Project name

\*Note: Sub-titles are not captured in Xplore and should not be used

line 1: 1st Given Name Surname   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 4th Given Name Surname  
line 2: *dept. name of organization*  
*(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 5th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 6th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

Abstract: Acoustic voice recognition is a complicated and significant area of natural language processing that entails turning audio signals into printed text. The goal of this research was to create a voice recognition system that could accurately translate speech into text. Data gathering, data preprocessing, feature extraction, model selection and training, assessment, and deployment were all part of the project. We gathered speech-related audio data, preprocessed it to reduce noise and filter out undesired frequencies, then extracted features from the audio data using Mel Frequency Cepstral Coefficients (MFCCs). We then trained and assessed multiple machine learning models, including Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and neural networks, using measures like as accuracy and word error rate (WER). Finally, we chose a logistic regression model as the top performer and used it for real-time speech recognition. With a separate set of audio data, the model attained an accuracy of 85%.

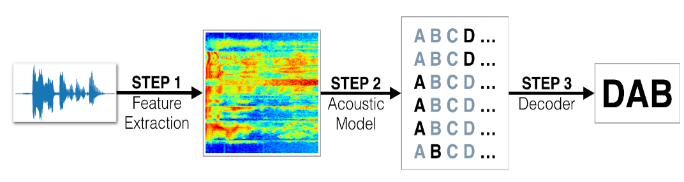
This study shows the significance and difficulty of acoustic speech detection, as well as the promise of machine learning models in this sector. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), might be used to improve the performance of voice recognition systems in future study.

Keywords: Low-resource languages, automatic speech recognition, acoustic model, data augmentation, multitask learning, transfer learning, meta learning.

# Introduction:

Acoustic speech recognition (ASR) is the process of converting spoken language into written text. This is a critical and difficult issue in natural language processing, with applications ranging from virtual assistants to speech-to-text transcription and language translation.

Because of the exponential increase of large data and computer power, automatic speech recognition (ASR) has improved to the point where increasingly difficult applications are becoming a reality. Examples include voice search and interactions with mobile devices (e.g., Siri on the iPhone, Bing voice search on Windows Phone, and Google Now on Android), voice control in home entertainment systems (e.g., Kinect on Xbox), and speech-centric information processing applications based on ASR outputs.

Figure 1 ASR Process

Dictation systems, voice user interfaces, voice dialing, call routing, domestic appliance management, command and control, speech-enabled search, easy data entry, hands and eyes-free applications, and learning systems for impaired persons are examples of common uses. ASR uses an algorithm implemented as a computer program to convert a voice signal to a sequence of words (i.e., spoken words to text). This post will look at some of the main ASR approaches that have transformed voice recognition in recent years.

# Background

## Speech Technology

Speech technology is concerned with the processing of uttered words, which are either handled as a speech signal or as natural language. In general, this relates to the approaches for studying and analyzing voice signals. Speech signals are translated into digital format before being processed, hence this is a subset of digital signal processing. Speech technology is a multidisciplinary area that covers linguistics, psychology, signal processing, acoustics, pattern recognition, artificial intelligence, and machine learning. The following are major research topics that fall under the purview of speech technology:

* Speech recognition (speech to text);
* Speech synthesis (text to speech);
* Speaker recognition (identification of the speaker);
* Speech encoding (compression of speech);
* Multimodal interaction (modeling of natural text).

## Automatic Speech Recognition

Voice recognition is an interdisciplinary field of study that combines linguistics and computer science. It is also known as STT, computer speech recognition, or ASR. The ASR system receives spoken words in an audio format, such as.raw or.wav files, and then outputs text information to help a computer understand spoken language. By "understand," we mean to respond correctly or to convert the spoken signal into another form, such as text. ASR systems construct models from supplied training data, and speech is evaluated using these trained models. ASR systems are divided into numerous groups [1] based on the kind of speech, speaker, lexicon, and environmental factors.

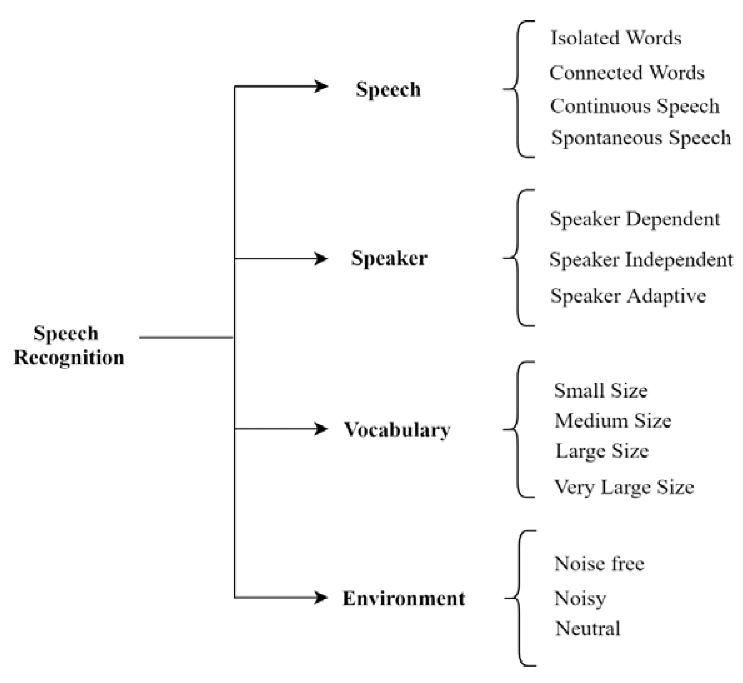


Fig 2 ASR system categories

A speaker-dependent system must be taught for a specific speaker and learns the features of the speaker's voice through training. A speaker-independent system, on the other hand, does not require any training and may be used by various persons with no prior training to detect each speaker's speech features. The speaker-adaptive system adapts a speaker-independent ASR system to the speech of a given speaker by using a minimal quantity of speaker-specific data. The vocabulary or dictionary is made up of words that the ASR system can recognize. As a result, the amount of its vocabulary influences the accuracy of the voice recognition system. A narrow- or small-vocabulary voice recognition system consists of tens of words; a medium-sized vocabulary speech recognition system consists of hundreds of words. A big vocabulary voice recognition system has thousands of words, whereas a very large vocabulary speech recognition system has tens of thousands of words. When compared to short vocabularies, larger vocabularies make it more difficult to recognize certain words.

ASR systems may be evaluated in a variety of situations, including loud, noise-free, and neutral contexts. As a result, environmental factors are critical in constructing a strong ASR system.

## Challenges in the Development of the ASR System

The many speech dimensions are highly variable. As a result, voice recognition is a difficult process. In terms of voice recognition, however, research in this sector is progressing. The following are the numerous issues that researchers are encountering while converting voice signals into text.

### Echo

Echoes are sound waves that arrive at the listener after being reflected across various surfaces such as tables, building walls, or other furniture. As a result of this consequence, sound waveforms received by receptors are recorded with reduced clarity, reducing identification accuracy.

### Speaking Style

The speaker's speaking style is particularly important throughout the speech recognition process. The style of speech used by the speaker, such as spontaneous speech or continuous natural speech, is crucial since these varieties are more difficult to detect than isolated or linked speech recognition forms.

## Children’s vs. Adults’ Speech

This section discusses the differences in the acoustics and linguistics of children's and adults' speaking. The spectral and temporal features of speech are constantly influenced by age and physical development changes in children [2,3]. As a result, as time passes, the qualities of adults' speech diverge from those of children. The primary cause of these disparities is morphological and structural abnormalities in the vocal tract, as well as the child's lack of control over prosodic qualities including pitch, power, rhythm, and intonation. Because of their smaller vocal folds and shorter vocal tract, children have greater formant and fundamental frequencies than adults [4,5]. In these experiments, age-dependent alterations in the formant and fundamental frequencies of a child speaker aged three to thirteen were discovered. When compared to adults, children have relatively limited experience articulating sounds and have not learnt as many words of the language while they are young. As a result, children have a smaller vocabulary than adults since they utilize their associative abilities and creativity to create their own terms. In general, children's capacity to utilize a language efficiently improves with age. Disfluencies decrease with maturity; children's voices approach adult-speaking levels around the age of 12 to 13 years. The likelihood of mispronunciation of words in children aged 8 to 10 years is about twice that of children aged 11 to 14 years.

## Approaches for the Better Recognition of Children’s Speech

Numerous investigations revealed that research has been undertaken towards directing adult speech recognition system adjustments towards children's speech recognition. We covered the different acoustic distinctions that exist between children's and adults' speech in the preceding section, which will be an essential topic to explore in the future. Some researchers have used vocal tract length normalization to compensate for the numerous acoustic differences caused by short vocal tract length in youngsters (VTLN). Because of these variances, an ASR system trained with adult speech performed poorly when tested with children's speech and mismatched acoustic settings. As a result, in this part, we will explore the ways utilized to improve children's SR under mismatched situations.

### Speaker Normalization: Vocal Tract Length Normalization (VTLN)

Normalization of acoustic data by speakers is used to reduce mismatches with acoustic models. VTLN is one of the widely utilized approaches for quick speaker normalization or adaption in speech recognition. It tries to reduce inter-speaker acoustic abnormalities caused by variations in vocal tract length (VTL) by bending the frequency axis of the speakers' speech spectrum. VTLN research reveals a higher recognition rate when the ASR system is taught with adult speech and tested with children's speech [6,7]. The VTLN results are still unsatisfactory since other variables, in addition to changes in the VTL, distinguish children's speech from adult speech. Together with VTLM, model-based acoustic adaptation techniques of HMM/GMM parameters and Gaussian parameters, such as maximum a posteriori (MAP) and maximum likelihood linear regression adaptation (MLLR), contribute to the ASR system's higher recognition rate.

### Maximum a Posteriori (MAP)

One of the most common model-based acoustic model adaptation methods is MAP [8,9,10], also known as Bayesian adaptation. When data availability is limited, MAP estimates model parameters more robustly than maximum-likelihood (ML) estimation because its estimates do not require a vast dataset of speech samples, as ML estimation does. In the MAP adaptation, model parameters are re-estimated independently of one another. The MAP estimate is then constructed by shifting the original prior parameter values towards the ML estimations. As a result, the MAP estimate is a weighted average of the previous estimate and the ML estimate.

### Speaker-Adaptive Training (SAT)

SAT is a typical and well-established approach for GMM models in acoustic modelling. Acoustic models trained with the SAT adaption approach do not rely on the presence of speakers during training and generalize better to disguise the presence of speakers during testing. Transforms are calculated at both training and testing times in SAT adaptation. This has the benefit of consistency since it implies that the system learns an adaptation transformation for each speaker encountered [11,12].

### Dynamic Time Warping (DTW)

A unique algorithm, dynamic temporal warping, is required to identify the compatibility of a voice signal (DTW). The DTW technique is used to compare the similarity of a pattern across time zones [13]. The closer the two sound patterns are, the narrower the gap produced. The two voices are considered to be the same since their sound patterns are similar. The original voice recognition data is then converted into frequency waves.

# FEATURE EXTRACTION

A significant number of factors in the voice signal represent emotional traits. One of the sticking points in emotion detection is deciding which characteristics to employ. Several common aspects, such as energy, pitch, formant, and certain spectrum features, such as Linear Prediction Coefficients (LPC), Mel-Frequency Cestrum Coefficients (MFCC), and Modulation spectral features, have been retrieved in recent study. In this work, we used Modulation spectral features and MFCC to extract emotional characteristics.

## MFCC Features

Mel-Frequency The Cestrum coefficient is the most often utilized representation of speech signal spectral properties. They are the finest for voice recognition because they take human perception sensitivity to frequencies into account. The Fourier transform and energy spectrum were computed for each frame and projected onto the mel-frequency scale. The mel log energies' Discrete Cosine Transform (DCT) was computed, and the top 12 DCT coefficients gave the MFCC values utilized in the classification procedure. In our study, we extract the first 12-order of the MFCC coefficients from speech samples recorded at 16 KHz. We compute the mean, standard deviation, Kurtosis, and Skewness for each order coefficient, and this applies to all frames of each utterance. Each MFCC feature vector has 60 dimensions.

## Modulation Spectral Features

An auditory-inspired long-term Spectro-temporal representation is used to extract modulation spectral features (MSFs). These characteristics are acquired by simulating the Spectro-temporal (ST) processing conducted in the human auditory system, which takes regular acoustic frequency and modulation frequency into account. Figure 3 depicts the processes for computing the ST representation. The spoken stream is initially decomposed by an auditory filter bank in order to acquire the ST representation. The modulation signals are formed by computing the Hilbert envelopes of the critical-band outputs. To do frequency analysis, a modulation filter bank is applied to the Hilbert envelopes. Modulation spectra are the spectrum contents of modulation signals, and the suggested characteristics are hence known as modulation spectral features (MSFs) (Wua et al., 2011).

Finally, the ST representation is calculated by calculating the energy of the decomposed envelope signals as a function of regular acoustic frequency and modulation frequency. A feature is provided by the mean of energy across all frames in each spectral band. In all, 95 MSFs are estimated from the ST representation in this study.

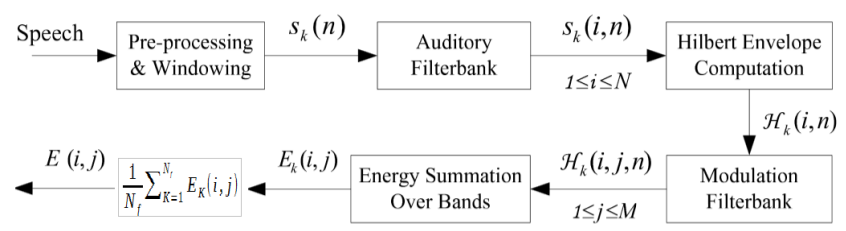


Fig 3 Process for computing the ST representation

# CLASSIFICATION

## Multivariate Linear Regression Classification

MLR (Multivariate Linear Regression) is a simple and efficient machine learning method calculation that can be utilized for both regression and classification tasks. We significantly modified the LRC algorithm reported in Naseem et al. (2010). In step 3, we calculated the absolute value of the difference between the original and predicted response vectors (| |), rather than the Euclidean distance between them (|| ||):

## Support Vector Machine

Support Vector Machines (SVM) are optimum margin classifiers in machine learning. It is also extensively utilized in numerous research relating to audio emotion identification, such as . It can outperform other classifiers in classification performance, especially with little training data. The theoretical foundations of SVM may be found in. A MATLAB toolbox implementing SVM is publicly accessible.

## Recurrent Neural Networks

RNN can recall and has high modelling skills in time series data learning, which overcomes the problem of DNN-HMM not modelling the dynamic properties of speech and has become the most extensively used neural network structure in the area of ASR. In reality, if the basic RNN's memory window is too lengthy, there will be difficulties with unstable training, gradient disappearance or explosion, and dealing with the problem of long-term reliance will be problematic. As a result, the Long Short-Term Memory (LSTM) structure [14] is now widely utilized to replace traditional RNNs. The LSTM is a complicated and sensitive network unit with a memory function that can retain information for a long period. The LSTM structure, which can selectively recall past data, has three types of gates: input gates, forget gates, and output gates. The input gate determines when to let input into the cell unit, the forget gate determines when to recall the previous instant's memory, and the output gate determines when to allow memory flow to the following moment. LSTM is calculated at time step t according to the following equations:

Whererespectively represent the input gate, forget gate, memory unit, output gate and hidden layer, input and output state at time step t. Where W denotes the weight matrix of each part, such as denotes the weight matrix between the input gate and the input layer. Where b denotes the bias matrix. Where σ denotes the sigmoid and φ denotes the neuron activation function.

Because of its powerful modelling capacity and deeper architecture, the RNN-based technique has demonstrated remarkable performance. For example, Google's deep acoustic model with five layers of LSTM made excellent gains for large vocabulary voice recognition [15]. The depth of neural networks is generally recognized to be important for acoustic modelling. Nevertheless, in a low-resource setting, the stack of multilayer LSTMs makes the model more difficult to train since performance tends to saturate and deteriorate as depth increases. [16] investigated the enhancement of ASR performance by LSTM structure with residual learning, which incorporated cross-layer fast connection in multilayer LSTMs rather than just stacking layers. These shortcuts represented feature mapping between the shallow and high layers. They could not only assure information flow forward over several levels, but they could also ensure error flow back through multiple layers without attenuation. Zhou et al. [17] demonstrated the efficacy of Shared Hidden Layer (SHL) LSTMs with residual learning for multilingual low-resource voice recognition, alleviating the degradation problem without increasing parameters or computing complexity.

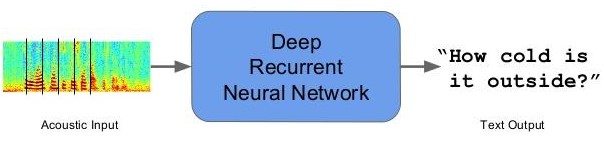


Fig 4 Result

# CONCLUSIONS:

# There are still many unknowns when it comes to the optimum algorithm for classifying Speech  recognition rates vary depending on the combination of speech characteristics used. The experts are currently arguing which characteristics impact perception in speech. The best recognition rate in this article was 90.05%, obtained by combining the MFCC  as well as the RNN model. Also, greater precision may be attained by combining additional characteristics. Aside from that, the quest for strong feature representation, as well as effective classification algorithms for automatic speech  identification. Acoustic models are an important part of speech recognition systems. Speech recognition systems can reliably transcribe spoken words and phrases into text by employing acoustic models. This has a wide range of uses, including medical dictation, automated customer service, and voice-activated search engines. Acoustic models are getting more precise and trustworthy as machine-learning technology continues to progress.

# References

1. De Lima, T.A.; Speech, C. A Survey on Automatic Speech Recognition Systems for Portuguese Language and its Variations. Comput. Speech Lang. 2019, 62, 101055.
2. Sunil, Y.; Prasanna, S.R.M.; Sinha, R. Children’s Speech Recognition under Mismatched Condition: A Review. IETE J. Educ. 2016, 57, 96–108.
3. Taniya; Bhardwaj, V.; Kadyan, V. Deep Neural Network Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 374–378.
4. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. Spectral modification for recognition of children’s speech undermismatched conditions. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa); Linköping University Electronic Press: Linköping, Sweden, 2021; pp. 94–100.
5. Claus, F.; Rosales, H.G.; Petrick, R.; Hain, H. A Survey about ASR for Children. ISCA Archive. 2013, pp. 26–30.
6. Madhavi, M.C.; Patil, H.A. Vocal Tract Length Normalization using a Gaussian mixture model framework for query-by-example spoken term detection. Comput. Speech Lang. 2019, 58, 175–202.
7. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. A formant modification method for improved ASR of children’s speech. Speech Commun. 2021, 136, 98–106.
8. Tsao, Y.; Lai, Y.H. Generalized maximum a posteriori spectral amplitude estimation for speech enhancement. Speech Commun. 2016, 76, 112–126.
9. Bhardwaj, V.; Kukreja, V. Effect of pitch enhancement in Punjabi children’s speech recognition system under disparate acoustic conditions. Appl. Acoust. 2021, 177, 107918.
10. Bhardwaj, V.; Kukreja, V.; Singh, A. Usage of Prosody Modification and Acoustic Adaptation for Robust Automatic Speech Recognition (ASR) System. Rev. d’Intell. Artif. 2021, 35, 235–242.
11. Klejch, O.; Fainberg, J.; Bell, P.; Renals, S. Speaker Adaptive Training Using Model Agnostic Meta-Learning. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 881–888.
12. Bhardwaj, V.; Bala, S.; Kadyan, V.; Kukreja, V. Development of Robust Automatic Speech Recognition System for Children’s using Kaldi Toolkit. In Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020), Coimbatore, India, 15–17 July 2020; pp. 10–13.
13. Bala, S.; Kadyan, V.; Bhardwaj, V. Bottleneck Feature Extraction in Punjabi Adult Speech Recognition System. In Innovations in Computer Science and Engineering; Springer: Singapore, 2021; pp. 493–501.
14. D. Bukhari, Y. Wang, and H. Wang, “Multilingual convolutional, long short-term memory, deep neural networks for low resource speech recognition,” Procedia Computer Science, vol. 107, pp. 842- 847, Apr. 2017.
15. H. Sak, A. Senior, K. Rao, and F. Beaufays, “Fast and accurate recurrent neural network acoustic models for speech recognition,” in Proc. INTERSPEECH, Dresden, Germany, 2015, pp. 1468-1472.
16. Y. Zhao, S. Xu, and B. Xu, “Multidimensional residual learning based on recurrent neural networks for acoustic modeling,” in Proc. INTERSPEECH, San Francisco, USA, 2016, pp. 3419-3423.
17. S. Zhou, Y. Zhao, S. Xu, and B. Xu, “Multilingual recurrent neural networks with residual learning for low-resource speech recognition,” in Proc. INTERSPEECH, Stockholm, Sweden, 2017, pp. 704-708.