

A Systematic Literature Review of Machine Learning Approaches for Speech Feature Analysis: Trends, Applications, And Comparative Insights

Deepak Kumar^{1*}, Dr. Monika Khatkar²

¹Ph.D. Scholar, Department of Computer Science, K. R. Mangalam University, Sohna Rural, Gurugram, Haryana – 122103, India

*Corresponding Email Id: javadevdeepak@gmail.com

ORCID ID: <https://orcid.org/0009-0004-7124-873X>

²Assistant Professor, Department of Computer Science, K.R. Mangalam University, Sohna Rural, Gurugram Haryana- 122103, India
Email Id: khatkarmonika@gmail.com

ABSTRACT

Speech feature analysis is essential to intelligent human–computer interaction systems for automatic speech recognition, speaker identification, emotion detection, and healthcare diagnostics. In the recent decade, machine-learning techniques have advanced rapidly, replacing statistical modelling with data-driven methods that can handle high-dimensional and dynamic acoustic patterns. Support Vector Machines (SVM), Random Forests, and different Neural Network architectures have been integrated into speech processing frameworks faster due to increased computational power and big voice corpora. Given this backdrop, this work carefully reviews and synthesises current research on machine-learning algorithms for speech feature analysis, concentrating on their methodological foundations, performance outcomes, and application fields. To assess strengths, weaknesses, and contextual applicability, standard machine-learning models are compared to deep learning methods. This study uses a systematic literature review to examine peer-reviewed journal publications and conference proceedings from the past five to six years. Relevant voice feature extraction, categorisation, dataset utilisation, and algorithmic innovation studies were chosen. The findings show a clear shift toward deep neural networks, including convolutional and recurrent architectures, which perform better in large-scale and complex speech tasks, while SVM and Random Forests remain effective in low-resource and noise-robust environments. The review found that data availability, computational resources, and application needs determine model selection, not algorithm performance. The study helps academics and practitioners understand current trends, select models, and identify future research directions in robust, interpretable, and resource-efficient speech analysis systems.

Keywords: Speech Feature Analysis; Machine Learning; Support Vector Machines (SVM); Neural Networks; Random Forests

INTRODUCTION

Speech processing encompasses the computational analysis, interpretation, and creation of human speech signals, facilitating robots' ability to recognise, synthesise, and respond to spoken language. This domain relies on speech feature analysis to extract Mel-frequency cepstral coefficients (MFCCs), pitch, formants, and spectral properties from raw audio data for classification and recognition. Automatic speech analysis began with rule-based and statistical models, then evolved into Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), which greatly enhanced ASR systems [1, 2, 3]. Natural language processing and computational linguistics improved speech technology by merging syntactic and semantic modelling [4]. End-to-end architectures and neural-based text-to-speech systems have reached near-human quality in recent years, transforming speech technology [2]. ML's capacity to predict complicated, non-linear patterns in big datasets makes it important in voice systems, enabling stable performance in different

and noisy contexts. Modern AI-driven frameworks offer advanced speech classification, enhancement, and disorder detection applications, demonstrating the growing role of intelligent algorithms in speech analysis [5].

Problem statement

Environmental noise, channel distortions, speaker variability, accent diversity, emotional modulation, and the high-dimensional complexity of acoustic features make robust and generalised speech analysis difficult despite advances in speech processing technologies. Background noise and recording errors reduce recognition accuracy and model stability in real-world voice signals. Features modelling and categorisation are also complicated by inter- and intra-speaker heterogeneity in age, gender, dialect, speaking rate, and physiological parameters. Traditional statistical methods and rule-based systems struggle to capture non-linear and context-dependent patterns, requiring handcrafted features and domain-specific tuning that limit scalability and adaptability [6]. Deep learning has enhanced speech system representation, but data scarcity, interpretability, processing demands, and cross-domain generalisation still hamper effective implementation in varied situations [7]. These constraints require a comprehensive review of machine-learning methods to find more resilient and adaptive solutions.

Objective of this review

Support Vector Machines, Neural Networks, Random Forests, and other developing models used in speech feature analysis are critically examined in this paper. The study compares their performance and appropriateness across datasets, processing constraints, and adaptation to complicated speech situations. The review also seeks to highlight contemporary research trends, methodological advances, and hybrid modelling techniques. Finally, it discusses how these algorithms are used in automatic voice recognition, speaker verification, emotion detection, healthcare diagnostics, and intelligent human-computer interaction platforms.

METHODOLOGY - RESEARCH DESIGN OF THE SYSTEMATIC LITERATURE REVIEW

This study follows PRISMA for transparency, repeatability, and methodological rigour. The review examines 2020–2026 research in SVM, Neural Networks, Random Forests, and hybrid models for voice feature analysis.

Peer-reviewed journal and conference papers, studies using machine-learning or deep-learning approaches to model or analyse speech features, English-language publications, and empirical assessment metrics were included. Review papers, non-English publications, speech feature research, and untested pieces were excluded.

Searches were conducted on Scopus, Web of Science, IEEE Xplore, and ACM Digital Library using structured Boolean keyword combinations like ("speech features" OR "speech signal processing") AND ("machine learning" OR "SVM" OR "neural networks" OR "random forest") AND ("classification" OR "recognition" OR "analysis").

This systematic review adhered to PRISMA principles to guarantee transparency and rigour. A total of 160 records were discovered from Scopus (52), Web of Science (38), IEEE Xplore (46), and ACM Digital Library (24). Following the elimination of 32 duplicates, 128 records were evaluated based on title and abstract, resulting in the exclusion of 64 irrelevant research. Sixty-four full-text studies were evaluated for eligibility; thirty-six were rejected (fifteen not focused on speech features, eleven lacking empirical validations, and ten being review papers). A total of 28 high-quality empirical studies (2020–2026) were incorporated. The screening process utilised established inclusion and exclusion criteria, while quality assessment was based on dataset

transparency, methodological rigour, performance indicators, and standards of repeatability.

Identifying, screening, eligibility, and inclusion were PRISMA steps. Following duplication removal and eligibility checks, 28 high-quality studies were chosen for final synthesis. Author information, dataset, feature kinds, methods, evaluation metrics, and significant conclusions were extracted.

Quality was assessed by dataset transparency, methodological clarity, performance reporting, and reproducibility. To balance interpretation, publishing bias and selective reporting were evaluated.

Figure 1 shows the PRISMA flow diagram from 160 initially recognised records (2020–2026) to 28 final included studies following duplication removal, screening, and eligibility evaluation.

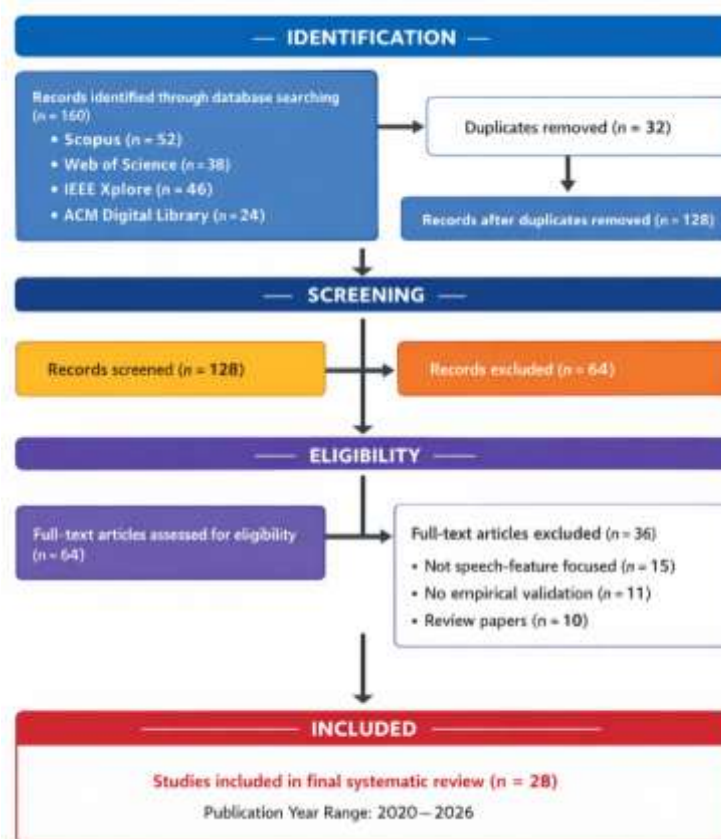


Figure 1. PRISMA Diagram followed in this study.

Fundamentals of Speech Feature Extraction

Any speech analysis system relies on speech feature extraction to shape and discriminate raw acoustic signals for machine-learning models. Signal processing must be resilient to collect relevant linguistic and paralinguistic information from dynamic, nonlinear, context-dependent speech. Modern voice recognition systems improve categorisation and prediction using systematic pre-processing, segmentation, and feature engineering [8].

Speech Signal and Characteristics

The lungs, vocal folds, and articulators work together to produce speech. Voice signals are analysed in short frames under quasi-stationarity because their statistical features shift with time. Acoustic signal representation can be done in the time domain to show amplitude fluctuations or in the frequency domain using spectral analysis to show energy distribution [9].

Phonetics and phonology refer to speech properties. Phonetics reflect articulatory and auditory aspects including pitch, voice, and articulation method, while phonological elements convey abstract linguistic contrasts that distinguish meaning. Automatic speech recognition and emotion detection require accurate modelling of these traits [10]. Variations in pronunciation, speaking rate, and context make representation difficult [6].

Feature Extraction Techniques

Speech feature extraction creates concise, useful vectors for machine-learning algorithms. Time-domain features, including zero-crossing rate, energy, and amplitude-based measurements, are computationally efficient but less resilient in noise. Frequency-domain features from Fourier or filter bank analysis capture spectrum properties and are more discriminative for speech recognition [9].

Mel-Frequency Cepstral Coefficients (MFCCs), which simulate human auditory perception, Perceptual Linear Prediction (PLP), and Linear Predictive Coding in vocal tract layout are popular characteristics. Prosodic elements, including pitch, intensity, and length, are essential for emotion and speaker recognition, while spectrograms show speech signals' time–frequency mapping [10]. Modern machine-learning frameworks use deep feature extraction to automatically generate hierarchical representations from raw waveforms [7]. Feature selection improves classification accuracy, reduces redundancy, and reduces computational cost, boosting model efficiency [11].

Figure 2 shows the evolution of voice processing models from rule-based to statistical, neural, deep learning, and generative AI models.

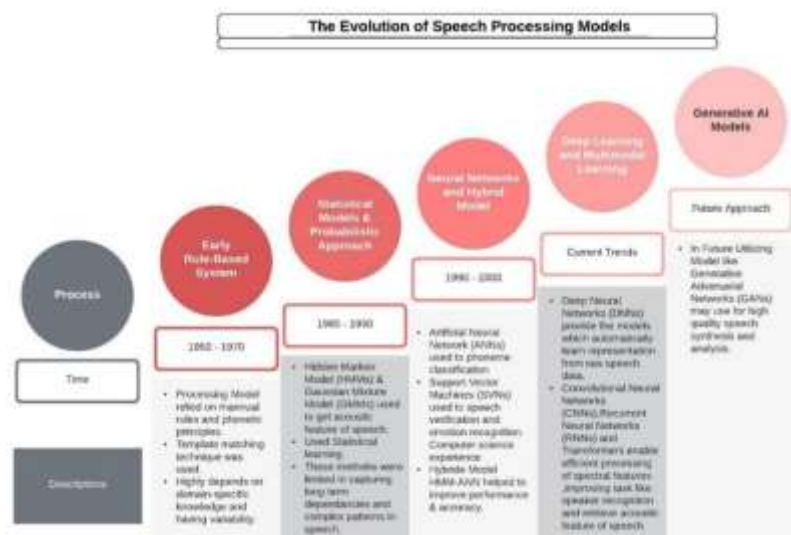


Figure 2. Evolution of Speech Processing Models from Rule-Based Systems to Generative AI Approaches.

Challenges in Feature Representation

Speech representation remains difficult despite feature engineering breakthroughs. Acoustic signals are distorted by environmental noise, reverberation, and recording artefacts, lowering recognition accuracy [6]. Speech signals are high-dimensional and

often contain redundant and correlated information, which can slow learning and complicate computation [7]. Inconsistencies from accent, dialect, gender, age, and mood complicate model generalisation [10]. By automatically learning robust features, deep representation learning has addressed some of these drawbacks, although data reliance, interpretability, and cross-domain adaptation remain research challenges [7]. In modern speech analysis systems, designing adaptive and noise-resilient feature extraction approaches is crucial.

Overview of Machine-Learning Methods for Speech Analysis

Machine-learning (ML) algorithms can predict complicated auditory and linguistic patterns from enormous datasets, making them essential to speech analysis. Traditional machine-learning methods use supervised and unsupervised learning. Using labelled voice data, supervised learning models like SVM, k-NN, Decision Trees, and Random Forests perform speaker identification, speech recognition, and emotion classification. Unsupervised learning methods like clustering and dimensionality reduction can find latent structures in unlabelled voice data, compressing and discovering patterns [12, 13]. These classical methods inspired newer deep learning frameworks and remain important in resource-constrained contexts. According to extensive reviews, conventional and deep learning models complement each other based on data and computing restrictions [12]. ML techniques have also been applied to part-of-speech tagging and semantic speech analysis, proving their versatility [14]. Figure 3 details speech collection, processing, and machine learning classification.

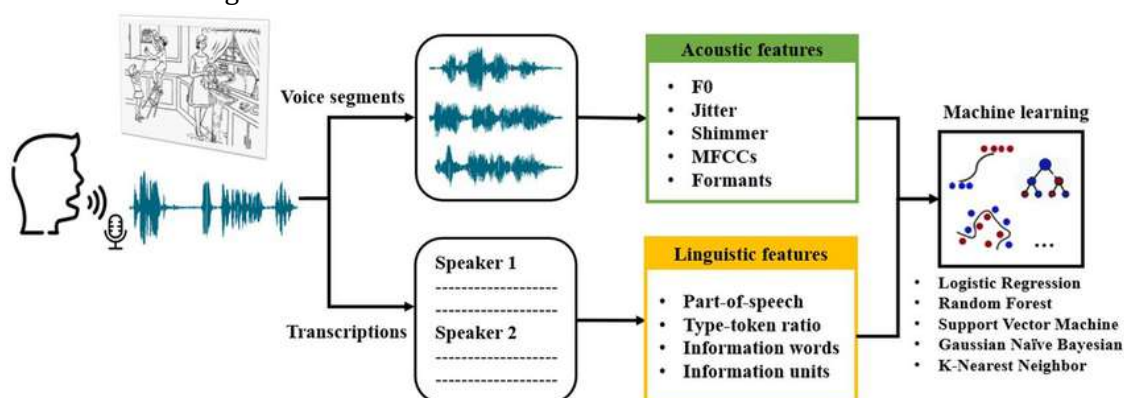


Figure 3. Overview of speech collection and analysis steps as well as machine learning classification [15].

Speech analysis system performance depends greatly on algorithm characteristics and job requirements, so model selection is crucial. The bias-variance trade-off is important because simpler models have high bias but low variance, while complicated models like deep neural networks have low bias but high variance if not regularised. When dealing with restricted or imbalanced speech datasets, model selection must balance generalisation power with overfitting concerns [12]. Deep learning architectures require large annotated corpora to build hierarchical representations, whereas classical ML models may work with handcrafted features and smaller datasets [13]. Real-time speech applications like voice assistants, call-centre analytics, and assistive devices involve latency and computing limits that affect algorithm choice. Cloud-based processing environments demand high-capacity architectures, while embedded systems prefer lightweight models. Thus, effective and scalable speech analysis systems require strategic model selection informed by bias-variance trade-offs, dataset scale, computational resources, and application demands [12], [14].

Figure 4 depicts how Speech Emotion Recognition (SER) uses machine learning to automatically identify and classify a speaker's emotional state by analysing acoustic and prosodic elements from speech signals.

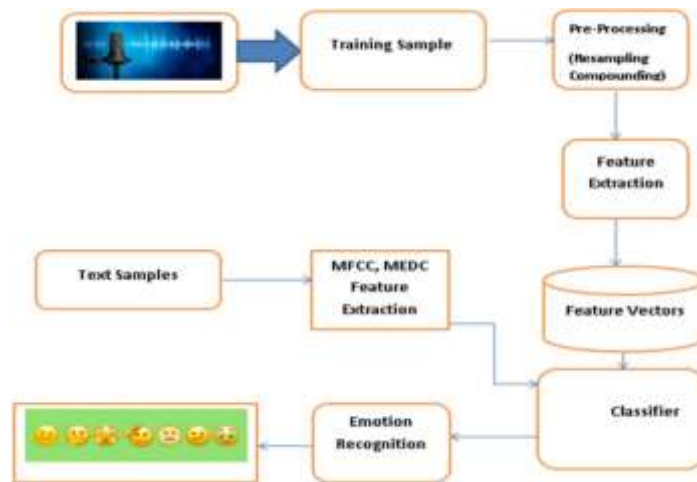


Figure 4. Speech Emotion Recognition using ML techniques – An overview [12].

Support Vector Machines (Svm) for Speech Analysis

SVMs are commonly used in speech analysis due to their strong theoretical foundation in statistical learning theory and capacity to handle high-dimensional feature spaces. SVM improves generalisation by finding the best hyperplane to maximise class margin. SVM can classify high-dimensional vectors from MFCCs, LPC, and prosodic characteristics in speech-related tasks with little training data. The SVM might be linear or non-linear. Linear SVM works well for basic classification tasks or well-processed feature sets with linearly separable voice features. However, coarticulation, speaker variability, and environmental noise make speech signals nonlinear. In non-linear SVM, kernel functions translate input features into higher-dimensional spaces to reduce complexity. Radial Basis Function (RBF) kernels capture localised patterns well, Polynomial kernels model feature interactions, and Sigmoid kernels resemble neural network behaviour. SVM is computationally efficient relative to deep learning because the kernel method allows nonlinear decision limits without high-dimensional mappings.

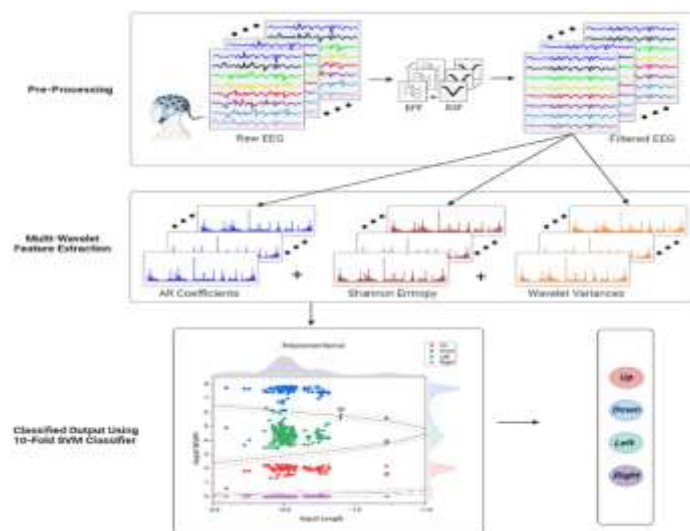


Figure 5. depicts the proposed SVM model for inner speech classification. A sequence of Band-Pass and Band Stop Filters the input time-series raw EEG signals. Before categorising EEG data with 10-Fold SVN Classifier, multi-wavelet feature extraction is done.

Figure 5: Architecture of proposed Support Vector Machine (SVM) model for classifying inner speech [16]

Speaker recognition often uses SVM classifiers to classify persons based on vocal characteristics using MFCC features. SVM models use spectral and prosodic information to classify speech emotions as happiness, anger, sadness, and neutrality. SVM also performs well in phoneme classification, utilising short-time spectral analysis feature vectors. Recently developed hybrid SVM models use feature selection methods like PCA or Genetic Algorithms to reduce dimensionality and enhance accuracy. A new trend is to use deep neural networks to extract high-level features and SVM as the final classifier. Deep learning's representational power is combined with SVM's margin-based classification performance in this hybrid architecture. SVM generalises well with small or medium datasets and over fits less due to margin maximisation. In high-dimensional spaces, it excels in binary classification. The kernel choice, regularisation parameter C, and gamma in RBF must be tuned for SVM, which might be computationally costly. Very big datasets increase training time relative to parallel-optimized deep neural networks, making scalability a problem.

Table 1. Key Studies & Findings on SVM in Speech Analysis [16, 17, 18].

Study / Approach	Feature Technique Used	Application Domain	Findings / Performance Outcome
SVM + MFCC-based Classification	MFCC (Mel-Frequency Cepstral Coefficients)	Speaker Identification	Achieved high accuracy (>90%) in controlled environments; robust to moderate noise; suitable for small datasets.
SVM with Prosodic & Spectral Feature Fusion	MFCC + Pitch + Energy + Formants	Speech Emotion Recognition	Improved emotion classification accuracy (5–10% gain) compared to single-feature systems; effective in multi-class problems.
Kernel-based SVM (RBF) for Phoneme Recognition	LPC + Spectral Features	Phoneme Classification	Demonstrated strong nonlinear separation capability; outperformed linear classifiers in complex acoustic datasets.

Hybrid Deep Feature Extraction + SVM	CNN-extracted deep features + MFCC	Emotion & Speaker Recognition	Hybrid model achieved higher precision and F1-scores compared to standalone SVM or CNN; reduced overfitting.
SVM with PCA Feature Reduction	MFCC + PCA	Speech Recognition Subtasks	Reduced computational cost and improved classification stability; effective for embedded and real-time systems.

Overall, SVM is a robust and reliable machine-learning method for speech analysis, especially with little data, well-engineered features, and moderate processing resources. SVM is useful in specialised tasks like emotion detection, speaker verification, and phoneme classification where interpretability and generalisation are important, even though deep learning models dominate large-scale voice recognition systems.

Neural Networks for Speech Feature Analysis

Automatic hierarchical and nonlinear representation extraction from speech data using neural networks has greatly improved speech feature analysis. The early Perceptron model used linear classification, while the Multilayer Perceptron (MLP) used hidden layers to simulate complex decision boundaries. Nonlinear activation functions like sigmoid, tanh, and ReLU let networks capture complex acoustic variations. Backpropagation enabled gradient descent-based network weight optimisation, becoming the computational backbone of deep learning [19].

Deep Neural Networks (DNNs) revolutionised speech processing. In automatic voice recognition, fully connected DNNs improved acoustic modelling. Learning spatial patterns from spectrograms and MFCC features helped CNNs perform better in emotion recognition [20]. RNNs and LSTM networks addressed speech sequence temporal relationships, improving contextual information modelling. Modern speech frameworks use transformer designs with self-attention methods to efficiently capture long-range dependencies.

Neural network models are used in ASR, speaker verification, emotion detection, and language identification. They adapt well to healthcare settings and are widely employed in clinical speech analysis for neurological and pathological assessment [19, 21]. CNN-RNN hybrid designs combine spatial and temporal modelling strengths, while attention processes optimise model emphasis on prominent audio regions. Neural networks outperform standard machine-learning approaches in accuracy, especially when trained on huge datasets, but they demand a lot of computer power and data.

The figure 6 below depicts how a vocal signal is extracted (time and frequency domain data), trained using machine learning or neural network models, and identified as angry, happy, sad, neutral, or terrified.

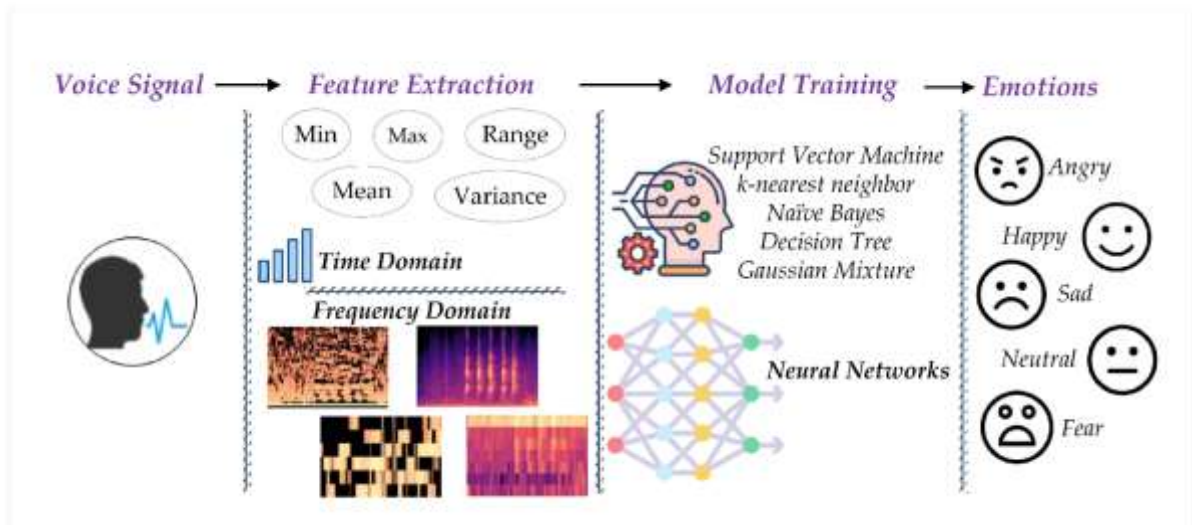


Figure 6. An illustration of the SER process using neural networks.

Table 2. Key Deep Speech Frameworks and Applications [19, 20, 21].

Framework / Model Type	Core Architecture	Application Area	Key Contribution
Deep Acoustic Modeling System	Fully Connected DNN	Automatic Speech Recognition	Improved acoustic modeling accuracy over traditional statistical models.
Deep CNN-Based Emotion System	Convolutional Neural Network	Speech Emotion Recognition	Enhanced classification accuracy through spatial feature learning and feature optimization.
Clinical Speech Analysis Model	Hybrid Deep Neural Networks	Neurological & Pathological Speech Assessment	Enabled robust detection of speech impairments for clinical decision support.
wav2vec Framework	Transformer-based Self-Supervised Model	ASR & Low-Resource Speech Processing	Learned powerful pretrained speech representations with minimal labeled data.
DeepSpeech System	End-to-End RNN Architecture	Large-Scale Speech-to-Text Systems	Simplified speech pipeline with competitive end-to-

			end transcription performance.
--	--	--	--------------------------------

Overall, Neural networks dominate voice feature analysis, providing scalable, flexible, and high-performing solutions for a variety of speech applications.

Decision Tree and Ensemble Models for Speech Recognition

Smart supervised models like Decision Trees identify speech using MFCCs, pitch, energy, and formants, partitioning data by Gini impurity or information gain. Pruning reduces speaker and environmental variability-induced overfitting, improving generalisation. Bagging and random feature selection build robustness in high-dimensional audio applications like emotion and language detection in Random Forests [23, 24, 25]. AdaBoost and Gradient Boosting identify misclassified samples to reduce bias and improve difficult paralinguistic categorisation [24]. Compared to single trees, ensemble models improve accuracy and stability but increase computing cost and impair interpretability.

Figure 7 shows an audio processing pipeline that feeds CNN and Bi-LSTM models pre-processed and normalised features, then averages their outputs to predict emotions.

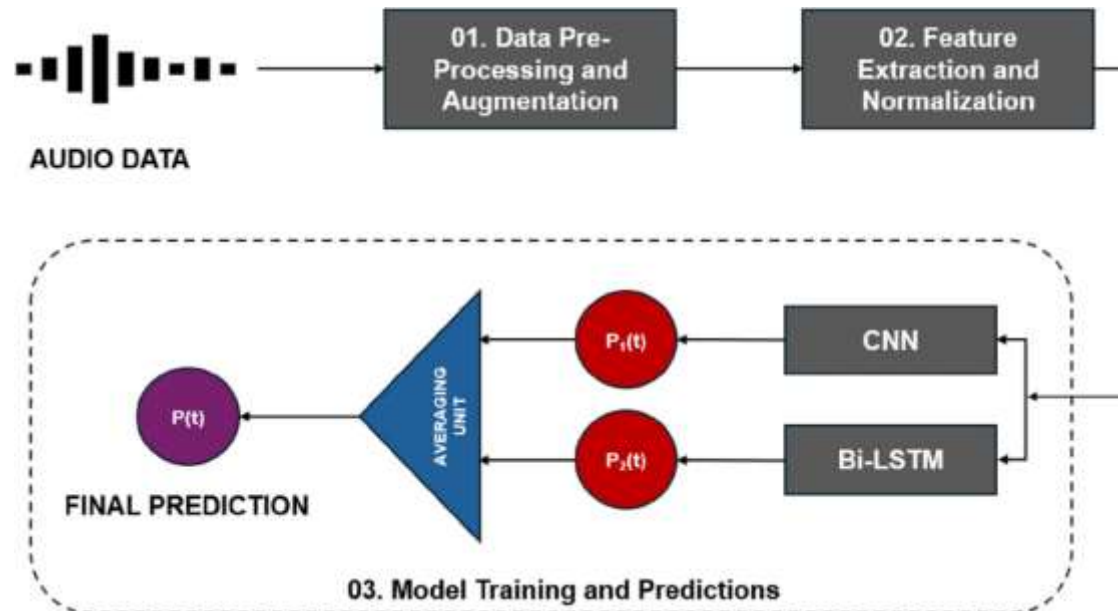


Figure 7. Audio processing pipeline using CNN and Bi-LSTM models [26].

Table 3. Key Studies on Random Forest in Speech Emotion Recognition [23, 24, 25].

Study Focus	Feature Type	Model Used	Key Outcome
Combined spectral and prosodic features	Spectral + Prosodic	Random Forest	Improved classification accuracy through multimodal feature fusion [23]
Cross-corpus emotion recognition	Multilingual acoustic features	Hybrid Ensemble + Deep Models	Enhanced generalization

			across datasets [24]
Feature representation comparison	TF-IDF vs Count Vectorizer	Random Forest	Demonstrated stability under random state variation [25]
Emotion detection system	Acoustic & statistical features	Random Forest Ensemble	Robust handling of noisy speech environments

Overall, Decision Trees and ensemble algorithms are strong, interpretable, and competitive speech feature analysis tools, especially when paired with different acoustic representations and huge datasets.

Comparative Evaluation of Algorithms for Speech Recognition

Multidimensional voice recognition algorithm evaluation considers predicted performance, resilience, and computational efficiency. Accuracy, Precision, Recall, and F1 Score quantify phoneme and word-level categorisation performance in traditional and deep learning-based ASR systems. In imbalanced datasets, accuracy may be deceiving, whereas F1 Score balances false positives and negatives. ROC and AUC curves evaluate keyword detection and speaker verification discrimination [26, 27, 28]. Practical systems must be scalable, low-latency, and computationally affordable beyond precision. Transformer-based models are more accurate, although HMM and GMM models are computationally efficient but less robust in noise.

The following table summarizes key evaluation criteria used in comparing speech recognition algorithms:

Table 4. Key evaluation criteria's.

Evaluation Criterion	Definition	Purpose in Speech Recognition	Impact on System Design
Accuracy	Ratio of correct predictions to total samples	Measures overall correctness	May mask class imbalance
Precision	True positives / predicted positives	Evaluates prediction reliability	Important for keyword detection
Recall	True positives / actual positives	Measures detection capability	Reduces missed speech events
F1 Score	Harmonic mean of precision and recall	Balances false positives and negatives	Suitable for imbalanced datasets
ROC Curve	Plot of true positive vs false positive rate	Visual performance comparison	Helps threshold tuning
AUC	Area under ROC curve	Measures separability	Higher AUC indicates better discrimination

Word Error Rate (WER)	Edit distance between predicted and reference text	Standard ASR metric	Lower WER = better transcription
Phoneme Error Rate	Error rate at phoneme level	Fine-grained acoustic evaluation	Useful in low-level modeling
Scalability	Ability to handle increasing data/users	Evaluates robustness	Influences architecture choice
Latency	Time delay in producing output	Critical for real-time systems	Affects user experience
Training Time	Time required for model training	Measures efficiency	Impacts deployment cycle
Inference Time	Time per prediction	Determines responsiveness	Important for streaming ASR
Memory Usage	RAM/VRAM consumption	Hardware requirement indicator	Limits edge deployment
Computational Cost	CPU/GPU resource consumption	Operational expense measure	Impacts scalability
Robustness to Noise	Performance under noisy conditions	Evaluates generalization	Essential for real-world use

Overall, Speech recognition systems are accurate, efficient, scalable, and suitable for real-world implementation after comparative examination.

Emerging Trends and Future Directions

Deep representation learning, especially end-to-end and self-supervised learning (SSL) models that learn directly from raw audio and decrease labelled data, drives speech processing trends. Hybrid CNN frameworks and spectral-prosodic features improve speech emotion recognition [22, 23], whereas cross-corpus deep models increase generalisation [24]. Lightweight ensembles from transfer learning and pretrained models reduce data needs and boost efficiency [25, 26]. Explainability and reliable benchmarking are needed as healthcare and enterprise deployments expand [27, 28]. Future study must address low-resource languages, cross-domain adaptation, and noise/bias resilience.

CONCLUSION

Machine learning and deep learning have improved speech recognition and emotion interpretation. From feature-based models to end-to-end deep neural networks and self-supervised learning frameworks, current systems improve accuracy, scalability, and robustness. Ensemble approaches, hybrid architectures, and pretrained models improve performance without big labelled datasets. Low-resource languages, cross-domain adaptation, interpretability, and prejudice remain issues. Future research will produce efficient, explainable, and fair speech systems that can be deployed reliably in different real-world and multilingual situations.

Acknowledgement

The authors acknowledge the support of the Centre of Excellence in Artificial Intelligence (CoE-AI), K.R. Mangalam University, Gurugram, for providing a conducive research environment and technical support for this study.

REFERENCES

1. Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6), 9411-9457.
2. Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., ... & Liu, T. Y. (2024). NaturalSpeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6), 4234-4245.
3. Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., ... & Almojel, M. (2021). Automatic speech recognition: Systematic literature review. *Ieee Access*, 9, 131858-131876.
4. Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021, March). Natural language processing: History, evolution, application, and future work. In *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020* (pp. 365-375). Singapore: Springer Singapore.
5. Remya, M. S., Raman, R., Sankaran, R., Namboodiri, V., & Nedungadi, P. (2025). Artificial intelligence for speech classification and enhancement of speech and language disorders: Techniques, applications, and future directions. *IEEE Access*.
6. Basak, S., Agrawal, H., Jena, S., Gite, S., Bachute, M., Pradhan, B., & Assiri, M. (2023). Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems. *CMES-Computer Modeling in Engineering and Sciences*.
7. Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2020). Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378*.
8. Xiong, X. I. A. O. (2023, March). Fundamentals of speech recognition.
9. Kacur, J., Puterka, B., Pavlovicova, J., & Oravec, M. (2021). On the speech properties and feature extraction methods in speech emotion recognition. *Sensors*, 21(5), 1888.
10. Sönmez, Y. Ü., & Varol, A. (2024). In-depth investigation of speech emotion recognition studies from past to present–The importance of emotion recognition from speech signal for AI–. *Intelligent Systems with Applications*, 22, 200351.
11. Abdusalomov, A. B., Safarov, F., Rakhimov, M., Turaev, B., & Whangbo, T. K. (2022). Improved feature parameter extraction from speech signals using machine learning algorithm. *Sensors*, 22(21), 8122.
12. Tyagi, S., & Szénási, S. (2024). Semantic speech analysis using machine learning and deep learning techniques: a comprehensive review. *Multimedia Tools and Applications*, 83(29), 73427-73456.
13. Palo, H. K., & Sarangi, L. (2020). Overview of machine learners in classifying of speech signals. In *Handbook of research on emerging trends and applications of machine learning* (pp. 461-489). IGI Global Scientific Publishing.
14. Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1), 10.
15. Huang, L., Yang, H., Che, Y., & Yang, J. (2024). Automatic speech analysis for detecting cognitive decline of older adults. *Frontiers in Public Health*, 12, 1417966.
16. Abdulghani, M. M., Walters, W. L., & Abed, H. K. (2024). Enhancing the classification accuracy of EEG-informed inner speech decoder using multi-wavelet feature and support vector machine. *IEEE Access*, 12, 147929-147941.
17. Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*.
18. Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification a review. *Qubahan Academic Journal*, 1(2), 81-90.

19. Zaidi, B. F., Selouani, S. A., Boudraa, M., & Sidi Yakoub, M. (2021). Deep neural network architectures for dysarthric speech analysis and recognition. *Neural Computing and Applications*, 33(15), 9089-9108.
20. Farooq, M., Hussain, F., Baloch, N. K., Raja, F. R., Yu, H., & Zikria, Y. B. (2020). Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors*, 20(21), 6008.
21. De Silva, U., Madanian, S., Olsen, S., Templeton, J. M., Poellabauer, C., Schneider, S. L., ... & Rubaiat, R. (2025). Clinical decision support using speech signal analysis: systematic scoping review of neurological disorders. *Journal of Medical Internet Research*, 27, e63004.
22. Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech emotion recognition through hybrid features and convolutional neural network. *Applied Sciences*, 13(8), 4750.
23. Alamri, F. S., Mirdad, A. R., Saba, T., Raza, M. A., & Siddique, U. (2026). Investigating the Impact of Combined Spectral and Prosodic Features on Speech Emotion Recognition. *IEEE Access*, 14, 5719-5732.
24. Almadhor, A., Belhaj, S., Alsubai, S., Baili, J., Hejaili, A. A., Gregus, M., & Ivanochko, I. (2026). Cross-corpus language-independent speech emotion recognition using hybrid deep learning framework. *Complex & Intelligent Systems*.
25. Kooptiwoot, S., & Kooptiwoot, S. (2026). A Comparative Study of TF-IDF and Count Vectorizer under Random State Changes in a Random Forest Classifier for Emotion Detection. *Engineering, Technology & Applied Science Research*, 16(2), 33247-33252.
26. Chowdhury, J. H., Ramanna, S., & Kotecha, K. (2025). Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. *Scientific Reports*, 15(1), 11824.
27. Ng, J. J. W., Wang, E., Zhou, X., Zhou, K. X., Goh, C. X. L., Sim, G. Z. N., ... & Ng, Q. X. (2025). Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. *BMC medical informatics and decision making*, 25(1), 236.
28. Trabelsi, A., Warichet, S., Aajaoun, Y., & Soussilane, S. (2022). Evaluation of the efficiency of state-of-the-art Speech Recognition engines. *Procedia Computer Science*, 207, 2242-2252.