International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

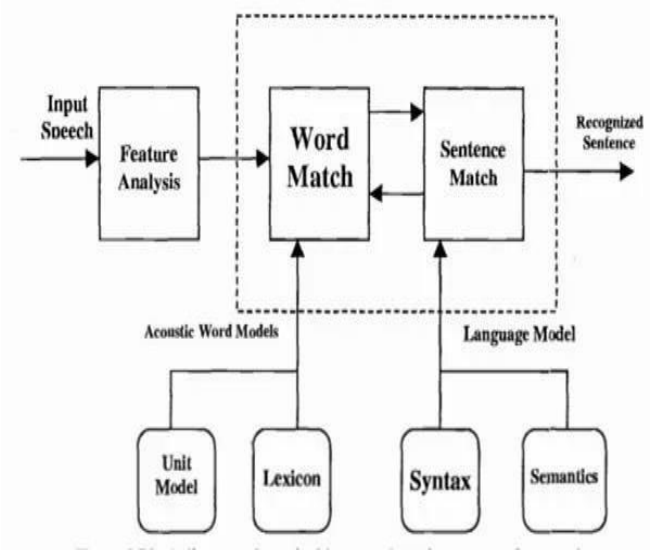
SPEECH RECOGNITION SYSTEM

Rathod Sudam Manik[1], Admane Rajendra Kashinathrao[2], Zameer Farooqui Abdul Hadi Farooqui [3]

[1]Department of Electronics & Telecommunication Engineering,
Aditya Polytechnic, Beed, Maharashtra, India - 431 122
[2,3]Department of Electronics & Telecommunication Engineering,
Mitthulalji Sarda Polytechnic, Beed, Maharashtra, India - 431 122

ABSTRACT

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.



(Fig 1:- Block Diagram of Generic Speech Recognition System)

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.

Generic Speech Recognition System:

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.

Almost every aspect of the continuous speech recognizer of Figure 1 has been studied and optimized over the years. As a result, we have obtained a great deal of knowledge about how to design the feature analysis module, how to choose appropriate recognition units, how to populate the word lexicon, how to build acoustic word models, how to model language syntax and semantics, how to decode word matches against word models, how to efficiently determine a sentence match, and finally how to eventually choose the best recognized sentence.

Building Good Speech-Based Applications:

In addition to having good speech recognition technology, effective speech based applications heavily depend on several factors, including:

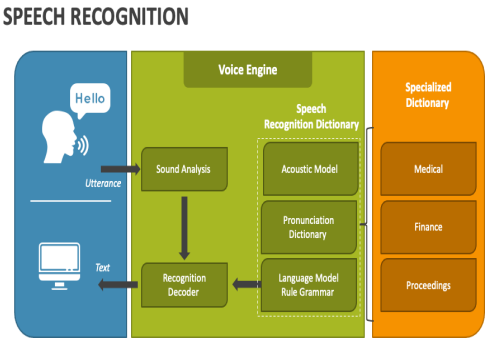
- **Good models of dialogues** that keep the conversation moving forward; periods of great uncertainty on the parts of either the user or the machine.

We now expand somewhat on each of these factors:

User Interface Design: In order to make a speech interface as simple and as effective as Graphical User Interfaces (GUI), 3 key design principles should be followed as closely as possible, namely:

- Provide a continuous representation of the objects and actions of interest.
- Provide a mechanism for rapid, incremental, and reversible operations whose impact on the object of interest is immediately visible.
- Use physical actions or labeled button presses instead of text commands.

Dialogue Design Principles:



(Fig 2:- Block Diagram Of Speech Recognition System)

For many interactions between a person and a machine, a dialogue is needed to establish a complete interaction with the machine. The „ideal" dialogue allows either the user or the machine: to initiate queries or to choose to respond to queries initiated by the other side. (Such systems are called "mixed initiative" systems.) A complete set of design principles for dialogue systems has not yet evolved (it is far too early yet). However, much as we have learned good speech interface design principles, many of the same or similar principles are evolving for dialogue management. The key principles that have evolved are the following:

- Summarize actions to be taken, whenever possible.
- Provide real-time, low delay, responses from the machine and allow the user to barge in it at any time.
- Orient users to their „location" in task space as often as possible.
- Use flexible grammars to provide incrementally of the dialogue.
- Whenever possible, customize and personalize the dialogue (novice/expert)

Match Task to the Technology:

(Fig 3:- Block Diagram of Match Task to the Technology)

It is essential that any application of speech recognition be realistic about the capabilities of the technology, and build in failure correction modes. Hence building a credit card recognition; application before digit error rates fell below 0.5% per digit is a formula for failure, since for a 16-digit credit card, the string error rate will be at the 10% level or higher, thereby frustrating customers who speak clearly and distinctly, and making the system totally unusable for customers who slur their speech or otherwise make it difficult to understand their spoken inputs. Utilizing this principle, the following successful applications have been built:

Game/aids-to-the-handicapped: voice control of selective features of the game, the wheelchair, the environment (climate control).

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 shown in the Figure below.

Telecommunication Applications of Speech Recognition

Speech recognition was introduced into the telecommunications network in the early 1990"s for two reasons, namely to reduce costs via automation of attendant functions, and to provide new revenue generating services that were previously impractical because of the associated costs of using attendants.

Examples of telecommunications services which were created to achieve cost reduction include the following:

Voice Dialing Systems have been created for voice dialing by name (so-called alias dialing such as Call Home, Call Office) from AT&T, NYNEX, and Bell Atlantic, and by number (AT&T SDN/NRA) to enable customers to complete calls without having to push buttons associated with the telephone number being called.

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. Callers dial one number and are greeted by the message "Welcome to XYZ Company, how I can help you?" The caller is routed to the right agent within 20 to 30 seconds of

the call being answered with misdirected calls reduced to as low as 3-5 percent.

By introducing Natural Language Speech Recognition (NLSR), general insurance company Suncorp replaced its original push button IVR, enabling the customer to simply say what they want. Using a financial services" statistical language model of over 100,000 phrases, the system can more accurately assess the nature of the call and transfer it the first time to the appropriate department or advisor. The company reduced its call waiting times to around 30 seconds and misdirected calls to virtual nil.

In-car systems

Typically a manual control input, for example by means of a finger control on the steering wheel, enables the speech recognition system and this is signaled to the driver by an audio prompt. Following the audio prompt, the system has a "listening window" during which it may accept a speech input for recognition.

Simple voice commands may be used to initiate phone calls, select radio stations or play music from a compatible smartphone, MP3 player or music-loaded flash drive. Voice recognition capabilities vary between car make and model. Some of the most recent car models offer natural-language speech recognition in place of a fixed set of commands, allowing the driver to use full sentences and common phrases. With such systems, there is, therefore, no need for the user to memorize a set of fixed command words.

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. In these programs, speech recognizers have been operated successfully in fighter aircraft, with applications including setting radio frequencies, commanding an autopilot system, setting steer-point coordinates and weapons release parameters, and controlling flight display.

Performance of speech recognition systems- It is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate , whereas speed is measured with the real time factor. Dictation machines can achieve very high performance in controlled conditions and require only a short period of training. Optimal conditions usually assume that users -have speech characteristics which match the training data. Can achieve proper speaker adaption. Work in clean and no noise environment. There are 2 models on statistically-based Speech Recognition-Hidden Markov Model (HMM model)Dynamic Time Wrapping (DTW model)

Helicopters - As in fighter applications overriding issue for voice in helicopters is the impact on pilot effectiveness. Battle Management – Speech recognition equipment was tested in conjunction with an integrated information display for naval battle management applications. Telephony and other domains – ASR in the field of computer gaming and simulation is becoming more

Key features of effective speech recognition

Many speech recognition applications and devices are available, but the more advanced solutions use AI and machine learning. They integrate grammar, syntax, structure, and composition of audio and voice signals to understand and process human speech. Ideally, they learn as they go — evolving responses with each interaction.

The best kind of systems also allow organizations to customize and adapt the technology to their specific requirements — everything from language and nuances of speech to brand recognition. For example:

Language weighting: Improve precision by weighting specific words that are spoken frequently (such as product names or industry jargon), beyond terms already in the base vocabulary.

Speaker labeling: Output a transcription that cites or tags each speaker's contributions to a multi-participant conversation.

Acoustics training: Attend to the acoustical side of the business. Train the system to adapt to an acoustic

environment (like the ambient noise in a call center) and speaker styles (like voice pitch, volume and pace).

Profanity filtering: Use filters to identify certain words or phrases and sanitize speech output.

Meanwhile, speech recognition continues to advance. Companies, like IBM, are making inroads in several areas, the better to improve human and machine interaction.

Speech recognition algorithms

The vagaries of human speech have made development challenging. It's considered to be one of the most complex areas of computer science – involving linguistics, mathematics and statistics. Speech recognizers are made up of a few components, such as the speech input, feature extraction, feature vectors, a decoder, and a word output. The decoder leverages acoustic models, a pronunciation dictionary, and language models to determine the appropriate output.

Speech recognition technology is evaluated on its accuracy rate, i.e. word error rate (WER), and speed. A number of factors can impact word error rate, such as pronunciation, accent, pitch, volume, and background noise. Reaching human parity – meaning an error rate on par with that of two humans speaking – has long been the goal of speech recognition systems. Research from Lippmann (link resides outside ibm.com) estimates the word error rate to be around 4 percent, but it's been difficult to replicate the results from this paper.

Various algorithms and computation techniques are used to recognize speech into text and improve the accuracy of transcription. Below are brief explanations of some of the most commonly used methods:

Natural language processing (NLP): While NLP isn't necessarily a specific algorithm used in speech recognition, it is the area of artificial intelligence which focuses on the interaction between humans and machines through language through speech and text. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or provide more accessibility around texting.

Hidden markov models (HMM): Hidden Markov Models build on the Markov chain model, which stipulates that the probability of a given state hinges on the current state, not its prior states. While a Markov chain model is useful for observable events, such as text inputs, hidden markov models allow us to incorporate hidden events, such as part-of-speech tags, into a probabilistic model. They are utilized as sequence models within speech recognition, assigning labels to each unit—i.e. words, syllables, sentences, etc.—in the sequence. These labels create a mapping with the provided input, allowing it to determine the most appropriate label sequence.

N-grams: This is the simplest type of language model (LM), which assigns probabilities to sentences or phrases. An N-gram is sequence of N-words. For example, "order the pizza" is a trigram or 3-gram and "please order the pizza" is a 4-gram. Grammar and the probability of certain word sequences are used to improve recognition and accuracy.

Neural networks: Primarily leveraged for deep learning algorithms, neural networks process training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (or threshold) and an output. If that output value exceeds a given threshold, it "fires" or activates the node, passing data to the next layer in the network. Neural networks learn this mapping function through supervised learning, adjusting based on the loss function through the process of gradient descent. While neural networks tend to be more accurate and can accept more data, this comes at a performance efficiency cost as they tend to be slower to train compared to traditional language models.

Speaker Diarization (SD): Speaker diarization algorithms identify and segment speech by speaker identity. This helps programs better distinguish individuals in a conversation and is frequently applied at call centers distinguishing customers and sales agents.

Speech recognition use cases

A wide number of industries are utilizing different applications of speech technology today, helping businesses and consumers save time and even lives. Some examples include:Automotive: Speech recognizers improves driver safety by enabling voice-activated navigation systems and search capabilities in car radios.

Technology: Virtual agents are increasingly becoming integrated within our daily lives, particularly on our mobile devices. We use voice commands to access them through our smartphones, such as through Google Assistant or Apple's Siri, for tasks, such as voice search, or through our speakers, via Amazon's Alexa or Microsoft's Cortana, to play music. They'll only continue to integrate into the everyday products that we use, fueling the "Internet of Things" movement.

Healthcare: Doctors and nurses leverage dictation applications to capture and log patient diagnoses and treatment notes.

Sales: Speech recognition technology has a couple of applications in sales. It can help a call center transcribe thousands of phone calls between customers and agents to identify common call patterns and issues. AI chatbots can also talk to people via a webpage, answering common queries and solving basic requests without needing to wait for a contact center agent to be available. It both instances speech recognition systems help reduce time to resolution for consumer issues.

Security: As technology integrates into our daily lives, security protocols are an increasing priority. Voice-based authentication adds a viable level of security.

Learning Tasks

From analog audio via mircophone audio interface to digital audio - step 1 before speech analysis

(Applications of Speech Recognition) Analyze the possible applications of speech recognition and identify challenges of the application!

(Human Speech Recognition) Compare human comprehension of speech with the algorithmic speech recognition approach. What are the similarities and differences of human and algorithmic speech recognition?

(Speech and Detection of Emotions) Speech contains more information than the encoded text. Is it possible to detect emotions in speech with methods developed in computer science?

What are similarities and differences between text and emotion recognition in speech analysis?

What are possible application areas in digital assitants for both speech recognition and emotion recognition?

Analyze the different types of information systems and identify different areas of application of speech recognition and include mobile devices in your consideration!

(History) Analyze the history of speech recognition and compare the steps of development with current applications. Identify the major steps that are required for the current applications of speech recognition!

(Risk Literacy) Identify possible areas of risks and possible risk mitigation strategies if speech recognition is implemented in mobile devices, or with voice control for Internet of Things in general? What are required capacity building measures for business, research and development!

(Commercial Data Harvesting) Apply the concept of speech recognition to commercial data harvesting. What are potential benefits for generation of tailored advertisments for the users according to their generated profile? How is speech recognition contributing to user profile? What is the difference between offline and online speech recognition systems due to submission of recognized text or audio files submitted to remote servers for speech recognition?

(Context Awareness of Speech Recognition) The word "Fire" with a candle in your hand and with burning house in the background creates a different context and different expectations of people listening to what someone is going to tell you. Exlain why context awareness can be helpful to optimize the recognition correctness? How can a speech recognition system detect a context to the speech recognition. I.e. detecting the context without a user setting that switches to a dictation mode e.g. for medical report for X-Ray images.

(Audio-Video-Compression) Go to the learning resource about Audio-Video-Compression and explain how Speech Recognition can be used in conjunction with Speech Synthesis to reduce the consumption of bandwidth for Video conferencing.

(Performance) Explain why the performance of speech recognition and accurancy is relevant in many applications. Discuss application in cars or in general in vehicles. Which voice commands can be applied in a traffic situation and which command (not accurately recognized) could cause trouble or even an accident for the driver. Order the theortical application of speech recognition (e.g. "turn right at crossing", "switch on/off music",...) in terms of required performance and accuracy resp. to current available technologies to perform the command in an acceptable way.

(HTML5 Speech Recognition) Analyze the source code of the OpenSource web application demo with PocketSphinx (use browser Firefox/Chromium or Chrome).

Explain how the recognized words are encoded for speech recognition in the demo application (digits, cities, operating systems).

Explain how the concept of speech recognition can support handicapped people with navigating in a WebApp or offline AppLSAC for digital learning environments.

(Size of Vocabulary) Explain how the size of the recognized vocabulary determines the precision of recognition.

(People with Disabilities) Explore the available frameworks Open Source offline infrastructure for speech recognition without sending audio streams to a remote server for processing. Identify options to control robots or in the context of Ambient Assisted Living with voice recognition.

(Version Control) Explore the concept of Version Control and apply that specifically to the Open Community Approach:

Collaborative development of the Open Source code base of the speech recognition infrastructure,

Application on the collaborative development of a domain specific vocabulary for speech recognition for specific application scenarios.

Application on Open Educational Resources that support learners in using speech recognition and Open Source developers in integrating Open Source frameworks into learning environments.

Definition

Speech recognition is the interdisciplinary subfield of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields.

Training of Speech Recognition Algorithms

Some speech recognition systems require "training" (also called "enrollment") where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent".

Models, methods, and algorithms

Both acoustic modeling and language modeling are important parts of modern statistically-based speech recognition algorithms. Hidden Markov models (HMMs) are widely used in many systems. Language modeling is also used in many other natural language processing applications such as document classification or statistical machine translation.

Hidden Markov Model

Dynamic Time Warping

Neural Networks

End-to-End Automated Speech Recognition

Learning Task: Applications

The following learning tasks focus on different applications of Speech Recognition. Explore the different applications.

In-Car Systems

People with Disabilities

Health Care

Telephone Support Systems

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.

Students who are physically disabled or suffer from Repetitive strain injury/other injuries to the upper extremities can be relieved from having to worry about handwriting, typing, or working with scribe on school assignments by using speech-to-text programs. They can also utilize speech recognition technology to freely enjoy searching the Internet or using a computer at home without having to physically operate a mouse and keyboard

Speech recognition can allow students with learning disabilities to become better writers. By saying the words aloud, they can increase the fluidity of their writing, and be alleviated of concerns regarding spelling, punctuation, and other mechanics of writing. Also, see Learning disability.

Use of voice recognition software, in conjunction with a digital audio recorder and a personal computer running word-processing software has proven to be positive for restoring damaged short-term-memory capacity, in stroke and craniotomy individuals.

Further applications

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 emotion recognition

Automatic translation

Court reporting (Real time Speech Writing)

eDiscovery (Legal discovery)

Hands-free computing: Speech recognition computer user interface

Home automation

Interactive voice response

Mobile telephony, including mobile email

Multimodal interaction

Pronunciation evaluation in computer-aided language learning applications

Real Time Captioning

Robotics

Speech to text (transcription of speech into text, real time video captioning, Court reporting)

Telematics (e.g. vehicle Navigation Systems)

Transcription (digital speech-to-text)

Video games, with Tom Clancy's EndWar and Lifeline as working examples

Virtual assistant (e.g. Apple's Siri)

Further information

Conferences and journals

Popular speech recognition conferences held each year or two include SpeechTEK and SpeechTEK Europe, ICASSP, Interspeech/Eurospeech, and the IEEE ASRU. Conferences in the field of natural language processing, such as ACL, NAACL, EMNLP, and HLT, are beginning to include papers on speech processing. Important journals include the IEEE Transactions on Speech and Audio Processing (later renamed IEEE Transactions on Audio, Speech and Language Processing and since Sept 2014 renamed IEEE/ACM Transactions on Audio, Speech and Language Processing—after merging with an ACM publication), Computer Speech and Language, and Speech Communication.

Software

In terms of freely available resources, Carnegie Mellon University's Sphinx toolkit is one place to start to both learn about speech recognition and to start experimenting. Another resource (free but copyrighted) is the HTK book (and the accompanying HTK toolkit). For more recent and state-of-the-art techniques, Kaldi toolkit can be used. In 2017 Mozilla launched the open source project called Common Voice to gather big database of voices that would help build free speech recognition project DeepSpeech (available free at GitHub) using Google open source platform TensorFlow

A demonstration of an on-line speech recognizer is available on Cobalt's webpage

CONCLUSION

This paper presents the Speech Recognition in Artificial intelligence systems and it is important to consider the environment in which the speech recognition system has to work. The grammar used by the speaker and accepted by the system, noise level, noise type, position of the microphone, and speed and manner of the user's speech are some factors that may affect the quality of speech recognition.

REFERANCES

1. ohn Levis and Ruslan Suvorov, "Automatic Speech Recognition".

2. B.H. Juang and Lawrence R. Rabiner, "Automatic Speech Recognition - A Brief History of the Technology Development".

3. S. Xue, X. Y. Kou and S. T. Tan, "Natural Voice-Enabled CAD: Modeling via Natural Discourse".

4. Ekenta Elizabeth Odokuma and Orluchukwu Great Ndidi, "Development Of A Voice-Controlled Personal Assistant For The Elderly And Disabled".