

**NUMBER**

**IS**

**SEVEN**

**THREE**

**NINE**

**ZERO**

**SEVEN**

**FOUR**

**A Brief History of Voice Recognition**

**From the 1950’s to the 1960’s**

In the history of speech recognition software technology, this was the era of ‘baby talk’; only numbers and digits could be comprehended. In 1952, ‘Audrey’ was invented by Bell Laboratories which could only understand numbers. But in 1962, the ‘shoebox’ technology was able to understand 16 words in English. Later, voice recognition was enhanced to comprehend 9 consonants and 4 vowels.

**In the 1970’s**

The U.S. Department of Defence contributed heavily towards the development speech recognition systems and from 1971 to 1976, it funded the DARPA SUR (Speech Understanding Research) program. As a result, ‘Harpy’ was developed by Carnegie Mellon which had the ability to comprehend 1011 words. It employed a more efficient system of searching for logical sentences.

There were also parallel advancements in the technology such as the development of a device by Bell Laboratories that could understand more than one person’s voice.

**In the 1980’s**

A major breakthrough was the development of the hidden Markov model which used statistics to determine the probability of a word originating from an unknown sound. It did not rely on speech patterns or fixed templates. Many of these programs made their way into industries and business applications.

A doll was also made for children in 1987; it was known as ‘Julie’ and it could be trained by children to respond to their speech. But speech recognition systems of the 80s had one flaw: you had to take a break between each spoken word.

**In the 1990’s**

With the introduction of faster microprocessors, speech software became feasible. In 1990, the company Dragon released ‘Dragon Dictate’ which was the world’s first speech recognition software for consumers. In 1997, they improved it and developed ‘Dragon NaturallySpeaking’; you could speak 100 words in a minute.

In 1996, the first voice activated portal (VAL) was made by BellSouth. However, this system is inaccurate and still is a nuisance for many people.

**In the 2000’s**

By 2001, speech recognition development had hit a plateau, until Google came along. Google invented an application called ‘Google Voice Search’ for iPhones which utilized data centres to compute the enormous amount of data analysis needed for matching user queries with actual examples of human speech.

In 2010, Google introduced personalized recognition on Android devices which would record different users’ voice queries to develop an enhanced speech model. It consists of 230 billion English words. Eventually, Apple’s Siri was invented which relied on cloud computing as well, and you have a personal assistant who is not only intelligent, but funny and witty too.

**Acoustic Modeling**

An acoustic model is used in automatic speech recognition to represent the relationship between an audio signal and the phonemes or other linguistic units that make up speech. The model is learned from a set of audio recordings and their corresponding transcripts. It is created by taking audio recordings of speech, and their text transcriptions, and using software to create statistical representations of the sounds that make up each word.

Acoustic modeling of speech typically refers to the process of establishing statistical representations for the feature vector sequences computed from the speech waveform. Hidden Markov Model (HMM) is one most common type of acoustic models. Other acoustic models include segmental models, super-segmental models (including hidden dynamic models), neural networks, maximum entropy models, and (hidden) conditional random fields, etc.

Acoustic modeling also encompasses “pronunciation modeling”, which describes how a sequence or multi-sequences of fundamental speech units (such as phones or phonetic feature) are used to represent larger speech units such as words or phrases which are the object of speech recognition. Acoustic modeling may also include the use of feedback information from the recognizer to reshape the feature vectors of speech in achieving noise robustness in speech recognition.

Speech recognition engines usually require two basic components in order to recognize speech. One component is an acoustic model, created by taking audio recordings of speech and their transcriptions and then compiling them into statistical representations of the sounds for words. The other component is called a language model, which gives the probabilities of sequences of words.  Language models are often used for dictation applications. A special type of language models is regular grammars, which are used typically in desktop command and control or telephony IVR-type applications.

**Applications of Voice Recognition**

**Building Good Speech-Based Applications** Good user interfaces which make the application easy-to-use and robust to the good models of dialogue that keep the conversation moving forward, even in matching the task to the technology. Kinds of confusion that arise in human-machine communications by voice.

**The Telecommunications need for Speech Recognition** the telecommunications network is evolving as the traditional POTS (Plain Old Telephony Services) network comes together with the dynamically evolving Packet network, in a structure which we believe will look something like the one

**High-performance fighter aircraft** Substantial efforts have been devoted in the last decade to the test and evaluation of speech recognition in fighter aircraft. Of particular note have been the US program in speech recognition for the Advanced Fighter Technology Integration (AFTI)/F-16 aircraft (F-16 VISTA), the program in France for Mirage aircraft, and other programs in the UK dealing with a variety of aircraft platforms.

**Usage in education and daily life** for language learning, speech recognition can be useful for learning a second language. It can teach proper pronunciation, in addition to helping a person develop fluency with their speaking skills. Students who are blind (see Blindness and education) or have very low vision can benefit from using the technology to convey words and then hear the computer recite them, as well as use a computer by commanding with their voice, instead of having to look at the screen and keyboard.

**Replacing complicated and often frustrating ‘push button’ IVR** Due to poorly implemented and managed systems, IVR and automated call handling systems may be often unpopular and frustrating with customers. However, there is a way to improve this scenario. Termed ‘intelligent call steering’ (ICS), it does not involve any ‘button pushing’. The system simply asks the customer what they want (in their words, not yours) and then transfers them to the most suitable resource to handle their call.

**Advantages of Voice Recognition**

1. **Improve reliability**

Voice recognition systems today are so reliable that such software is now widely used in the health service, the legal profession, the security industry and the military, to name a few. It is now common practice for a doctor to dictate his case notes for them to be converted into digital or paper documents for later use (solving the problem of illegible handwriting), while solicitors, barristers and legal secretaries use voice software to record client information and other notes to be converted into legal files.

1. **Save time**

Dictating is, on average, three times faster than typing, so when time is of the essence and deadlines are looming it makes sense to resort to methods that will help to speed things up. This is where speech recognition software can be especially helpful, by enabling you to dictate, rather than type out, your work. This is especially handy if you’re on the road or short of time – simply voice your thoughts and let your computer do the hard work by transcribing what you say.

1. **Increase work productivity**

Workload piling up? Increase your productivity by spending less time typing, giving you more time to focus on other work. Voice notes enable you to produce a large amount of writing in a relatively short amount of time. By speaking naturally into the microphone, and letting the software do the rest, you can easily get your initial thoughts onto paper – leaving you more time for editing, drafting and revising.

1. **No more mistakes**

No longer will spelling or writing hold you back. Voice recognition software, as well as being faster to complete tasks, is increasingly accurate when it comes to vocabulary and spelling. Although the first systems that came onto the market were often under 90% accurate in terms of the words recognised and the way they were used by the software, today’s systems can reach accuracy levels of 99%+ – thanks to the hundreds of thousands of words now being stored in their database

1. **Greater Mobility**

Use voice recognition technology to dictate on the go, giving you greater mobility and more efficient use of time. With speech notes applications like Dragon Notes by Nuance, you can record your notes immediately after your meeting while conversations are still fresh in your mind, or even dictate your ideas on your way to the meeting.

**Disadvantages of Voice Recognition**

1. **Lack of Accuracy and Misinterpretation**

Voice recognition software won't always put your words on the screen completely accurately. Programs cannot understand the context of language the way that humans can, leading to errors that are often due to misinterpretation. When you talk to people, they decide what you say and give it a meaning. Voice recognition software can do this but may not be capable of choosing the correct meaning. For example, it cannot always differentiate between homonyms, such as "their" and "there."

1. **Time Costs and Productivity**

You might think that computerizing a process speeds it up, but this isn't necessarily true of voice recognition systems, and you may have to invest more time than you expected into the process. You'll have to factor in time to review and edit to correct errors. Some programs adapt to your voice and speech patterns over time; this may slow down your workflow until the program is up to speed. You'll also have to learn how to use the system.

1. **Accents and Speech Recognition**

Voice recognition systems can have problems with accents. Even though some may learn to decode your speech over time, you have to learn to talk consistently and clearly at all times to minimize errors. If you mumble, talk too fast or run words into each other, the software will not always be able to cope. Programs may also have problems recognizing speech as normal if your voice changes, say when you have a cold, cough, sinus or throat problem.

1. **Background Noise Interference**

To get the best out of voice recognition software, you need a quiet environment. Systems don't work so well if there is a lot of background noise. They may not be able to differentiate between your speech, other people talking and other ambient noise, leading to transcription mix-ups and errors. This can cause problems if you work in a busy office or noisy environment. Wearing close-talking microphones or noise-cancelling headsets can help the system focus on your speech.

1. **Physical Side Effects**

If you use voice recognition technology frequently, you may experience some physical discomfort and vocal problems. Talking for extended periods can cause hoarseness, dry mouth, muscle fatigue, temporary loss of voice and vocal strain. The fact that you aren't talking naturally may make this worse and you may need to learn how to protect your voice if you'll use a program regularly.

**Limitations of Voice Recognition**

Considering how popular the two voice-controlled intelligent personal assistants from Google and Amazon—Google Home and Amazon Echo, respectively—have become, it may seem that voice is set to become the default input method of the future. After all, it takes little effort to say what we think, and even the current technology doesn’t seem to have too much trouble understanding our commands. Are there really no limitations or disadvantages of voice control? There are, of course then disappears as a result of the DE coherence process and the information in a quantum bit is lost.

Speech recognition relies upon two components to generate the accuracy levels that are reported. The first is the language model. Although it takes time and money to generate a language model, they are valid for most users of that specific language. They are limited in that current speech technologies cannot handle the complexities of a free speech recognition application outside of maybe a dictation application. The second is the acoustic model. The acoustic model is in simple terms an internal representation of how people using a specific language from a specific country or region speak, i.e., US English is not the same as UK English. UK English is in fact not even real as the UK has a number of very different regional accents. These models take lots and lots of training data to create and for many users they work just fine. However for many others they do not. This is because the data used to generate the model contains samples from tens of thousands of different speakers, so they are generic. Making specific models for individuals is not economical, neither is making models for accents with small populations. Acoustic models are a limitation of the technology. There are many other limitations with the underlying science but it would take a major paper to go into details.

Speech recognition technology refers to a broad set of tools and techniques. I'll specifically address speech rec in the context of IVR and telephony: the big limitation there is the public switched telephone network itself and, to a lesser extent, the highly compressed codecs used by mobile carriers to squeeze every last drop of bandwidth out of their wireless spectrum. You can think of this two ways. In technical terms, a speech recognition engine is happiest when it has more data to work with. Most personal computing devices (laptops, tablets, even your smartphone) use 16-bit samples from two channels captured at 44 kHz -- that's around 1411kbps of bandwidth. The traditional PSTN uses 8-bit samples from a single channel captured at 8 kHz for 64kbps of data. Modern wireless networks take that 64kbps data stream and apply some magic to get that down to a third of that bitrate even. The garbage-in/garbage-out rule applies here: in the end there's only so much a speech recognition system can do with audio coming over a telephone.

The other way to look at this is qualitatively. Is it easier to understand someone talking on a mobile phone or a landline? A landline is almost always better, clearer and easier for a conversation. Would you rather listen to a podcast over a telephone or directly on your computer? Of course it's easier to listen to something like that on your computer. Guess what? A speech recognition engine feels exactly the same way you do: it always performs better when listening to speech over higher quality, higher bandwidth mediums.

**Future Scope**

1. **More Personalized Responses with Contextual Understanding**

The last few years have been about what the user is saying, and now it will be more about why and where they are saying it. Contextual understanding is the next step for voice in order for it to become an integral part of consumers’ lives.

Users want to feel like they’re getting a personalized experience when they interact with technology, specifically voice.

1. **Individualized Experiences**

Voice assistants will also continue to offer more individualized experiences as they get better at differentiating between voices. Last year, Google announced that their Assistant on Google Home is able to support up to six user accounts and detect unique voices, which allows Google Home users to customize many features.

1. **Search Behaviors Will Change**

Voice search has been a hot topic of discussion. Visibility of voice will undoubtedly be a challenge. This is because the visual interface with voice assistants is missing. Users simply cannot see or touch a voice interface unless it is connected to the Alexa or Google Assistant app. Search behaviors, in turn, will see a big change.

1. **Voice Notifications**

In terms of mobile app marketing, voice raises a new challenge- user engagement and retention. Developers will need to find ways to capture and maintain their users’ attention by giving them a push to continue using the app. In September, Amazon added a notification capability to the Amazon Voice Services API. The Notifications feature allows Alexa to proactively indicate that there is new content from core features, such as Shopping updates, and Alexa skills.

1. **Touch Interaction**

CES 2018 proved that voice and displays are merging into one seamless experience. Google showcased their assistant with the Lenovo screen display. In fact, Google’s CES 2018 installation housed of other devices showcasing the Google Assistant, many of which were screen displays so users could further interact with the assistant.

1. **Security Will Be a Focus**

This year, voice payments will become more secure and convenient for users to make voice payments. Users are increasingly aware and familiar with concerns around their smart home devices. With so many devices in the space of our own home, it’s more important than ever to keep information private. Privacy and security need to be a focus as voice makes its way to mainstream, entering the homes of millions more. Speaker verification and ID will become paramount as part of the voice assistant experience with more security being built around the user.