KAMPALA INTERNATIONAL UNIVERSITY
EXPLORING THE HEIGHTS

KIU
Journal of Science, Engineering and Technology

KJSET Review Article

Review of Techniques used in Speech Signal Processing

Arinaitwe Prosper [1], Murungi Emelta [2], Ogenyi Fabian Chukwudi [3], Asiimwe Robert[4] and Mohammed Dahiru Buhari[5]
Department of Electrical, Telecommunication and Computer Engineering, Kampala International University, Uganda[1,2,3,4,5]
Electrical and Electronic Engineering Department, Abubakar Tafawa Balewa University, Bauchi, Nigeria[5]
Corresponding Author: prosper.arinaitwe@studwc.kiu.ac.ug[1]

Paper history:
Received 22 December 202
Accepted in revised form 14 April 2024

Keywords
Signal processing techniques, Fourier transform, Mel-Frequency Cepstral Coefficients (MFCCs), Hidden Markov Models (HMMs), Deep Neural Networks (DNNs), Waveform coding, Speech signal processing.

Abstract

This paper provides an in-depth examination of crucial signal processing methods essential for analyzing and interpreting various signals, particularly in the realm of speech. The reviewed techniques include the Fourier transform, Mel-Frequency Cepstral Coefficients (MFCCs), Hidden Markov Models (HMMs), Deep Neural Networks (DNNs), and waveform coding. The applications of these methods in speech signal processing are elucidated, highlighting their specific advantages and inherent limitations. The paper also explores challenges associated with signal processing, such as the impact of noise, equipment quality, and computational demands. Emphasizing the need to carefully choose the appropriate signal processing technique for a given task, the review underscores the importance of striking a balance between the strengths and weaknesses of each method to achieve effective signal enhancement and analysis.

Nomenclature and units

k	Kurtosis
γ	Skewness
τ_{MED}	Mean Excess Delay
τ_{RDS}	Root Mean Squared Delay Spread
μ	Mean

1.0 Introduction

Speech signal processing stands as a cornerstone in contemporary technology and communication, playing a pivotal role in facilitating interaction between humans and machines through the interpretation of spoken language. At its core, this field involves the transformation of analog audio waveforms into digital data, a process referred to as analog-to-digital conversion (ADC) [1]. Once acoustic signals are digitized, a diverse range of techniques and methodologies comes into play for extracting, analyzing, and manipulating spoken language. An exemplary application of speech signal processing is found in Automatic Speech Recognition (ASR) systems, which transcribe spoken words into text, powering voice-activated assistants. These systems leverage advanced techniques like hidden Markov models (HMMs) and deep learning [2]. Furthermore, speech signal processing plays a pivotal role in speaker recognition, contributing to enhanced security through the identification of unique vocal characteristics [3]. This proves valuable in applications requiring user authentication and access control. Additionally, the field of speech signal processing contributes significantly to accessibility applications, where Text-to-Speech (TTS) systems bring written text to life by generating natural-sounding speech.

Text-to-Speech (TTS) systems, vital in various applications, rely on a blend of techniques such as speech synthesis and prosody modeling [4]. In all these applications, speech signal processing seamlessly operates, acting as the bridge between humans and machines, fundamentally supporting effective communication. The techniques employed in speech signal processing highlight the intricate interplay between sound waves, algorithms, and human comprehension, shaping our technological landscape and promising further innovations in an increasingly voice-centric world. This comprehensive review delves into the complex realm of techniques used in speech signal processing, where the transformation of spoken language into a format understandable and generatable by machines is achieved [5]. In an era where voice interfaces and natural language understanding play an ever-growing role in daily life, understanding these underlying techniques becomes paramount.

The journey commences with the acquisition of acoustic signals through microphones, humble devices serving as the ears of the digital world, capturing the subtleties of human speech and converting them into electrical signals [1]. This analog-to-digital conversion marks the foundation of speech signal processing, representing the pivotal moment when the sounds of human communication are translated into the language of machines. As the acoustic signals undergo digitization, they undergo sophisticated analysis. Among the critical domains in this journey is Automatic Speech Recognition (ASR). ASR systems leverage advanced technologies, including hidden Markov models (HMMs) and deep learning, to transcribe spoken words into text [1]. These systems meticulously analyze pronunciation, intonation, and language intricacies, enabling voice-activated assistants to accurately comprehend and respond to your voice commands. The significance of speech signal processing goes beyond recognition, extending into the domain of speaker identification. Techniques within this field can differentiate individuals based on their unique vocal characteristics, providing an additional layer of security [3]. In an era where identity verification is paramount, especially in secure systems and digital transactions, speaker recognition adds a robust authentication layer.

Furthermore, the transformative impact of speech signal processing is evident in Text-to-Speech (TTS) systems. These sophisticated technologies employ diverse synthesis techniques, including neural networks, to faithfully recreate the richness and subtleties of human speech. Text seamlessly transforms into lifelike speech, enhancing accessibility applications and making information more readily available to those reliant on spoken content. Beyond mere convenience, these techniques have profound implications across various domains, including healthcare, where speech processing aids in early disease diagnosis, telecommunications benefiting from improved voice quality, and human-computer interaction, where natural language understanding is increasingly vital [4]. This field serves as the silent enabler behind the scenes, making human-machine communication more intuitive and effective. As we delve deeper into the multifaceted world of speech signal processing, we uncover the core techniques that underpin its functionality [5]. Whether approached from a technological, linguistic, or curiosity-driven perspective, this exploration provides a glimpse into the inner workings of a domain serving as the connective tissue between humans and machines, all through the remarkable power of speech. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) play pivotal roles in the domain of speech signal processing [6]. RNNs excel in modeling sequential features in speech, making them highly effective in tasks such as speech recognition, where the context of previous phonemes is crucial for accurate transcriptions [6].

In contrast, CNNs excel in capturing spectral features by processing audio data in the time-frequency domain, rendering them invaluable for tasks such as audio classification, noise reduction, and audio-based machine learning [6]. Prosody modeling introduces an additional layer of richness to speech processing [7], honing in on the subtleties of rhythm, pitch, and intonation—essential elements for making Text-to-Speech (TTS) systems sound more natural and expressive [7]. By incorporating prosodic information, TTS systems can imbue synthesized speech with appropriate emotional and linguistic cues, elevating the overall quality and human-like nature of the output [7].

In conjunction with RNNs, CNNs, and prosody modeling, a spectrum of complementary techniques further enhances the capabilities of spoken language processing [8]. Speaker diarization, for instance, aids in distinguishing between different speakers in a conversation, enabling systems to transcribe and attribute speech to specific individuals—an essential function in

tasks like transcription services and meeting recordings [8]. Voice Activity Detection (VAD) serves as a crucial preprocessing step in speech processing by identifying segments of audio containing actual speech and filtering out silence or non-speech regions [9]. This not only enhances the efficiency of ASR and TTS systems but also contributes to noise reduction and signal enhancement [9]. Mel-Frequency Cepstral Coefficients (MFCCs) represent essential features used to depict the spectral characteristics of speech [10]. Widely employed in ASR systems, MFCCs effectively capture the distinctive aspects of speech sounds, facilitating models in distinguishing between phonemes and words [10]. Collectively, these techniques form the bedrock of modern speech signal processing, enabling accurate and natural speech recognition and synthesis [6]. Their pivotal roles span applications from voice assistants to transcription services, contributing to the seamless communication between humans and machines in an increasingly voice-centric world [6]. A systematic examination of signal processing techniques, including filtering, Fourier analysis, wavelet analysis, time-frequency analysis, and adaptive signal processing, is instrumental for extracting, analyzing, and interpreting signals, providing a distinct advantage for enhanced comprehension. In the next following sections, the overview of techniques used in speech signal processing will be explained with comparative studies of their merits and demerits. Then the challenges and limitations, recommended methods and conclusion will be discussed.

2.1 The Fourier transform:

This is a method employed in the analysis of speech signals, breaking them down into sine and cosine waves as in figure 1. While extensively utilized for speech analysis and feature extraction, it encounters challenges in capturing the time-varying aspects of speech signals. The Fourier transform presupposes signal stationarity, potentially resulting in imprecise frequency estimates and diminished spectral resolution. The selection of window shape and size also plays a critical role, influencing the accuracy of the Fourier transform, with different shapes and sizes being more suitable for various speech features. Furthermore, the Fourier transform may introduce artifacts, such as spectral leakage or aliasing, posing a risk of signal distortion and inaccurate analysis outcomes. Hence, a meticulous assessment of Fourier transform techniques is essential, ensuring they effectively capture pertinent speech information, mitigate artifacts, and accurately represent non-stationary signals.

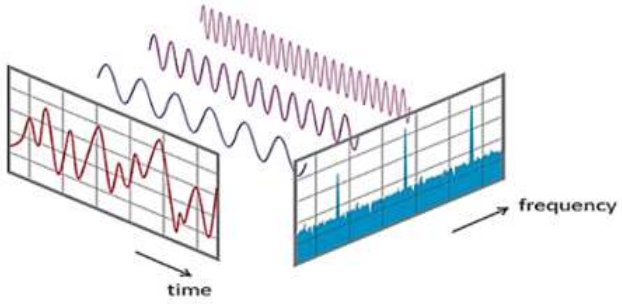

Figure 1: View of a signal in the time and frequency domain as broken down into cosine and sine waves [11].

2.2 The Automatic Speech Recognition System (ASR):

This is a technology that converts spoken words into written representation through the use of software and computer-based methods, employing techniques to identify and analyze human voice [12].

ASR systems find applications in various contexts such as voicemail transcription, YouTube closed captioning, Google voice search, dictation, and voice-operated systems. It's important to note that the primary function of ASR is speech translation; it does not encompass speech understanding [13]. If a machine receives the command "Urgently call me a doctor" and responds with "From now on, I will call you a doctor" instead of actually making the call, any misunderstanding lies within the speech comprehension module, not the ASR system itself. ASR is specifically focused on the conversion of speech into text and does not extend to comprehending the semantic meaning of spoken words.

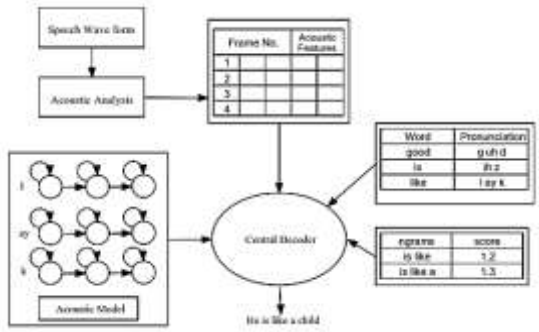

Figure 2: *Structure of an ASR system [13].*

Based on the illustration in Figure 2, the utilization of Artificial Neural Networks (ANN), specifically the Deep Neural Network-Based Speech Recognition System introduced in 1996, remains a prevalent approach in contemporary speech recognition systems. Prominent examples include SIRI, Cortana, Google Voice Search, and others, indicating the enduring influence of this technology across various platforms.

2.3 The Hidden Markov Model (HMM):

This stands out as a crucial machine learning model in the realm of voice and language processing, representing an extension of the Markov. The Markov model and Markov chain find their basis in weighted finite state automata. In this framework, the transition probability from one state to another is denoted by the weights assigned to each arc in the state transition. The Markov model operates as a completely observable model [14], determining state transition probabilities based on long-term observations of these transitions. As individual states are considered nodes, the probability of each arc departing from a node is stipulated to be 1, with the input sequence governing the state transition. However, the Markov model is constrained to identify entirely observable events [14], limiting its application to clear sequences and preventing it from effectively addressing inherently complex issues. In this model, each state is linked to a deterministic observable event, making it unsuitable for numerous real-world

challenges like automatic voice recognition. The block diagram of this technique is shown in Figure 3. To overcome these limitations, the hidden Markov model (HMM) [15] emerges as a variant, viewing the system under examination as a Markov process transitioning between hidden and unobserved states. The resulting HMM integrates an observable stochastic process with another stochastic process, providing a more versatile framework capable of modeling complex real-world phenomena.

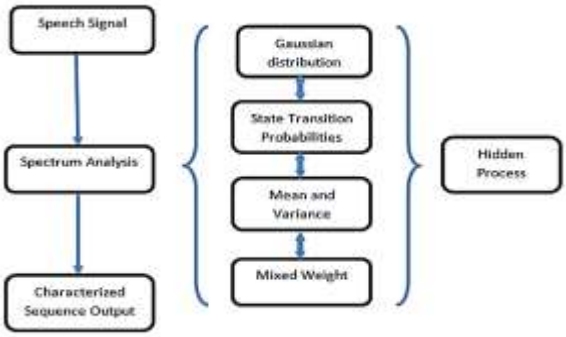

Figure 3: The Hidden Markov Model (HMM) [16].

2.3.1 Limitations

The application of the Hidden Markov Model (HMM) in speech recognition comes with inherent limitations [17]. Despite the assumption that subsequent observations are independent, in reality, they exhibit significant interdependence. The foundational concept of HMM rests on the Markov property, asserting that the state at a given time solely dictates the probability of a specific state at that moment. However, it's observed that dependencies often extend across multiple states, especially in the context of speech sounds. Additionally, the frame size is fixed in this scenario, lacking a precise formal procedure for determining the optimal architecture to address a given problem. Moreover, establishing the requisite number of states and transitions for a model lacks a standardized approach. Training an HMM demands an extensive number of decision parameters and a substantial volume of data.

2.4 Neural Network and Speech Recognition System

An artificial neural network (ANN) is an information processing system designed to emulate the functioning of the biological nervous system, such as the human brain. Its block diagram is as shown in figure 4. It addresses specific problems through a vast network of interconnected processing nodes, or neurons that collaborate in a manner akin to human learning processes. Artificial Neural Networks (ANNs) learn from their past experiences, similar to human learning. In the implementation of an ANN [18–21], it handles a multitude of inputs and produces a single output. The neurons operate in two distinct modes within this framework. Initially, the neurons undergo a training phase where they learn to recognize specific input patterns. Subsequently, the neurons transition to an execution mode where the current output is derived when a previously taught input pattern is detected. In cases where the input pattern is not found in the taught list, a firing rule is applied to determine the output. A firing rule is established to determine whether a neuron should activate for a given input pattern. This rule is universally applicable to all potential input patterns and is not specific to the patterns that have been taught.

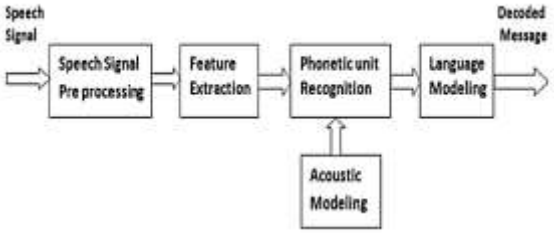

Figure 4: An artificial neural network (ANN) block diagram [22].

2.4.1 Application of Neural Network in Speech Recognition

The implementation of automatic voice recognition can be effectively achieved through the application of the recurrent neural network approach [23]. Initially, the audio voice input undergoes sampling. Directly inputting this sampled data into the neural network for speech pattern identification would be highly challenging. To facilitate this process, the sampled data undergoes specific preparation steps. The data is segmented into 20-millisecond segments, encompassing a brief period that includes a mix of high-pitched, mid-pitched, and low-pitched noises. The Fourier transform is then applied to this complex sound segment, separating it into its individual simple component parts. For example, the 20-ms frame is divided into distinct low-pitched and other components through the Fourier transform. Subsequently, the energy present in each frequency band is amalgamated, creating an audio sample fingerprint. This procedure is applied to every 20-ms audio chunk, generating a spectrogram that visually represents the pitch patterns in the audio data.

The spectrogram is particularly advantageous as it allows for a clearer visualization of pitch patterns, enabling a neural network to more readily identify patterns compared to raw acoustic data. The neural network receives this representation of audio data as input, analyzing audio segments every 20 milliseconds and attempting to identify the corresponding spoken sound's letter in each slice.

Understanding the human speech production system is crucial in building an Automatic Speech Recognition (ASR) system. The model of the human articulatory system considers each articulatory organ as a variable, producing various phones through diverse combinations of variable values. The historical development of the ASR system is briefly outlined, providing insights into its evolution. ASR systems consist of multiple components, utilizing various audio characteristics as input. The acoustic analysis stage extracts these characteristics, matching related phonemes to acoustic features. The Hidden Markov model, a frequently used statistical model, maps feature vectors to phonemes. The introduction of neural networks and deep recurrent neural networks has significantly advanced speech

recognition, enhancing prediction accuracy through feedback loops in the recurrent neural network.

The pronunciation model aligns identified phonemes with corresponding words, addressing prediction uncertainties arising from dialect variations by incorporating articulatory information. Meanwhile, a language model organizes words into coherent sequences to form sentences. A central decoder receives outputs from each part of the ASR system and is responsible for selecting the most accurate phrase from a vast collection. The search graph for this process is extensive, with exponential time complexity. Given that many pathways in the search graph are undesired, search graph optimization becomes imperative. The effective Viterbi method for search graph optimization is presented as a conclusion.

2.5 Spectral analysis:

This involves examining the frequency content of a speech signal to identify formants and other essential elements crucial for speech recognition. Feature extraction is a vital process in speech processing that eliminates unnecessary information and noise from a voice signal to retain essential details. The fundamental feature extraction process encompasses statistical modeling, parametric transformation, and spectral analysis [23], resulting in a parameter vector [24]. However, speech transmissions are intricate and may contain overlapping frequencies, posing challenges in precisely identifying specific components. Spectral analysis operates on the premise that the signal can be decomposed into its constituent frequencies. Nevertheless, there is a common trade-off where irrelevant information is eliminated, potentially resulting in the loss of important details [26]. Another challenge is that the technology used for spectral analysis can impact the resolution and accuracy of the analysis. Different methods, such as wavelet transform or short-time Fourier transform, come with their advantages and disadvantages in capturing specific speech feature types. Moreover, spectral analysis can be computationally demanding, requiring substantial memory and computing resources. This computational intensity may limit its applicability, particularly in low-power or real-time applications. Effective spectral analysis methods are essential to ensure the efficient collection of pertinent speech data while minimizing errors and processing demands. Figure 5 shows how spectral analysis is applied.

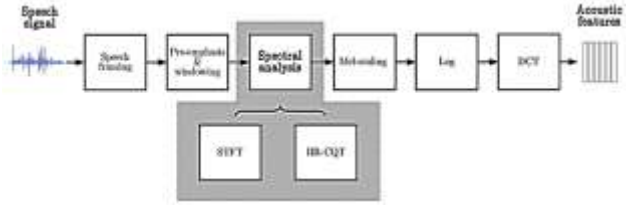

Figure 5: How Spectral analysis is applied in speech signal processing [27].

2.6 Mel-Frequency Cepstral Coefficients (MFCCs)

This serves as a method for extracting characteristics from speech signals to facilitate speech recognition. This approach is adept at capturing the most significant spectral and temporal aspects of speech, drawing inspiration from the human auditory system. An advantageous feature of MFCCs is their noise resistance and adaptability to various speaker configurations. Unlike methods reliant on linear properties [28], [29], MFCCs closely resemble the human auditory perception system. The correlation among coefficients is low [27], contributing to effective discrimination. However, a limitation of this approach is that it focuses solely on the power spectrum, potentially missing some crucial aspects of speech and providing only a limited representation of speech signals [28]. Additionally, MFCCs exhibit poor noise resistance [28], [29].

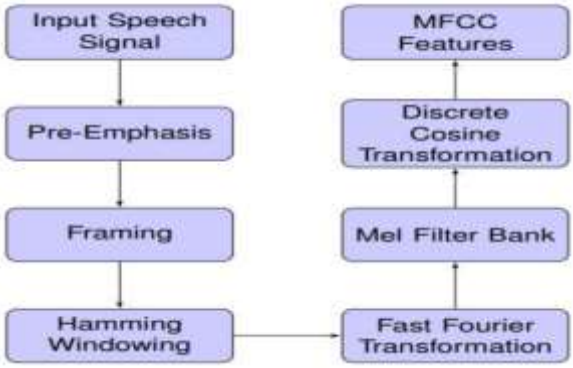

Figure 6: *Mel Frequency Cepstral Coefficient (MFCC) Feature Extraction Process [30].*

The initial voice signal input undergoes a process of segmentation into overlapping frames. Subsequently, windowing is applied, followed by the implementation of the Fast Fourier Transform. In the next stage, the frequency domain signal is converted to the Mel frequency scale. The transformation from the log Mel scale spectrum to the time domain is accomplished using the Discrete Cosine Transform (DCT) [31]. The result of this conversion is the Mel Frequency Cepstrum Coefficient (MFCC). MFCC primarily emphasizes the static properties of a signal.

2.7 Deep Neural Networks (DNNs)

Deep neural networks employ intricate mathematical calculations to process speech signals. These calculations are executed by interconnected nodes, or neurons, distributed across multiple layers. Each layer within the network serves a specific function, transmitting its output to the subsequent layer until the final output is generated. The initial layer in a deep neural network is the input layer, receiving the raw speech signal. The signal undergoes transformation into a series of numerical values, which are then passed on to the following layer. Subsequent layers, known as hidden layers, are responsible for learning the features embedded in the input data. Each neuron in a hidden layer receives inputs from multiple neurons in the preceding layer, conducting a weighted sum of these inputs. The weighted sum is then processed through an activation function, introducing non-linearity and enabling the network to comprehend intricate relationships between the input data and output labels. The ultimate layer in a deep neural network is the output layer, generating the desired

output based on the learned features from the hidden layers. In the context of speech signal processing, this output might represent phonemes, words, or other labels based on the specific task.

Throughout the training phase, the network refines its parameters, or weights, by evaluating the disparity between its predicted output and the actual output. This iterative adjustment, known as backpropagation, enables the network to learn from errors and enhance its performance progressively. Once trained on a substantial dataset, the network can make predictions on new, unseen data, a process termed inference. This utilization of deep neural networks in speech signal processing tasks underscores their effectiveness in learning intricate relationships within input data and generating accurate outputs. Deep neural networks have gained popularity in speech signal processing due to their capacity to automatically extract features from raw speech signals, eliminating the need for manual feature engineering [32]. This is accomplished through the intricate architecture of interconnected nodes across multiple layers, with each layer progressively learning more abstract representations of the input data.

In the realm of speech signal processing, deep neural networks find application in various tasks such as speech recognition, speaker identification, emotion recognition, and speech synthesis. For instance, in the context of speech recognition, deep neural networks play a crucial role in mapping the acoustic features of speech signals to corresponding phonemes or words [33]. However, a significant challenge in leveraging deep neural networks for speech signal processing lies in the necessity for substantial amounts of labeled training data [34]. The reason behind this lies in the extensive parameters deep neural networks require for effective training. Insufficient data may lead to overfitting on the training set, resulting in suboptimal performance on unseen data. To overcome this challenge, researchers have devised techniques like transfer learning, involving fine-tuning a pre-trained network on a large dataset for a specific task using a smaller dataset. Another strategy is data augmentation, which involves generating artificial data from existing data to augment the size of the training set. Despite these challenges, deep neural networks have exhibited promising results in various speech signal processing tasks and remain an active area of research [35]. As technology advances and access to larger datasets becomes more prevalent, there is an expectation that the performance of deep neural networks in these applications will continue to improve in the future [36].

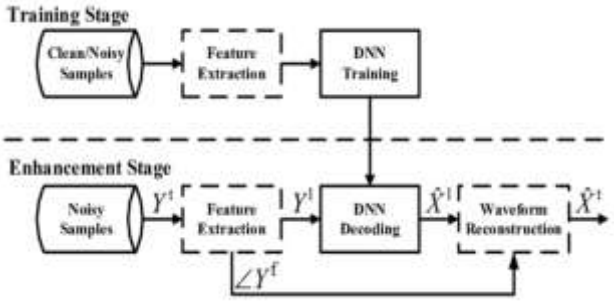

Figure 7: *Deep Neural Networks (DNNs) block diagram [36].*

2.8 Waveform Coding:

This technique is employed for speech transmission and compression, relying on the direct encoding of the waveform without the need for feature extraction. One advantage of this method is its capability to generate high-quality speech at low bitrates. However, a drawback is that both the encoding and decoding processes require substantial processing power.

3.0 Challenges and Limitations Associated with e Techniques in DSP

There are several challenges and limitations associated with speech signal processing techniques. Improving the quality of speech signals, particularly when the original signal is noisy or distorted, can be a complex task. The effectiveness of these techniques is influenced by factors such as the quality of recording equipment, the recording environment, and the type of processing applied.

Moreover, some of these techniques, especially complex ones like speech recognition or speaker identification, may demand significant computational resources and time. This can pose constraints, limiting their application in real-time scenarios or on low-power devices. Thus, a thorough evaluation and consideration of these techniques are essential to ensure they enhance speech signal quality effectively while minimizing computational demands and addressing inherent limitations.

4.0 Findings

Signal processing techniques play a crucial role in the analysis and interpretation of speech signals. While the Fourier transform is effective, it has limitations in capturing time-varying characteristics. Mel-Frequency Cepstral Coefficients (MFCCs) offer robust feature extraction but may not capture all relevant features. Hidden Markov Models (HMMs) are widely utilized but face challenges due to dependencies between observations and frame size. Deep Neural Networks (DNNs) provide advanced performance but require significant computational resources and may lack interpretability. Waveform coding proves effective for speech compression but demands high computational resources. Challenges in speech signal processing include noise, equipment quality, and computational demands, emphasizing the need to carefully choose the appropriate technique for each task by balancing strengths and weaknesses.

4.1 Recommendations.

Selecting the right signal processing technique necessitates careful consideration of data nature and analysis goals. A crucial aspect is balancing the advantages and limitations of each technique. Researchers should explore techniques addressing challenges like noise reduction, improving equipment quality, and developing efficient algorithms. Future research should focus on hybrid approaches that combine multiple signal processing techniques, such as Fourier transform and MFCCs, for a more comprehensive analysis. Integrating machine learning and artificial intelligence algorithms, especially deep learning techniques, can lead to more accurate and adaptable systems. Researchers should continue optimizing and interpreting deep neural networks for speech-related tasks.

5.0 Conclusions

Signal processing techniques are indispensable for analyzing speech signals, offering methods to enhance signal quality and extract valuable information. Each technique has its merits and drawbacks, underscoring the importance of choosing the appropriate one based on the task and available resources. While the Fourier transform is effective for frequency content analysis, it struggles with time-varying characteristics. MFCCs are robust for speech recognition, but their limited representation and noise resistance should be considered. Hidden Markov Models are widely used, but their limitations are related to dependencies between observations, frame size, and data requirements. Deep Neural Networks offer high performance but come with high computational demands and potential interpretability challenges. Waveform coding is effective for speech compression but requires significant computational resources. Challenges in speech signal processing, influenced by noise, equipment quality, and computational requirements, underscore the importance of careful evaluation and selection of techniques.

publishing platform. Additionally, the authors do not have any affiliation with any organization that has a direct or indirect financial stake in the subject matter discussed in this manuscript.

References

[1] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[2] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, PMLR, 2016, pp. 173–182.

[3] G. Heigold *et al.*, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 8619–8623.

[4] M. Kawata, M. Tsuruta-Hamamura, and H. Hasegawa, "Assessment of speech transmission index and reverberation time in standardized English as a foreign language test rooms," *Appl. Acoust.*, vol. 202, p. 109093, 2023.

[5] W. Álexander, "Phoneme Recognition Using Time-Delay Neural Network," *IEEE Trans. Acoust.*, vol. 37, no. 3, pp. 328–339, 1989.

[6] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 2149–2153, 2020.

[7] S. Latif, M. Shoukat, F. Shamshad, M. Usama, H. Cuayáhuitl, and B. W. Schuller, "Sparks of large audio models: A survey and outlook," *arXiv Prepr. arXiv2308.12792*, 2023.

[8] H. Zhou, A. Kan, G. Yu, Z. Guo, N. Zheng, and Q. Meng, "Pitch perception with the temporal limits encoder for cochlear Implants," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2528–2539, 2022.

[9] B. Gold, N. Morgan, and D. Ellis, *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[10] M. Lavechin *et al.*, "Given what babies really hear, both domain-general and domain-specific mechanisms are needed to simulate perceptual attunement".

[11] Harmonic analysis and the Fourier Transform available", [https://terpconnect.umd.edu/~toh/spectrum/Harmonic Analysis.html]. Accessed on 18th November, 2023.

[12] Retrieved oct, 31, 2023, from https://www.youtube.com/watch?v=q67z7PTGRi8&t=4294s.

[13] Khiatani, D., & Ghose, U. (2017, October). Weather forecasting using hidden Markov model. In 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), (pp. 220–225). IEEE.

[14] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech synthesis based on hidden Markov models. Proceedings of the IEEE, 101(5), 1234–1252.

[15] An Overview on Speech Recognition System and Comparative Study of its Approaches - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Block-diagram-of-hidden-markov-model_fig5_312578015 [accessed 18 Nov, 2023]

[16] Rabiner, L. R., & Juang, B. H. (1992). Hidden Markov models for speech recognition—strengths and limitations. In Speech recognition and understanding (pp. 3–29). Heidelberg: Springer.

[17] Hore, S., Bhattacharya, T., Dey, N., Hassanien, A. E., Banerjee, A., & Chaudhuri, S. B. (2016). A real time dactylology based feature extraction for selective image encryption and artificial neural network. In Image feature detectors and descriptors (pp. 203–226). Cham: Springer.

[18] Samanta, S., Kundu, D., Chakraborty, S., Dey, N., Gaber, T., Hassanien, A. E., & Kim, T. H. (2015, September). Wooden Surface classification based on Haralick and the Neural Networks. In 2015 Fourth International

Conference on Information Science and Industrial Applications (ISI), (pp. 33–39). IEEE.

[19] Kotyk, T., Ashour, A. S., Chakraborty, S., Dey, N., & Balas, V. E. (2015). Apoptosis analysis in classification paradigm: a neural network based approach. In Healthy World Conference (pp. 17–22).

[20] Agrawal, S., Singh, B., Kumar, R., & Dey, N. (2019). Machine learning for medical diagnosis: A neural network classifier optimized via the directed bee colony optimization algorithm. In U-Healthcare monitoring systems (pp. 197–215). Academic Press.

[21] Sarma, Kandarpa & Sarma, Mousmita. (2015). Acoustic Modeling of Speech Signal using Artificial Neural Network: A Review of Techniques and Current Trends. 10.4018/978-1-4666-8493-5.ch012.

[22] Retrieved oct. 29, 2023, from https://medium.com/@ageitgey/machine-learning-is-fun-part6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a.

[23] Pooja V. Janse, Ratnadeep R. Deshmukh, 2014, Design and Development of Database and Automatic Speech Recognition System for Travel Purpose in Marathi, OSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 5, Ver. IV, PP 97-104 https://doi.org/10.9790/0661-165497104

[24] Urmila Shrawankar, Techniques for Feature Extraction in Speech Recognition System: A Comparative Study, Available from: https://arxiv.org/ftp/arxiv/papers/1305/1305.1145.pdf

[25] Ms. Yogita A. More, Mrs. S. S. Munot(Bhabad), 2016, Effect Of Combination Of Different Features On Speech Recognition For Abnormal Speech, International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 5 Issues 8,Page No. 17590-17592

[26] Patino, Jose. Efficient speaker diarization and low-latency speaker spotting. 2019.

[27] Wenzhi Liao, Aleksandra Piurica, Paul Scheunders, Wilfried Philips, Youguo Pi,2013, Semisupervised Local Discriminant Analysis for Feature Extraction in Hyperspectral Images, IEEE Transactions On Geo Science And Remote Sensing , Vol. 51, No. 1. https://doi.org/10.1109/TGRS.2012.2200106

[28] Lahiru Dinalankara, 2017, Face Detection and Face Recognition Using Open Computer Vision Classifies. Available from: https://www.researchgate.net/publication/318900718

[29] Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal,2014, A Comparative Study of Feature Extraction Techniques for Speech Re cognition System, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Issue 12. https://doi.org/10.15680/IJIRSET.2014.031203

[30] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing 22, 10 (2014), 1533–1545.

[31] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012.

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29, 6 (2012), 82–97.

[32] Li, X., Wang, L., & Wang, Y. (2017). Deep neural network for speech and speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(11), 2086-2098.

[33] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. Advances in neural information processing systems 28 (2015).

[34] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing. Ieee, 6645–6649.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017).

[36] Sarma, Kandarpa & Sarma, Mousmita. (2015). Acoustic Modeling of Speech Signal using Artificial Neural Network: A Review of Techniques and Current Trends. 10.4018/978-1-4666 8493-5.ch012.