JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ENHANCING HUMAN-MACHINE INTERACTION: ADVANCEMENTS IN SPEECH RECOGNITION TECHNOLOGY

Ravi Prabhat Sharma[1] and Ritu Kadyan[2]

Department of CSE
Ganga Institute of Technology & Management, Haryana, India

Abstract. The vast majority of people has a natural voice. Conversation is made easier by communication. Speech recognition can refer to a method or a piece of equipment that understands spoken words and reacts to them. Computers can translate verbal cues into action-able commands using speech recognition. Users also have the option of speaking naturally. The goal of speech recognition was to enhance human-machine communication. The human-machine interface technology it uses is therefore effective. The history, principles, classifications, and methodologies of speech recognition technology are all covered in this essay. To make it easier to classify these speech technology techniques based on performance, their accuracy is presented.

Keywords: Speech Recognition, Equipment, Spoken Words, Verbal Cues, Human-Machine Communication, Human-Machine Interface Technology, History, Principles, Classifications, Methodologies, Speech Technology Techniques, Performance, Accuracy.

1 INTRODUCTION

Some people with disabilities finds it challenging to speak in their native language. The desire for machine-human communication that is not based on typing has grown over the past ten years. Using speech recognition software, people with disabilities can interact with machines. Listening, recognising, and understanding are made easier by speech recognition. the systematic conversion of spoken language into text. Speech recognition refers to a machine's ability to understand spoken commands and respond appropriately. Machines are now able to understand human speech thanks to two techniques: speech signal processing and pattern recognition. The voice is involved. With the aid of speech recognition technology, machines can now recognise and understand voice signals and translate them into text or instructions. Numerous subfields are included in speech recognition. Information theory and the theory of pattern recognition are related, as is the theory of pattern recognition to neurobiology, linguistics, acoustics, and other fields. Speech recognition technology is quickly becoming one of the most important technologies in the field of computer information processing due to the rapid development of computer hardware, software, and information technology. Communication between humans and machines can be improved thanks to automatic speech recognition. Speech wave analysis is required for both processes, making voice recognition more challenging than speech coding. Process automation, phone inquiries, automatic banking, and secure voice access are just a few of the uses for speech recognition. Research on speech recognition is made possible by allocation. There is a sizable gap between research and commercialization [2]. Complex mental processes like syntactic parsing, semantic analysis, and signal processing (word, syllable, and phoneme recognition) are necessary for understanding spoken language. Speakers can use words they learned as children more effectively thanks to language. Pattern recognition is the answer to the problem of speech recognition because language affects the acoustics of speech [3]. Systems for speech recognition can be discrete or continuous, dependent on the speaker or independent of them, or adaptive. To train a system that relies on the speaker, the user must record a word, phrase, or sentence. Some speaker-dependent systems can recognise the entire vocabulary from a small sample of the vocabulary. Systems without speakers don't require recordings. With each individual speaker, it works. Systems with adaptive speakers can change to accommodate new speakers [4].

1.1 Types of Speech

The following application tasks are made possible by the hierarchy of problems that speech recognition solves: [6]

1.1.1 Isolated Word

The Typically, silent utterances on both sides of the sample windows are necessary to perform isolated word recognitions. One word at a time is subjected to the operation. There is a "Listen as well as the Non-Listen state" here. The term "isolated_utterance" could be used instead [5]. Word recognition can be trained using speakers and is not voice-dependent. The development of "command and control"

applications was made possible by this technology. These programs allowed the computer to recognize and react to a predetermined list of single-word commands. The technology's sensitivity to background noise, which it frequently mistook for false words as well as to unimportant speech, which has accidentally been vocalized by means of the command word, caused serious problems. Many keyword detection algorithms have been developed to address these problems [6].

1.1.2 Connected Word

The connected word system "runs together with the barest minimum of pauses" [5], much like the isolated word system does. Regardless of who uses them, words have connections. Using word models that detected individual terms, this technology was constructed on top of existing word recognition software. Within the framework of the expanded model, a word string can be viewed as a collection of these word models. A class of applications made possible by a new technology that recognizes numeric and alphabetic strings makes it possible to voice dial, authorize credit cards, look up directory assistance, and order catalogues. [6] Voice dialing is also possible thanks to this technology. Voice dialing, information searches for directory assistance, and catalogue ordering all make use of this technology.

1.1.3 Continuous speech

Users of continuous speech recognizers can speak normally while a computer interprets their words. Because continuous speech recognition systems employ a special method to identify the boundaries of utterances, they are notoriously difficult to create [5]. Independent of the speaker, continuous or fluid speech recognition. DARPA was able to complete the Resource Management Task, the ATIS Task, the NAB Task, and desktop dictation systems for PC environments with the help of this technology. Large vocabularies, regulated information access, and dialogue could all be handled by these systems [6].

1.1.4 Spontaneous speech

The most basic use of improvisation is in improvised speech. An ASR system must be able to control word sequences for spontaneous speech [5]. Systems that can understand spoken language and recognise it for use in spontaneous conversations. New services like "Conversation Summarization," "Business Meeting Notes," "Topic Spotting," and language translation between any languages [6] will be made possible by these systems. The method by which speech recognition "identifies" human voices is referred to as "recognise." This is done by converting the sound card's digital audio signals into speech-recognition signals. Numerous mathematical and statistical methods are used to decipher these signals [4]. The four steps required for performing identification or verification during the speech recognition process are shown in Figure 1. The speech first needs to be transformed into a digital wave. converting analogue sound waves to digital sound waves effectively using a microphone. Signal processing is the term for the pertinent process. compute the features after that. Calculations are done within this function's parameters. "Framing" is the process of removing features from an image, and each frame corresponds to a 10-millisecond period of time. The third spot is occupied by the neural network. In this case, language models are in charge of the learning process and a grammar has been pre-built for neural learning. The fourth option is the proper answer. All that remains after speech recognition is finished are words.

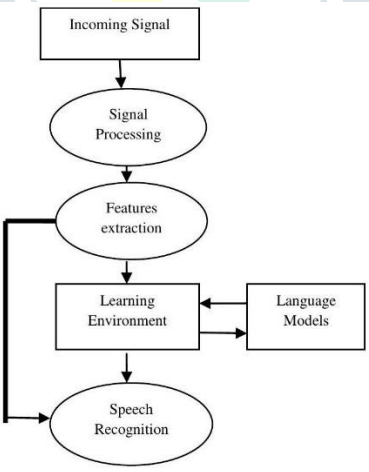

Figure 1: Validating Process for Speech Recognition

2 Basic Principle of Speech Recognition

The four pillars that support the voice recognition system [1] are pattern recognition, feature extraction, matching, and the reference model library.

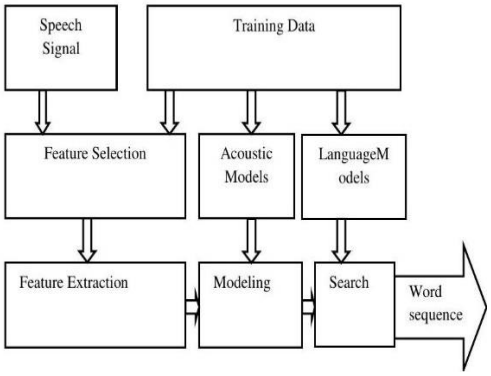

Figure 2: Core Concept of Speech Recognition

2.1 Speech Signal:

Graphical signals are initially used to represent speech. For spectral estimation using linear predictive coding, the signal image must be preprocessed. The amount of prediction error is decreased by this model. The signal spectrum plays an essential role in the Toeplitz Matrix (TM) analysis. Then, using TM as the feature vector, a description of each speech signal image is produced. The following stage is feature vector classification based on Toeplitz. Data can be classified using RBF and probabilistic neural networks [4].

2.2 Feature Selection:

Amplitude, zero-crossing rate, and spectral content are all used in speech recognition. By making it easier to distinguish between vowels and consonants, amplitude, also known as power, is a factor that aids in phonetic recognition. The zero crossings of spectrum balance can be used to distinguish between fricatives and sibilants. Formant and formant transition details can be found in high-resolution spectral data (pitch and format), LP parameters, and filter-bank output. Recognition patterns can be represented as time functions that cover the entire word or as feature values that represent the word's subdivisions. These two interpretations are both correct. Each pattern could be a matrix of formant frequencies for a number of samples taken all over the word in a system that sorts words according to their formant frequencies [7]. This would make it possible for the system to categorize words in a more logical way. These samples came from a variety of places around the world. Each digit of a number is converted into its corresponding initials by a system for number recognition. Then, each region's phonetic material is categorized according to the traits of its feature. The median, final, and reasons are calculated using the dower's variance.

2.3 Feature Extraction

Direct speech recognition is anticipated to improve with waveform digitization. It is preferable to have stabilising characteristics over those that lower the variability of the speech signal. The significance of other characteristics, such as periodicity, pitch, excitation signal amplitude, and fundamental frequency, is somewhat diminished when specific information sources, such as whether a sound is voiced or unvoiced, are removed. The fundamental frequency and amplitude of the excitation signal are essential. The human ear's cochlea is in charge of the quasi-frequency analysis that determines the short-term spectrum. For cochlear analysis, a nonlinear frequency scale known as the Bark or mel scale is typically used. This scale shifts from linear to logarithmic after 1000 Hz. Frequency axis warping usually occurs during the feature extraction process after spectral computation is finished. This makes data analysis easier.

2.4 Acoustic Models:

The individual sounds that make up a word are statistically preserved by acoustic models. The labels for the various phonemes are included in these statistical displays. The English language has forty different phonemes because speech recognition needs forty different sounds. Algorithms that produce statistical representations of each phoneme in a language are trained using speech corpora, which are enormous databases of words that have already been spoken. Another name for a corpus of spoken language is "word bank." HMMs are used to statistically represent these models. Phonemes each have their own HMM [9].

2.5 Language Models :

The two main categories of language modelling approaches are stochastic (also known as statistical) and deterministic (also known as grammar-based). On the basis of their comprehension of a language, their knowledge of how a language model (LM) is currently being developed, and their instincts regarding the most effective formal representation of linguistic entities and relations, specialists create grammar-based language models. An official grammar of a language can be created within this framework. An index of all possible terminal symbols is included in the lexicon that goes with the grammar. [10] The most common kind of grammar used in language applications is probably context-free grammars, or CFG for short. Unsupervised LM estimation produces statistical LMs when done on a training corpus. The initial elements of a statistical LM typically consist of a set of void parameters that are calculated by looking at language-related data. Which estimation parameters to use is still left to the designer's discretion. The use of statistical and grammar-based approaches is contested by a number of strong arguments.

2.6 Modelling:

The modelling approach creates speaker models using feature vectors unique to each individual speaker. Speaker modelling can more accurately recognise and identify speakers with categorization. Automatic speaker identification is made possible by the use of speech signal data. Speaker recognition can be both speaker-dependent and speaker-independent. The computer should ignore speaker-specific traits when performing speaker-independent speech recognition and instead extract the intended message.For a machine to recognise the speakers, the speaker characteristics must be extracted from the acoustic signal [5]. Use text-dependent and text-independent methods to distinguish between speakers. The same crucial phrases or sentences are used in text-dependent training and recognition.

3 Methods For Speech Recognition

3.1 Hidden Markov Model:

An approach to modelling is called hidden Markov modelling. The model, a method for estimating the likelihood that the model will produce a specific output, and a method for estimating the model parameters using examples of the target word that are known must all be considered in addition to the speech data. These approaches are interrelated. An HMM, also referred to as a stochastic hidden Markov model, produces the observable symbols. The random processes that produce the symbols are controlled by a finite state machine. The FSM will randomly choose an output symbol from the set of symbols connected to the state after a state transition is complete. The fact that the FSM's symbols display its state is taken as evidence that it is "hidden" from view. Each vocabulary word is represented by an HMM so that each word can be distinguished. These HMMs may be made up of other HMMs that combine phonemes and other subword units to form a word model. Domain grammar serves as the foundation for the HMM that represents continuous word recognition. Word-model Hidden Markov Models were employed in this grammar model. [11] Utilising quantized speech frame measurements, the observable symbols are synchronized.

3.2 Neural Network Model

The most popular neural network models that can enhance speech recognition include the one-layer perception model, multi-layer perception model, Kohonen self-organizing feature map model, radial basis function neural network, and predictive neural network. Create neural networks that can precisely reflect the dynamic and time-varying characteristics of speech signals using recurrent neural networks, delay neural networks, and other types of neural networks [1]. Deep neural networks, or DNNs, are distinguished by the presence of hidden unit layers in addition to inputs and outputs [12].

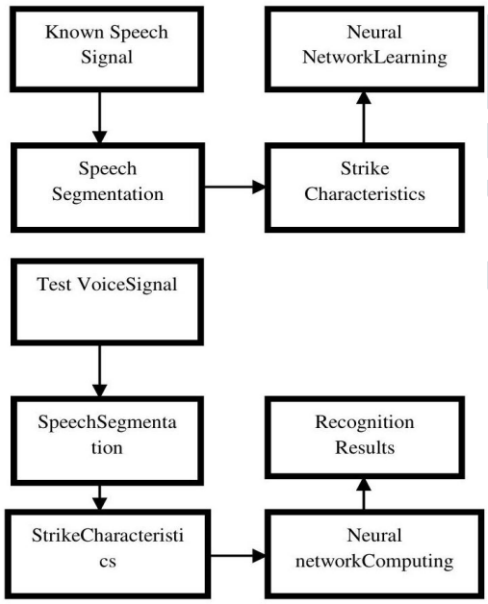

Figure 3: Speech Recognition based NN Architecture [1]

3.3 Dynamic Time Warping

When a speaker repeats a word, it is impossible for a linear-time scaled word-template machine to recognise the word. This was covered in the section before that. Due to variations in time, frequency, and amplitude, the feature vectors or patterns connected to an event will never be the same. Two patterns must therefore represent the same occurrences. It is possible to calculate the pattern distance using this alignment as a reference. The process of "dynamic time warping," or "DTW," allows for the temporal recording of both reference and test patterns. This correspondence is calculated by comparing individual frames of the test pattern and the reference pattern using a user-defined distance metric. The least total accumulated distance across all test pattern frames will be the optimal alignment function. The calculations for time alignment and distance are done at the same time in this stage. The input is identified as a member of the class of reference tokens with the shortest accumulated distance after all reference patterns have been scored [2].

4 Application of Speech Recognition

Set resource priorities. Because there are so many text documents available on the internet and because they are so simple to read, text-based information retrieval is both useful and appealing. Speak-content and multimedia are increasingly being included in voice-activated information retrieval systems. These resources will eventually match the depth and breadth of text-based resources as technology develops further. As a result, there are no issues. [13]. Users' preference for text-based information is not surprising. Retrieval engines do, in fact, correctly rank and filter documents, which increases precision. Particularly for spontaneous speech, voice-based information retrieval is much less accurate than it should be. Speaking under challenging circumstances affects the accuracy of the ASR system, whether in queries or target documents. The precision of voice-based information retrieval technologies is constrained by their memory and computational requirements. Please make sure you adhere to these standards. Therefore, precision is impacted by computation and memory costs. As a last step, think about how the user and the system communicate. The summarization of the text-based data on the screen makes it simple for the user to scan and pick out the pertinent information. The user can choose which of the suggested search terms the engine will return next in an interactive process. The best way to retrieve information is through text because it makes the crucial interaction between the user and the system simpler. On the other hand, voice-based information retrieval makes it difficult to search for and choose spoken and multimedia documents because it is challenging to summarise them on screen [13]. Sifting through and choosing information from multimedia and spoken documents is challenging.

Table 1: Comparison among text-based as well as Voice-based retrieval of information [13]

	Text based	Voice based
Resources can be realized even sooner given mature technologies	Rich resources huge quantities of text documents available over the internet. Quantity continues to increase exponentially due to convenient access.	Spoken/Multimedia contents are the new trend.
Accuracy	Retrieval Accuracy acceptable to users. Retrieving documents properly ranked and filtered.	Problems with speech recognition errors, especially for spontaneous speech under adverse environment.
User-System Interaction	Retrieved documents easily summarized on screens, thus easily scanned and selected by users. Users may easily select query terms suggested for next iteration retrieval in an interactive process.	Spoken/Multimedia documents not easily summarized on screen, thus difficult to scan and select. Lacks efficient user-system interaction.

a. Historically, it was the responsibility of operators to support customers of telecom services. With its aid, callers could be identified. Automation is dominated by DTMF input. Human interfaces and service capabilities may become as usable with voice recognition as with operator assistance [14].

b. Data transmission is carried out automatically. Voice recognition technology is used by directory assistance to find the caller's phone number. Access to account balances and stock market prices over the phone is now possible thanks to voice recognition [14].

c. Sprint PCS runs the biggest digital wireless network in the world. The creative customer service provided by Sprint PCS is also well-known. Voice-activated systems have been available to consumers since the year 2000. The system supports a number of features, such as voice dialling, number verification, and customer service. Another one of China Telecom's innovations is CELL-VVAS. In order to provide user-friendly customised services, the system also seamlessly integrated applications for telecommunications switching networks [1].

5 Comparative Analysis of Speech Recognition

The historical benchmarks are shown in Table 2. Despite the favourable results, the researchers only looked at a small sample of speech classes. Despite the fact that their results were generally favourable. This begs the question of how the speech traits will translate to the speech that wasn't included in their study.

Table 2: Comparison of Different Speech Recognition Studies on the basis of Accuracy Performance

Authors	Feature of Speech	Classifier	Accuracy Performance
Joseph Picone [15]	Continuous speech	HMM	95 – 99%
			84 – 96%
			42-92\% (No. Of Mixture)
			30-44\% (No. Of Models)
Geoffrey, Li Deng [12]	Large Vocabulary Continuous Speech	HMM	88%
			53%
			84%
			48%
Todd A., Herve [16]	Continuous	HMM/ANN	90 – 91%
			70 – 90%
			70 – 80%
			30 – 75%
Nihat, Ulvi [17]	Compared with Voice Command Input	Microprocessor	65 – 70%
E. Chandra, C. Sunitha [4]	Continuous Speech	RBF NNs	97.73%
			94.82%
Jose, David [18]	Impaired People Speech	Fourier Transform	10 Recordings per Word
Olli, David [19]	Noise Robust Speech	Single Mixture HMMs	83.7%
Yifan Gong [20]	Speech in Noisy Environments	HMM Vector Equalization	98.3%
Steve, Herve [21]	Connectionist-statistical speech	CI-HMM	92%
Ossama, Abdel [22]	Voice search large Vocabulary speech	CNN	92%
Alin, Jacek [23]	Visual and auditory Information for speech	PCA	81.11%

1. IBM's commitment to stochastic modelling for speech recognition resulted in the development of the Tangora System. Speech recognition ranging from 5,000 to 20,000 words based on the individual utterances of the speaker. A person's performance is measured by their ability to complete between 95% and 99% of the numerous large vocabulary recognition tasks successfully.
2. Triphone acoustic models, which are series of three phones, are the source of power for the speaker-independent continuous speech recognizer in the SPHINX System. The DARPA Resource Management SPHINX application is a finite state automaton with over 7,000

nodes and 65,000 arcs and a language with 1,000 words. The DARPA Task Domain database's word accuracy for SPHINX can range from 84% to 96% depending on which recognition unit is used.
3. The third component is the HPCDR. As a result of this research, a real-time HMM digit recognition system was demonstrated by the ASPEN multiprocessing system using a number of Digital Signal Processors (DSPs). The precision of the sentence depends on the number of mixtures present in each state because each model has 10 states, one for each digit. Performance of the system ranges from 42% to 92% this time.
4. The fourth element, "improved digit recognition," is referred to as IDR. Her system applied a single multivariate Gaussian distribution to each reference model frame to enhance speech recognition. The accuracy of the sentences varies from 33% to 44% depending on how many models were used.

Using hidden Markov models, Geoffrey and Li Deng presented Large Vocabulary Continuous Speech [12]. HMMs are used by speech recognition systems to handle the data's temporal variability. A frame or a small window of frames representing the acoustic input is used to evaluate the fit of each HMM state using a statistical method called a Gaussian mixture model (GMM). Speech temporal variability is handled using hidden Markov models. Often, DNN-HMMs trained on the same data outperform GMM-HMMs by a large margin. DNN-HMMs outperform GMM-HMMs in a number of tasks despite being trained on more data. In the present study, two HMM systems were used.

1. The DNN-HMM and GMM-HMM models will be evaluated independently by YouTube and Google Voice Input. Only 53% of YouTube's performance is accurate, compared to 88% of Google Voice Input in DNN-HMM. Google Voice Input has an accuracy score of 84% while YouTube has a performance accuracy score of 48% in GMM-HMM.

Todd A., Herve [16] presented Auxiliary Continuous Information Feature using HMM/ANNs model.

Look into different approaches to adding these extra data to hybrid HMM/ANN models or dynamic Bayesian networks (DBN). informational support at the intermediate level, which varies by state. On the high-level supplementary information, the state has no bearing. Intermediate auxiliaries are articulatory traits. It can be confusing because it isn't always obvious whether a characteristic contains high-level or mid-level information. DBNs with results for pitch, energy, and ROS.

Both HMM/ANNs and DBNs use clean and noisy speech systems to assess system performance.

1. First, the DBN accuracy ranges from 70 to 90% in a noisy system and from 90 to 91% in a clean system.
2. Second, the accuracy of HMM/ANNs ranges between 30 and 75% in noisy systems and 70 and 80% in clean systems.

Voice Command was compared to a voice recognition system model built using a PIC18F452 microcontroller by Nihat and Ulvi [17].

The PIC18F452 microcontroller did not record a voice command, only two words. We looked into the timing and counting capabilities of audio signals with zero crossings. Our mouths and tongues create each and every sound that we hear as a linear combination of sine waves with various frequencies. 300–3300 hertz are used in human speech. According to the Nyquist Theorem, doubling both the sampling frequency and the sound frequency yields an accurate sample. An increased number of people were sampled. The ability of electret microphones to record voices is improved by the LM386. A PIC18F452 microprocessor was used in this investigation. Microprocessors from the PIC18F452 family are used to power this device. The 10-MIPS-per-second 10-bit CMOS FLASH PIC18F452 microcontroller processes data and executes instructions at a speed of 100 ns per instruction. The PIC18F452's synchronous serial port can be set up as either a 2-wire inter-integrated circuit bus or a 3-wire serial peripheral interface. It also has an addressable universal asynchronous receiver transmitter, eight 10-bit A/D converter channels, and two 10-bit capture/compare/PWM functions. The PIC18F452 has eight A/D converters, each of which has an output resolution of 10 bits. During testing, voice recognition accuracy varies between 65 and 70%.

E. Chandra, C. Sunitha [4] presented Speech and Speaker identification using Neural Networks.

To classify the data, probabilistic and RBF neural networks were used. The conventional and neural network-based classification techniques were picked because of how simple they are to use. Both methods use Burg's curve or a minimal Eigen value algorithm as their input data. 220 classes are produced by twenty speakers and eleven digits. For speaker identification, we currently use 20 classes, each of which corresponds to one speaker in the author's database. The 11 classes used by voice recognition, in contrast, each represent a value between 0 and 10.

The authors carried out the next two experiments:

1. Nonetheless, when the TM Minimal Eigen values Algorithm is implemented, Classification's performance rate rises to 97.73%.
2. The second model is Burg's classification of standard speech. In this instance, 94.82% of attempts to identify the speaker were successful.

Using the Fourier transform, David and Jose [18] were able to distinguish hearing-impaired voices.

In this article, speech recognition is demonstrated using MATLAB as an example. When it comes to signal processing algorithms, MATLAB excels. The Fourier Transform provides additional assistance for this investigation. The signal was accurately portrayed in the frequency domain by the discrete Fourier transform (DFT). FFT is used in numerous DFT applications. The phonemes were recorded

using Sagebrush Systems Recall Version 2.4a. There are ten recordings for each phoneme. They played the five recordings from the collection that they thought were the best. The least amount of background noise was present in these five recordings. Using MATLAB, five different phoneme recordings were analyzed. This program showed the filtered signal's spectrum in a variety of frequency bands. The characteristic standard of the phoneme was found to be the spectrum with the highest frequency.

The system's accuracy was verified by the authors by recording ten fresh voice samples for each word. final modification.

In the article cited as [19], Olli, David, and Kari presented a noise-resistant speech recognition algorithm based on a single mixture HMM.

Repeating feature The speaker-dependent recognition of names was assessed in addition to the speaker-independent recognition of connected digits. During the evaluation process, the first, second, and z-coefficients of each of the thirteen MFCCs were determined using the incoming signal. Various N and SNR intensities are used to show the speaker-dependent name recognition results. Even with low SNR, recognition needs at least 20 feature vectors to be buffered, which takes 0.2 seconds.

Single mixture HMMs performed 83.7% better in noisy environments.

Yifan Gong [20] was successful in identifying speech in noisy environments using HMM vector equalization.

Speech recognition in noisy environments is facilitated by class-dependent processing, time and frequency correlations, task-specific knowledge of speech and noise, prioritizing speech segments with a high signal-to-noise ratio (SNR), time and frequency correlations, and auditory models. Training and testing environments are not the same. This is due to the possibility that the parameter distribution, which yields excellent recognition results for speech with no background noise, may be very sensitive to disturbances in noisy environments.

The authors' performance in noisy environments with a vocabulary of 10 digits is 77.3% better when Hmm vector equalization and multiple speakers are used.

The Connectionist-statistical Hidden Markov speech recognition model was introduced by Steve and Herve [21]. This system is compared to the DARPA Resource Management database without speaker information. This comparison employs a multilayer perceptron probability estimator. It will become commonplace to use feature vector-based models of statistical speech signals. To communicate, they make use of basic HMMs like phones. The lexicon and the language model both recommend combining these into networks before attempting to form words or sentences.

Word accuracy for the context-free MLP-HMM hybrid system was 92%. Osama and Abdel presented on Convolutional Neural Network Models for Large Vocabulary Speech Search in [22], which is a useful tool for investigating consistent modelling, speaker adaptation, robust front ends, and context-dependent phone modelling.

When modelling speech with CNNs and a limited weight-sharing scheme, it is possible to achieve higher accuracy and lower error rates. CNN helps to cut down on mistakes. On the TIMIT phone recognition and voice search large vocabulary speech recognition tasks, CNNs outperformed DNNs by a margin of 6–10%. CNNs demonstrated their speech recognition abilities using both TIMIT phone recognition and VS with a large vocabulary. These two tasks required a significant amount of vocabulary. CNN only had two layers, whereas DNN had three that were all hidden and connected. For each section, CNN uses 84 feature maps and a weight-sharing system with a set number of points. When it comes to reducing errors, CNN is 8% more efficient than DNN. The performance consequently increased to 92%.

Principal component analysis was used by Alin and Jacek [23] to present data on both aural and visual speech recognition.

Audio-only speech recognition technology is less accurate in settings with low Signal-to-Noise Ratios (SNR). Because video content is unaffected by background noise, it is a perfect medium for speech recognition data fusion. This paper shows that the majority of static visual feature extraction techniques yield features that are equivalent. They demonstrate how these processes result in this property. They show that optical flow analysis-based audio-video recognition performs remarkably well in low signal-to-noise environments. Realistic motion simulation is simulated in some of the game's elements. In most cases, PCA is used for this. Cosine transforms, discrete wavelets, and PCA have all been used in real-world applications. When static images and clear audio are given equal weight, the word recognition rate can reach 81.11%.

6 Conclusion

This paper provides an overview of the constantly evolving and improving field of speech recognition. Potential benefits of speech recognition technology include the ability to control digital systems and enhance communication with individuals with disabilities. Furthermore, there are numerous opportunities for speech recognition technology to advance, which could significantly enhance the services provided to people with disabilities. Speech recognition can also help to create a secure environment via voice authentication. The use of HMM and ANN models is prevalent in continuous speech recognition processes, as evidenced by the tabular presentation of various methods and their corresponding accuracy values. In the coming years, the accuracy and quality of speech recognition are anticipated to significantly increase, making communication simple and reliable for everyone, including those with disabilities. Future systems must aim for greater efficiency and capability than existing ones. Speech recognition has an exciting future because of its rapid

development. The technology's initial applications were dispersed. The adaptability of speech recognition technology to various speakers and background noise, however, can be improved. Many aspects of speech recognition technology, including this one, still need work. In the end, these advancements will make accessible across all telecommunications services robust and dependable voice interfaces.

7 References

1. Jianliang Meng, Junwei Zhang and Haoquan Zhao, "Overview of the Speech Recognition Technology", 2012 Fourth International Conference on Computational and Information Sciences, 978-07695-4789-3/12$26.00@2012 IEEE.
2. Andress S. Spanias, Frank H. Wu, "Speech Coding and Speech Recognition Technologies: A Review", CH3006-4/91/0000-0572$1.000 IEEE.
3. Jeff Zadeh, "Technology of speech for a computer system", DECEMBER 2003/JANUARY 2004, 02786648/03/$17.00 @ 2003 IEEE.
4. E. Chandra and C. "A review on Speech and Speaker Authentication System using Voice Signal feature selection and Extraction", 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.
5. Santosh K.Gaikwad, Bharti W.Gwali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 - 8887) Volume 10- No.3, November 2010.
6. Lawrence R. Rabiner, "Applications of speech recognition in the area of telecommunication", 07803-3698-4/97/$10.00 01997 IEEE.
7. Tingyao Wu, D. Van Compernolle, H. Van hamme, "Feature Selection in Speech and Speaker Recognition" June 2009. U.D.C. 681.3_I27. Phd Thesis.
8. Urmila Shrawankar, Vilas Thakar, "Techniques for Feature Extraction in Speech Recognition System : A Comparative Study".
9. Chris Biemann, Dirk Schnelle-Walka, "Unsupervised acquisition of acoustic models for speech-to-text alignment", Master-Thesis von Benjamin Milde 10. April 2014.
10. Maxim Khalilov, J. Adri'an Rodr' guez Fonollosa, "New Statistical And Syntactic Models For Machine Translation", TALP Research Center, Speech Processing Group, Barcelona, October 2009.
11. Richard D. Peacocke, Daryl H. Graf, "An Introduction to Speech and Speaker Recognition", Bell-Northern Research,IEEE August 1990.
12. Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition", Digital Object Identifier 10.1109/ MSP.2012.2205597,Date of publication: 15 October 2012.
13. Lin-shan Lee and Yi-cheng Pan, "Voice-based Information Retrieval How far are we from the textbased information retrieval?", IEEE ASRU 2009. Special Conference Issue: National Conference on
14. Masanobu Fujioka, Seiicbi Yamamoto, Naomi lnoue, Makoto Nakamura and Takashi Mukasa, "Experience and Evolution of Voice recognition applications for telecommunications services" 07803-4984-9/98/$10.00 01998 IEEE.
15. Joseph Picone, "Continuous Speech Recognition Using Hidden Markov Models", IEEE ASSP MAGAZINE JULY 1990.
16. Todd A. Stephenson, Mathew Magimai Doss and Hervé Bourlard,"Speech Recognition with Auxiliary Information", IEEE transactions on speech and audio processing, vol. 12, no. 3, May 2004.
17. Nihat Öztürk and Ulvi Ünözkan, "Microprocessor Based Voice Recognition System Realization", 9781-4244-6904-8/10/$26.00 @2010 IEEE.
18. José Leonardo Plaza-Aguilar, David Báez-López, Luis Guerrero-Ojeda and Jorge Rodríguez Asomoza, "A Voice Recognition System for Speech Impaired People", Proceedings of the 14th International Conference on Electronics, Communications and Computers (CONIELECOMP'04) 0-7695-2074X/04 2004 IEEE.
19. Olli Viikki, David Bye and Kari Laurila, "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise", 0-78034428-6/98 $70.08 01998 IEEE.
20. Yifan Gong, "Speech recognition in noisy environments: A survey", Speech Communication 16 (199.5) 261-291,0167-6393/95/$09.50 01995 Elsevier Science B.V.
21. Steve Renals, Nelson Morgan, Herve Bourlard and Michael Cohen, "Connectionist Probability Estimators in HMM Speech Recognition", IEEE Transactions on Speech and Audio Processing, VOL. 2, NO. 1, PART 11, JANUARY 1994.
22. Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, VOL. 22, NO. 10, OCTOBER 2014.
23. Alin G. Chit u u, Leon J.M. Rothkrantz, Pascal Wiggers and Jacek C. Wojdel, "Comparison between different feature extraction techniques for audio-visual speech recognition", Journal on Multimodal User Interfaces, Vol. 1, No. 1, March 2007.