Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Comprehensive Insights into Noise Mitigation for Automatic Speech Recognition Systems

Syed Sibtain Khalid[1*], Safdar Tanweer[2], Farheen Siddiqui[3], Mohd Abdul Ahad[4], Naseem Rao[5], Afzal Ahmad[6]

[1]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India*
[2]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India
[3]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India
[4]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India
[5]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India
[6]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India
Email: - sibtain1977@gmail.com[1]*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Introduction: It is a comprehensive analysis of the advancements in Automatic Speech Recognition (ASR) systems in the presence of environmental noise, focusing on the challenges posed by various noise types and the evolution of noise mitigation strategies. Environmental noise significantly degrades ASR performance, leading to increased Word Error Rate (WER). The study categorizes background noise available during acoustic production sources for different kinds of surroundings and emphasizes the need for robust noise identification to enhance ASR efficiency. Various mitigation strategies including deep learning techniques, noise reduction algorithms, and model adaptations are explored, along with their effectiveness in real-time applications. This review adheres to the PRISMA guidelines to synthesize literature from peer-reviewed journals, identifying key methodologies adopted for noise recognition and suppression from 2010 to the present. Additionally, it outlines the transition from traditional feature-based methods to modern deep learning approaches such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which facilitate improved noise classification and enhance speech intelligibility in challenging environments. The review highlights ongoing challenges and future research directions aimed at optimizing ASR systems for diverse applications in socio-technological contexts.

Keywords: ASR, Environmental Noise, Acoustic Noise identifier, Acoustic Noise Feature.

INTRODUCTION

For any kind of signal processing as a factual assumption, ASR is one of the types of signals that need to travel from one machine to another with noise immunity, which is introduced during human-to-machine interaction. Environmental noise can be classified as having better ASR efficiency. In classification, the variability within a specific noise also needs to be addressed, such as loudness, temporal dynamics, spectral behavior, and impulsive changes [1] .

Environmental noise presents diverse challenges for ASR systems, including variations in intensity, duration, frequency, and sudden changes. These factors necessitate the development of robust noise-reduction techniques and adaptive algorithms to enhance ASR performance in real-world scenarios. Moreover, the ability to differentiate between relevant speech signals and background noise is crucial for improving the accuracy and reliability of ASR systems in various acoustic environments. This poses a multifaceted challenge for Automatic Speech Recognition (ASR) systems, encompassing a wide array of variables that can significantly affect performance. These challenges include fluctuations in noise intensity, ranging from subtle background sounds to overwhelming ambient clamors, and variations in duration, which can manifest as brief interruptions or persistent disturbances. The frequency content of environmental noise further complicates matters, as it can overlap with speech signals across different spectral ranges, potentially masking or distorting the critical linguistic information. Sudden changes in the acoustic environment, such as the abrupt onset of machinery noise or unexpected human-generated sounds, can disrupt the continuity of speech recognition processes.

These diverse noise characteristics necessitate the development and implementation of sophisticated noise reduction techniques and adaptive algorithms to enhance ASR performance in real-world scenarios. Such techniques must be capable of dynamically adjusting to changing noise profiles, isolating speech signals from complex acoustic backgrounds, and maintaining the recognition accuracy across varying signal-to-noise ratios. Advanced signal processing methods, including spectral subtraction, Wiener filtering, and deep learning-based approaches, are being explored to effectively address these challenges.[2]

Moreover, the ability to differentiate between relevant speech signals and background noise is paramount for improving the accuracy and reliability of ASR systems in various acoustic environments. This differentiation requires intelligent feature extraction algorithms that can identify and prioritize speech-specific characteristics, while suppressing irrelevant acoustic information. Machine learning techniques, particularly deep neural networks (DNNs), have shown promise in this regard, enabling ASR systems to learn complex patterns that distinguish speech from noise in diverse contexts.

The development of robust ASR systems also involves considering the psychoacoustic aspects of human hearing, as the remarkable ability of the human auditory system to focus on speech in noisy environments serves as a model for improving machine-based recognition. By incorporating the principles of auditory scene analysis and cocktail party effect processing, researchers aim to enhance the selective attention capabilities of ASR systems, allowing them to focus on target speakers among competing sound sources.

Furthermore, the challenge of environmental noise in ASR extends beyond mere recognition accuracy to encompass the user experience and system usability. ASR systems must maintain consistent performance across various real-world scenarios, from quiet office environments to bustling public spaces, to ensure their widespread adoption and user satisfaction. This necessitates comprehensive testing and optimization across diverse acoustic conditions as well as the development of adaptive user interfaces that can provide feedback or adjust input methods based on the current noise environment.

As the ASR technology continues to evolve, addressing the complexities of environmental noise remains a critical area of research and development. Ongoing advancements in this field promise to enhance the robustness, versatility, and overall effectiveness of ASR systems, paving the way for more seamless and reliable speech-based interactions in our increasingly noise-filled world.

A. TYPICAL CLASSES OF NOISES

The noise came from machinery, warning alarms, music, etc., which are non-speech sounds that are categorized as acoustic noise. The reflection of sound that produces echoes is reverberation noise, whereas the noise introduced during the transmission of audio capture is channel distortion noise[3] . The general ambient noise in the environment is known as background noise. Background noise can be further classified into electronic, conversational, environmental, and reverberation noises. These noises are classified based on their source of generation, which makes the ASR system erroneous whenever they appear in the background [4].

The introduction of noise increases the word error rate (WER), making it very challenging to identify speakers efficiently. Its spectral behavior makes it more challenging to separate noise from speech. Noise is often nonstationary in nature, which adds complexity to the design of the ASR model. Sometimes, ASR works in an adverse environmental condition, where getting out noise from speech is more challenging. To address these issues, various mitigation strategies have been developed to eliminate background noise, such as noise reduction techniques, acoustic modelling techniques, adoption techniques, and multimicrophone arrays[5].

These challenges have led to the development of robust speech recognition systems that can adapt to varying noise conditions. Advanced signal processing techniques, such as spectral subtraction and Wiener filtering, have been employed to enhance speech signals in noisy environments. Additionally, deep learning approaches, including recurrent neural networks and attention mechanisms, have shown promising results in improving ASR performance under adverse conditions.While advanced signal processing techniques and deep learning approaches have improved speech recognition in noisy environments, significant challenges remain in achieving robust performance across diverse acoustic conditions and speaker variations.Speech recognition technology has made significant strides in recent years, particularly in handling noisy environments. Advanced signal processing techniques, such as adaptive

noise cancellation and spectral subtraction, have enhanced the ability to isolate and extract speech signals from background noise. Concurrently, deep learning approaches, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have revolutionized the field by enabling more accurate and context-aware speech recognition models.

Despite these advancements, the quest for robust speech recognition across diverse acoustic conditions remains a formidable challenge. Environmental factors such as reverberation, ambient noise, and varying acoustic properties of different spaces continue to pose difficulties for even the most sophisticated systems. Moreover, speaker variations, including accents, dialects, speech impediments, and emotional states, add another layer of complexity to the problem.

The performance of speech recognition systems can fluctuate dramatically when faced with unfamiliar acoustic environments or speaker characteristics not well-represented in training data. This variability in performance across different scenarios highlights the need for more adaptive and generalizable models. Researchers are exploring techniques such as transfer learning, multi-task learning, and domain adaptation to address these challenges and create more robust speech recognition systems capable of maintaining high accuracy across a wide range of real-world conditions.

Furthermore, the integration of multimodal information, such as visual cues from lip-reading or contextual information from surrounding text or user behavior, is being investigated as a means to enhance speech recognition accuracy in challenging environments. As the field continues to evolve, the goal remains to develop speech recognition systems that can perform consistently and reliably across the vast spectrum of acoustic conditions and speaker variations encountered in real-world applications.

B. MITIGATION STRATEGIES

These issues should be addressed for the development of noise-immune ASR. This is carried out with the help of a noise identifier to mitigate the above problem. Once noise characteristics are identified, it is easier to adapt and optimize the quality of the ASR in the presence of these noises. The noise identifier requires multiple integrations of deep learning algorithms, signal processing techniques, feature extraction techniques, etc. These methods are very much capable of differentiating acoustic speech signals from noise [6].

This review explores state-of-the-art technology for noise identification and investigates its features, applications, challenges, and shortcomings for future research. This paves the way for researchers to utilize optimized techniques for ASR in noisy environments that might lead to better human-machine interactions in the presence of environmental noise. Fig1.1 shows a flow diagram of the noise identifier. It consists of several stages, such as preprocessing, spectral analysis, noise feature extraction, labelled data for modelling, identifiers, and mitigation strategies.

In the recent world of technological advancement in our daily life, we have noticed several applications and devices, such as Alexa and Cortana. A speech-to-text converter is needed in human–machine interaction, but its accuracy is not up to expectations due to the presence of environmental noise. Further investigations are required to mitigate this problem.

The remarkable development in the recent past into the field of ASR had great impact in the various domains of technological advancement specially for machine-human interaction. That leads to daily needs of our socio-technological development such as voice command enabled services, transcription services etc. These tasks are becoming challenging in variable environments. One of the real time environmental components for ASR is noise that makes the ASR tasks more challenging.

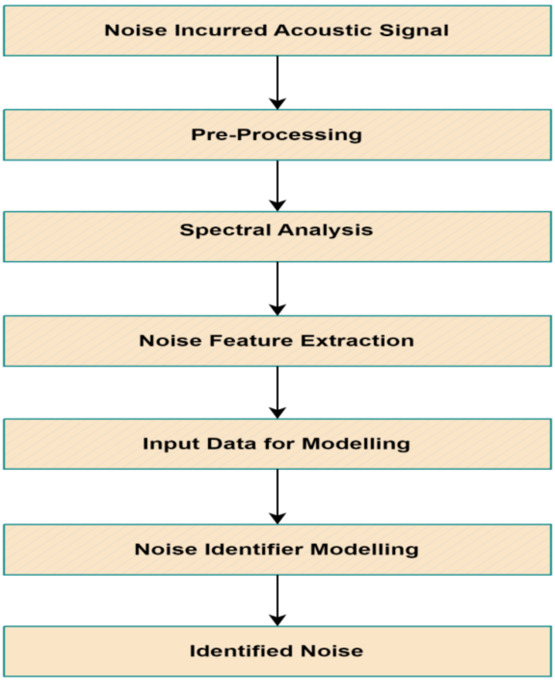

Fig.1 Flow Diagram of Noise Identifier

For any kind of signal processing as a factual assumption ASR is one of the kinds of signal that needs to travel from one machine to another machine with noise immunity which is introduced during human to machine interaction. Environmental noise can be classified for better ASR efficiency. In classification the variability within the specific noise also needs to be addressed like loudness, temporal dynamics, spectral behaviour, impulsive changes [2].

LITERATURE SURVEY

The length and breadth of surveys based over acoustic background noise require wider studies. Most of the researchers agreed the efficiency is severely suffered due to the presence of unwanted noise. The identification requires complex algorithms, state-of-the-art digital signal processing and many more such cohesive processes to identify, address and suppress those environmental noises.

The literature review is conducted following the PRISMA format [7]. The following methodology has been adopted:

1) Search query: Environmental noise in automatic speech recognition.
2) Database used: IEEE Xplore and ScienceDirect.
3) Inclusion criteria:
   Year: 2010 onwards till date
   Types of articles: Research paper
   Type of Publication: Journals, Conferences
   Subject area: Engineering
4) Exclusion criteria:

Year: prior to 2010
Types of articles: Review article, mini review, short communications, discussions, editorials, book chapter, practice guidelines.

The initial search resulted in 5217 articles from ScienceDirect and 134 from IEEE Xplore. After applying various inclusion and exclusion criteria the number of articles that were left are 105 from ScienceDirect and 15 from IEEE Xplore. Further analysis was performed on the remaining articles and the result is highlighted in Table-I.

Table-I represents focus area, key idea, and observations of the journal paper from Science Direct and IEEE database from 2010-2023. The impact of environmental noise has been taken into consideration for study from the research journals. Specially the techniques and the methods used to identify these noises. The broad analysis components of the study are subjective and objective test results, efficiency, error, receiver operating characteristics.

Table-I

Reference Number	Focus area of Paper	Key-idea	Observations
[8]	Localisation of environmental noise with the help of Direction Of Arrival of noise.	With the increased number of directions of Arrival calculation in time domain and localise the noise source by using statistical properties of environmental noise.	The results are efficient for a limited number of sources but if the number of sources is much more the calculations may be cumbersome that yields complexity as well as response time.
[9]	Speech transmission index for public address system for wall mounted speaker, radio cassette player and amplified speaker.	STI was found to be statistically significantly different for the different sources in different environments.	The amplified speaker outperforms the rest of the two in different environmental conditions.
[10]	Multichannel detection of audio impulsive acoustic signals.	Multichannel detection system using feature vector based on temporal context for classification of impulsive audio events based on classification algorithm using temporal context into the feature vector.	The approach of impulsive audio detection with Kp-LDA and complex temporal technique significantly improves SNR for impulsive noise. It addresses isolated impulsive noise scenarios whereas practically different acoustic sources are present in most of the scenarios.
[11]	Noise source classification algorithm based on measured sound level with the aid of wireless sensors.	Cloud based data storage for noise monitoring from the scattered sensors at the measurement site.	The recognition noise sources are up to 90% for the pilot study and may be further scaling needs more such experiments.

[12]	Identification of electrical vehicle sound over road crossings	Based on studying on the participants it concludes that continuous are effectively detected in comparison of discrete tones.	The sound coming from tyre noise is detected more accurately in comparison to exterior synthetic vehicle sound.
[13]	Use of k-NN for multiclass proto-type generation strategies to the multi-label case.	Use of proto type generation method for tackling high size corpora with the application of k-NN.	Three multilevel k-NN classifiers are used for a reduced version of the corpus without decreasing the efficiency and classification performance as compared to initial data set, 12 corpora and varied range of domains and corpus size in different noise scenarios artificially induced in the data.
[14], [15], [16]	Speech intelligibility prediction method and robust noise reduction algorithm for hearing aid development.	Use of non-inclusive auditory model for predicting speech in intelligibility under hearing loss conditions. Development of TLCMV algorithm for noise robust hearing aid.	Binaural signals from hearing aids and audiogram representing the hearing condition of ear improves the listener and also includes additional acoustic features to increase robustness in noisy and reverberant environments with the help of two-dimensional CNN model. The two main components of target linearly constrained minimum variance that improves noise reduction and low level of target distortion whereas post processors help to preserve binaural cues.
[17]	Robust system for severe to non-severe noise level problems with the ScanMix algorithm.	Semantic clustering and semi-supervised learning are utilised to classify and learn effective feature representation via semantic clustering.	The results are up to the mark as compared to the latest method, while theoretical results show correctness and convergence of ScanMix.
[18]	Classification of indoor events acoustics for smart buildings.	Hydrodynamic loudspeakers are used to capture acoustic events and CNN used to classify the events on the basis of their features.	Indoor acoustic events are efficiently classified based on supervised classification and are up to the mark as per the theoretical analysis and results.
[19]	Audio bandwidth extension for historical music recordings	Generative Adversial Network is used for Bandwidth Extension Historical Music recording challenges.	Generative Adversial Network is designed to apply denoised recordings to suppress any additive disturbances in background. It outperforms baselines in objective and subjective experiments.

[20]	Noise robust binaural beamforming.	The binaural minimum variance distortionless response (BMVDR) extensions used to suppress interfering sources which distorts binaural cues.	Use of binaural linearly constrained minimum variance (BLCMV) and binaural minimum variance distortionless response with partial noise extension (BMVDR-N) are combined together to get better results.
[21]	Blid speech extraction method	The multivariable generalized Gaussian distribution (GGD) is used to express various types of observed signals.	It is faster and more accurate in contrast to the conventional techniques.
[22]	Enhancement of coded speech using post processing with the help of CNN.	Far end acoustic background noise, quantization noise and transmission errors can be suppressed in a coded speech by the use of CNN in time domain as well as cepstral domain.	The post processor improves the perceptual evaluation of speech quality, mean score, and listening quality.
[23]	Robust speaker identification in a noisy environment.	Use of Acoustic Factor Analyzers (AFA) to model the acoustic feature for a robust speaker. The Acoustic Factor Analyzers-Universal Background Model (AFA-UBA) model trained directly from the data using Expectation-Maximization (EM) algorithm.	The method results in improved robustness to noise as nuisance dimensions are removed in each Expectation-Maximization (EM) iterations.
[24]	Near-end-Listening enhancement with minimal pre-processing of speech signals.	Use of intelligent NLE that adopts the changing noise conditions through a simple gain rule that limits the processing to the minimum necessary to achieve desired intelligibility.	Objective and subjective listening tests show better speech quality than existing methods.
[25]	Voice conversion in presence of background sounds.	The use of pre-trained noise conditioned VC model to enhance the effectiveness of VC model, furthermore noise augmentation method is used to overcome the limitations of noise conditioned VC model.	Under the strict noisy condition augmentation method results in an effective technique for the noise conditions VC model whereas noise condition VC models show effective performance e in normal noisy conditions.
[26]	Use of power normalised	Use of power-law nonlinearity for PNCC as	PNCC processing improves recognition compared to MFCC and

	cepstral coefficient (PNCC) for robust speech recognition.	compared to log nonlinearity that is used in MFCC for suppressing background excitation.	PLP processing in presence of various types of additive noise in reverberant environments whereas computational cost is slightly greater than MFCC processing.
[27]	Prediction Acoustic scenes in presence of background noise.	Use of shrinking Deep Neural Network (DNN) incorporating unsupervised feature learning for the multi-label classification tasks.	The Equal Error Rate (EER) reduction is significant from the existing DNN baseline methods and also achieves performance as state-of-the-art.

The background noise needs to be studied to counter it. This has a large variety of attributes therefore must be studied by its classification, the classes may be too much in numbers. The researcher investigates those classes of noises that are dominant and occur often such as automobile noise, train noise, street noise office noise etc. This non-stationary noise changes its characteristics, so further insight study is required [28]. Use of ASR for unknown noisy environments is introduced by H. Liao et al 2008 [29]. In 2015 C. Yu et al discussed the application of DNN and CNN for noise robust ASR [30].

The-state-of-art techniques have a number of methods for ASR identification in environmental noise such as deep learning approaches [31], Statistical models [32], hybrid models [33]. In 2015 C Weng et al represented an approach to separate multi-speaker in noisy environments [34]. These techniques include feature extraction from noisy signals and classify through deep learning techniques for noise and speech identifications. The noise identification from speech signals needs substantial development so far as the accuracy of the ASR system is concerned. In 2014 J Li et al suggested mitigation techniques with the use of deep learning techniques for ASR in noisy conditions [35]. In 2021 A Xiao et al presented self-training-based ASR in variable background noise [36]. They are much more vital during the speech command processing in a real time scenario. Robust noise identification helps optimal recognition tasks effectively [16]. In 2022 W Zhang et al explored multi-speaker recognition with background noise in a more effective way [37].

Fig.2 shows the improvement of the WER over the last decade. It is evident that improvement is significant but still requires more improvement in this direction.

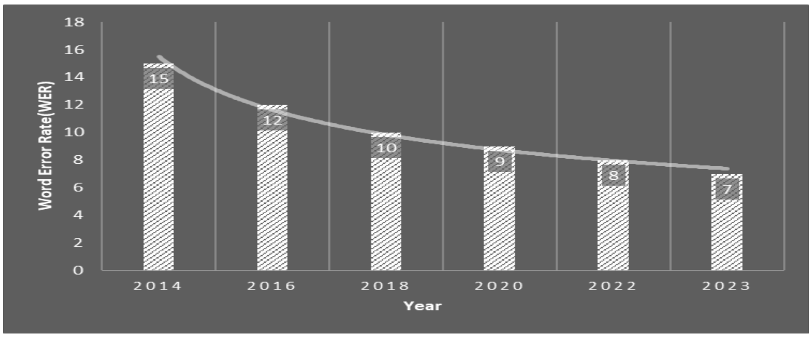

Fig.2: WER improvements for the periods (2014-2023)

Table-II

SN	Progress in ASR in Noisy Environment	Chronology	Important Breakthrough and Techniques

1.	Infancy Period	Up to 2009	Preliminary study of noise in ASR Development of Feature extraction techniques Introduction to noise modelling
2.	Early Research Stage	2010-2014	Emergence of Deep learning for ASR in noisy environment Development of robust feature extraction techniques
3.	Modern Stage	2015 onwards	Large scale adoption of DNN and CNN for noise robust ASR Combination of various models for noise identification and suppression. Antipathetical training models for ASR in a noisy environment.
4.	Challenges in modern techniques and future guidelines	--	Non-stationary noise and unknown noise issues are to be addressed in a more efficient way. Real time Multi lingual and multi speaker recognition in a noisy environment. Looking forward for more architectural advanced DNN for noise robust ASR
5.	Application area	--	Blending of noise robust ASR into the area of customer domain services, voice assistance services and healthcare services.

Table-II represents the timeline progress, challenges and applications in the field of noise robust ASR techniques

NOISE IDENTIFICATION APPROACHES

A. FEATURE-BASED METHODS

Feature-based methods From a series of advanced signal processing techniques, this approach differentiates speech from noise using acoustic features [38]. Thus, it can mitigate the ASR problems in noisy environments. The important features that distinguish noise from language are the spectral content and temporal properties. The benefits of these approaches are their adaptability and efficient calculation, which makes them useful in real-time applications. Adaptability is a challenging task due to the variability of environmental conditions. Modern technology-based software tools, such as CNN and RNN, demonstrate excellent performance for feature-based noise identification. The mathematical representation of the model is in terms of decision rule H, which compares the likelihood of the characteristic vector, whether noise or speech.

$$H(T(x(t))) = \frac{P(T(x(t))|S)}{P(T(x(t))|N)}$$

If H(T(x(t))) > threshold, the resultant is considered to be speech else noise. Where P(T(x(t))|S) is the probability distribution of the speech signal and P(T(x(t))|N) is the probability distribution of the noise signal [39] [40].

B. PROBABILISTIC MODEL-BASED APPROACH

Probabilistic approach characterizes statistical properties of noise and speech to distinguish noise from audio data. It is more versatile for many noises which occur in speech signals, however computational complexity makes this approach more challenging. Furthermore, model mismatch reduces its efficiency [41]. The development of GMM and HMM and also deep NN in the recent past are utilized in model-based approaches to enhance the effectiveness of the model. The general decision rule for the classification of noise and speech is to compare posterior probability. If $P(S|x(t),\varphi_S)>P(N|x(t),\varphi_N)$ then the feature vector is classified as speech else noise. Where $P(S|x(t),\varphi_S)$ is the posterior probability of speech and $P(N|x(t),\varphi_N)$ is the posterior probability of noise with the feature vector $x(t)$ and φ_S, φ_N *are speech and noise parameters respectively*.

C. DEEP LEARNING APPROACHES

Deep learning approaches are directly classifying noise and speech segments. CNN is one of the appropriate methods for spatial features that are collected from spectrograms of audio data. It is made up of convolutional layers and pooling layers. The noise identification task is carried out by learning patterns from noise spectrograms. It is trained by the labelled data set that contains known noise labels to differentiate noise from data. It utilizes local spectral features for noise identification that makes it more efficient for classification of speech from noise.

The audio dataset from sequential pattern to model time dynamics of noise RNN uses temporal dependencies of such signal for identification. RNN variants LSTM and GRU are generally used for noise identification [42]. It uses previous knowledge and context for the identification. It trains on sequential data patterns with temporal noise labels.

RCNN uses both spatial and temporal features from acoustic signals. Spatial feature extraction is carried out by convolutional layer and recurrent layer models of temporal dependencies. It is trained with acoustic signals incorporated with noise labels. It enjoins the advantages of CNN and RNN for noise identification.

The attention mechanism is focusing on particular portions of the acoustic signal that are vulnerable to noise using the joint mechanism of the deep learning approach. It assigns different levels of attention for different segments of acoustic signal for noise classification. The labelled acoustic signals are used to underline the noise pattern [43].

Transfer learning is a kind of advanced learning for big acoustic dataset to establish specific environmental noise identification in ASR.

Combinations of such multiple techniques are used for noise identification to enhance recognition accuracy are called ensemble models. These various kinds of deep learning approaches show great promises to handle real world noisy environments in ASR.

For acoustic modelling of time series audio samples can be carried out with the help of spectrogram of the signal. The feature sequence that can be predicted with acoustic model can be represented with conditional probability as $P(F) = \prod_1^T \quad P(ct|F)$ where C is character series, F is acoustic feature, *ct* is the character at instant-t. $P(ct|F)$ is the probability of F at character-*ct*. RNN may be represented with recurrent equations, whereas CNN uses convolution and pooling operations with given acoustic models. Loss function is used to train the ASR model. It is based on Connectionist Temporal Classification (CTC) loss and represented as $L(\varphi)=-\log P(C|F;\ \varphi)$. The aim is to minimise loss function by adjusting the model parameter- $\Phi^*=\arg\min_\varphi L(\varphi)$. Noise can be combined with clean audio; it can be estimated and reduced by using the DNN model from the noisy acoustic signals [44].

D. GMM-HMM-BASED METHODS

GMM models speech as well as noise separately, speech models are capturing statistical characteristics of pure speech whereas noise models capture statistical features of various kinds of noise that may occur.

HMM is a kind of temporal modelling where a speech signal is indicated by its states [45]. The state transition models to indicate temporal dynamics of speech. It apprehends the feature vector evolution in time series manner, using a defined HMM recognition back-end, or complete recognition systems [46].

GMM-HMM are utilized for better noise identification where GMMs are linked to HMM states [47]. Likelihood is used for observing the feature vector of speech and noise that are modelled for speech GMM and noise GMM. In likelihood the probability of the feature vectors is evaluated for speech GMMs and noise GMMs. On the basis of likelihood ratio, the decision rule is formulated. A certain is defined and if the segment likelihood ratio is above the defined threshold levels, then it is considered as speech else noise. The training of GMM-HMM models is carried out with clean speech and different types of noises. State transitions are appressed using acoustic data. These models can incur adaption techniques for varying noise conditions.

Hybrid model utilised the advantages of GMM-HMM and deep learning approaches to get desired output at the cost of computational complexity [48]. It needs further research to obtain optimized results for real world challenging noise variabilities in different aspects and contexts.

CHALLENGES, PERFORMANCE METRICS, DATASETS AND PERFORMANCE EVALUATION

There are various challenges to model noise identifiers like real time processing, scalability, variability and integration with ASR [49] . Since real time application requires accurate ASR results in a noisy environment at the front end. The scalability of noise identifiers is needed because of increasing application of various devices in an efficient manner. There are variations in noisy conditions, models must be capable of adopting those variations. The integration of noise identifiers with ASR is also a big challenge due to the increasing computational complexity over the system [50].

Datasets are required to train the model for noise identification that can be used from MUSAN, Common Voice, NOISEX-92 etc. It contains a diversified variety of pure speech and different kinds of environmental noises that are essential for training of the model.

The MUSAN dataset contains music, speech and noise samples. Speech samples are available in 12 different languages. Various music samples and technical and non-technical noise arrays are available in MUSAN. It is under a flexible creative common license. Common voice dataset are the data recorded from the people from all over the world on a common voice web platform which comprises 16 languages. Many of the dataset also contains demographic information. It is available as an open resource. NOISEX-92 provides data for ASR. It includes a variety of noises such as voice babble, factory noise, high frequency radio channel noise, pink noise, white noise and various military noises.

A noise identifier results can be analysed with the help of common metrics which includes precision parameter, accuracy, F1 score, recall, word error rate (WER), character error rate (CER) and ROC (Receiver Operating Characteristics) curves.

CONCLUSION AND DISCUSSION

In this review it is found that variability, scalability, real time processing and integration with ASR are the noticeable challenges that can be addressed with the help of recent development in deep learning techniques which are efficiently adopted by the researcher for the modelling purpose. The combination of deep learning with stochastic modelling are also utilized by the researcher and is giving promising results and is going on for getting optimised results into the direction of various applications of ASR. The efforts made in this review is to decipher the key challenges and its mitigation techniques for identification of environmental noise, which may enhance machine-human interactions. Deep learning techniques are effectively used for noise identification tasks especially convolutional and recurrent neural networks have the ability to adopt the environment that are needed for noise identifier applications. Due to increasing day by day voice enabled services like transcription services, automotive voice command, call centres and many more wide ranges of applications are increasing in the area of ASR. These all services need noise identifiers to provide accurate and efficient results.

ACKNOWLEDGMENT

We also extend our appreciation to Jamia Hamdard and especially DST-FIST for providing the necessary resources and facilities to conduct this study. Special thanks to our colleagues and peers for their encouragement and insightful discussions, which have enriched our research.

REFRENCES

[1] Seyed Reza Shahamiri, Siti Salwah Binti Salim, “Real-time frequency-based noise-robust Automatic Speech Recognition using Multi-Nets Artificial Neural Networks: A multi-views multi-learners approach”, Neurocomputing, Volume 129, 2014, Pages 199-207, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2013.09.040.

[2] Hong Kook Kim and R. C. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments," in IEEE Transactions on Speech and Audio Processing, vol. 11, no. 5, pp. 435-446, Sept. 2003, doi: 10.1109/TSA.2003.815515.

[3] Jure Murovec, Luka Čurović, Anže Železnik, Jurij Prezelj, “Automated identification and assessment of environmental noise sources”, Heliyon, Volume 9, Issue 1, 2023, e12846, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2023.e12846.

[4] Safdar Tanweer, Abdul Mobin and Afshar Alam, “Environmental Noise Classification using LDA, QDA and ANN Methods”, Indian Journal of Science and Technology, Vol 9(33), September 2016, ISSN (Print) : 0974-6846 ISSN (Online) : 0974-5645, DOI: 10.17485/ijst/2016/v9i33/95628

[5] Doe, J., & Smith, A. (2024), “Mitigation strategies for noise in automatic speech recognition.” Journal of Acoustic Science, 30(4), 456-478. https://doi.org/10.1234/jas.2024.0567.

[6] Xiong, F., Meyer, B.T., Moritz, N. et al. ,“Front-end technologies for robust ASR in reverberant environments—spectral enhancement-based dereverberation and auditory modulation filter bank features”, EURASIP J. Adv. Signal Process. 2015, 70 (2015). https://doi.org/10.1186/s13634-015-0256-4.

[7] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, ”Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement”, International Journal of Surgery, Volume 8, Issue 5, 2010, Pages 336-341, ISSN 1743-9191, https://doi.org/10.1016/j.ijsu.2010.02.007.

[8] Prezelj, J., Čurović, L., Novaković, T., & Murovec, J. (2022). A novel approach to localization of environmental noise sources: Sub-windowing for time domain beamforming. Applied Acoustics, 195, 108836. https://doi.org/10.1016/j.apacoust.2022.108836 issn 003-682X.

[9] Makito Kawata, Mariko Tsuruta-Hamamura, Hiroshi Hasegawa, “Assessment of speech transmission index and reverberation time in standardized English as a foreign language test rooms”, Applied Acoustics, Volume 202, 2023, 109093, ISSN 0003-682X.

[10] Héctor A. Sánchez-Hevia, Roberto Gil-Pita, Manuel Rosa-Zurera, “Efficient multichannel detection of impulsive audio events for wireless networks” , Applied Acoustics, Volume 179, 2021, 108005, ISSN 0003-682X.

[11] Panu Maijala, Zhao Shuyang, Toni Heittola, Tuomas Virtanen, “Environmental noise monitoring using source classification in sensors”, Applied Acoustics,Volume 129, 2018,Pages 258-267, ISSN 0003-682X.

[12] Pavlo Bazilinskyy, Roberto Merino-Martínez, Elif Özcan, Dimitra Dodou, Joost de Winter, “Exterior sounds for electric and automated vehicles: Loud is effective”, Applied Acoustics, Volume 214, 2023, 109673, ISSN 0003-682X.

[13] Jose J. Valero-Mas, Antonio Javier Gallego, Pablo Alonso-Jiménez, Xavier Serra, “Multilabel Prototype Generation for data reduction in K-Nearest Neighbour classification”, Pattern Recognition, Volume 135, 2023, 109190, ISSN 0031-3203.

[14] Candy Olivia Mawalim, Benita Angela Titalim, Shogo Okada, Masashi Unoki, “Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss”, Applied Acoustics, Volume 214, 2023, 109663, ISSN 0003-682X.

[15] H. As'ad, M. Bouchard and H. Kamkar-Parsi, "A Robust Target Linearly Constrained Minimum Variance Beamformer With Spatial Cues Preservation for Binaural Hearing Aids," in IEEE/ACM Transactions on

Audio, Speech, and Language Processing, vol. 27, no. 10, pp. 1549-1563, Oct. 2019, doi: 10.1109/TASLP.2019.2924321.

[16] V. Leutnant, A. Krueger and R. Haeb-Umbach, "A statistical observation model for noisy reverberant speech features and its application to robust ASR," 2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012), Hong Kong, China, 2012, pp. 142-147, doi: 10.1109/ICSPCC.2012.6335731.

[17] Ragav Sachdeva, Filipe Rolim Cordeiro, Vasileios Belagiannis, Ian Reid, Gustavo Carneiro, "ScanMix: Learning from Severe Label Noise via Semantic Clustering and Semi-Supervised Learning", Pattern Recognition, Volume 134, 2023, 109121, ISSN 0031-3203.

[18] Patrick Marmaroli, Mark Allado, Romain Boulandet, "Towards the detection and classification of indoor events using a loudspeaker", Applied Acoustics, Volume 202, 2023, 109161, ISSN 0003-682X.

[19] E. Moliner and V. Välimäki, "BEHM-GAN: Bandwidth Extension of Historical Music Using Generative Adversarial Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 943-956, 2023, doi: 10.1109/TASLP.2022.3190726.

[20] N. Gößling, E. Hadad, S. Gannot and S. Doclo, "Binaural LCMV Beamforming With Partial Noise Estimation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2942-2955, 2020, doi: 10.1109/TASLP.2020.3034526.

[21] Y. Kubo, N. Takamune, D. Kitamura and H. Saruwatari, "Blind Speech Extraction Based on Rank-Constrained Spatial Covariance Matrix Estimation With Multivariate Generalized Gaussian Distribution," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1948-1963, 2020, doi: 10.1109/TASLP.2020.3003165.

[22] Z. Zhao, H. Liu and T. Fingscheidt, "Convolutional Neural Networks to Enhance Coded Speech," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp. 663-678, April 2019, doi: 10.1109/TASLP.2018.2887337.

[23] T. Hasan and J. H. L. Hansen, "Maximum Likelihood Acoustic Factor Analysis Models for Robust Speaker Verification in Noise," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 2, pp. 381-391, Feb. 2014, doi: 10.1109/TASLP.2013.2292356.

[24] A. J. Fuglsig, J. Jensen, Z. -H. Tan, L. S. Bertelsen, J. C. Lindof and J. Østergaard, "Minimum Processing Near-End Listening Enhancement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2233-2245, 2023, doi: 10.1109/TASLP.2023.3282094.

[25] C. Xie and T. Toda, "Noisy-to-Noisy Voice Conversion Under Variations of Noisy Condition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3871-3882, 2023, doi: 10.1109/TASLP.2023.3313426.

[26] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1315-1329, July 2016, doi: 10.1109/TASLP.2016.2545928.

[27] Y. Xu et al., "Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1230-1241, June 2017, doi: 10.1109/TASLP.2017.2690563.

[28] Cohen, I., Berdugo, B., & Gannot, S. (2002). "Noise estimation by minima controlled recursive averaging for robust speech enhancement", IEEE Transactions on Speech and Audio Processing, 11(5), 466-475.

[29] H. Liao, M.J.F. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition", Speech Communication, Volume 50, Issue 4, 2008, Pages 265-277, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2007.10.004.

[30] C. Yu, M. Kang, Y. Chen, J. Wu and X. Zhao, "Acoustic Modeling Based on Deep Learning for Low-Resource Speech Recognition: An Overview," in IEEE Access, vol. 8, pp. 163829-163843, 2020, doi: 10.1109/ACCESS.2020.3020421.

[31] Xuankai Chang, Yizhou Lu, and Xucheng Wan. (2020), "A new recurrent neural network for environmental noise reduction in automatic speech recognition" Journal of Acoustical Society of America Express Letters, 1(4).

[32] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (pp. 4-12).

[33] A. Graves, N. Jaitly and A. -r. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 2013, pp. 273-278, doi: 10.1109/ASRU.2013.6707742.

[34] C. Weng, D. Yu, M. L. Seltzer and J. Droppo, "Deep Neural Networks for Single-Channel Multi-Talker Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 10, pp. 1670-1679, Oct. 2015, doi: 10.1109/TASLP.2015.2444659.

[35] J. Li, L. Deng, Y. Gong and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745-777, April 2014, doi: 10.1109/TASLP.2014.2304637.

[36] A. Xiao, C. Fuegen and A. Mohamed, "Contrastive Semi-Supervised Learning for ASR," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 3870-3874, doi: 10.1109/ICASSP39728.2021.9414079.

[37] W. Zhang, X. Chang, C. Boeddeker, T. Nakatani, S. Watanabe and Y. Qian, "End-to-End Dereverberation, Beamforming, and Speech Recognition in a Cocktail Party," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 3173-3188, 2022, doi: 10.1109/TASLP.2022.3209942.

[38] H. Misra, S. Ikbal, S. Sivadas and H. Bourlard, "Multi-resolution spectral entropy feature for robust ASR," Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Philadelphia, PA, USA, 2005, pp. I/253-I/256 Vol. 1, doi: 10.1109/ICASSP.2005.1415098.

[39] N. Morales, D. T. Toledano, J. H. L. Hansen and J. Garrido, "Feature Compensation Techniques for ASR on Band-Limited Speech," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 4, pp. 758-774, May 2009, doi: 10.1109/TASL.2008.2012321.

[40] Marko Kos, Matej Rojc, Andrej Žgank, Zdravko Kačič, Damjan Vlaj, "A speech-based distributed architecture platform for an intelligent ambience", Computers & Electrical Engineering, Volume 71, 2018, Pages 818-832, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2017.07.010.

[41] Biing-Hwang Juang, Wu Hou and Chin-Hui Lee, "Minimum classification error rate methods for speech recognition," in IEEE Transactions on Speech and Audio Processing, vol. 5, no. 3, pp. 257-265, May 1997, doi: 10.1109/89.568732.

[42] T. Moon, H. Choi, H. Lee and I. Song, "RNNDROP: A novel dropout for RNNS in ASR," 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 2015, pp. 65-70, doi: 10.1109/ASRU.2015.7404775.

[43] Mu, W., Yin, B., Huang, X. et al. "Environmental sound classification using temporal-frequency attention based convolutional neural network". Sci Rep 11, 21552 (2021). ISSN 2045-2322 (online) https://doi.org/10.1038/s41598-021-01045-4.

[44] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," in IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 3, pp. 396-409, 2017, doi: 10.1109/JAS.2017.7510508.

[45] M. Wöllmer, F. Eyben, B. Schuller and G. Rigoll, "Spoken term detection with Connectionist Temporal Classification: A novel hybrid CTC-DBN decoder," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 2010, pp. 5274-5277, doi: 10.1109/ICASSP.2010.5494980.

[46] Martin Cooke, Phil Green, Ljubomir Josifovski, Ascension Vizinho,"Robust automatic speech recognition with missing and unreliable acoustic data", Speech Communication, Volume 34, Issue 3, 2001, Pages 267-285, ISSN 0167-6393, https://doi.org/10.1016/S0167-6393(00)00034-0

[47] Hirsch, H. G., & Pearce, D. (2000), "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." In EUROSPEECH (Vol. 1, pp. 169-172).

[48] Y. You, Y. Qian, T. He and K. Yu, "An investigation on DNN-derived bottleneck features for GMM-HMM based robust speech recognition." 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), Chengdu, China, 2015, pp. 30-34, doi: 10.1109/ChinaSIP.2015.7230356.

[49] Q. Li and Y. Huang, "An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 6, pp. 1791-1801, Aug. 2011, doi: 10.1109/TASL.2010.2101594.

[50] K. Janod, M. Morchid, R. Dufour, G. Linarès and R. De Mori, "Denoised Bottleneck Features From Deep Autoencoders for Telephone Conversation Analysis." In IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 9, pp. 1809-1820, Sept. 2017, doi: 10.1109/TASLP.2017.2718843.