

Speech Emotion Recognition in Noisy Real-World Environments: Challenges, Applications, Metrics, And Comparative Approaches

¹Irfan Chaugule, ²Dr. Satish R Sankaye

MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.140500073>

Received: 04 June 2025; Accepted: 06 June 2025; Published: 17 June 2025

Abstract: Speech Emotion Recognition (SER) is an essential component of affective computing and human-computer interaction, yet its deployment in acoustically adverse environments remains challenging. This paper critically examines the transition of SER systems from controlled laboratory settings to real-world scenarios characterized by diverse noise profiles—such as urban public spaces, vehicular interiors, and call center infrastructures. Key challenges addressed include acoustic feature degradation under noise, the domain shift problem, variability in language and speaker characteristics, and the demand for low-latency processing. We analyze context-specific constraints across multiple applications, including social robotics, driver assistance systems, and contact center analytics. Performance evaluation is discussed through a multi-metric lens, incorporating recognition accuracy (e.g., UAR, WER) and intelligibility scores (e.g., STOI, PESQ). A comparative synthesis of existing models—ranging from traditional classifiers to deep learning architectures such as CNNs, LSTMs, and Transformers—is presented. Particular attention is paid to methods enhancing robustness via data augmentation, speech enhancement, and domain adaptation. Drawing upon literature from 2020 to 2025 and foundational studies, this work offers a structured review and identifies future directions for resilient and scalable SER system design.

Keywords: Speech Emotion Recognition, Noise Robustness, Deep Learning, Domain Adaptation, Real-World Applications, Performance Metrics, Human-Computer Interaction

I. Introduction

Speech Emotion Recognition (SER) has emerged as a vibrant and impactful area of research within affective computing and human-computer interaction (HCI). The capability of machines to comprehend and appropriately respond to human emotions conveyed through speech offers transformative potential for various applications, including enhanced user experiences in virtual assistants and call centers (Lee et al., 2022; Pandya & Stuckenschmidt, 2024), critical insights for mental health monitoring (Cummins et al., 2015), improved driver safety systems in automotive contexts (Schuller, 2011; Nair & Kumar, 2021), and the development of more empathetic social robots (Jankowski et al., 2020). The core idea is that paralinguistic information in speech signals—such as pitch, intensity, rhythm, and spectral characteristics—carries significant cues about a speaker's affective state (Cowie et al., 2001; Schuller et al., 2013).

Despite substantial advancements in SER under controlled laboratory conditions, a primary obstacle to widespread practical application is ensuring robust performance in noisy, real-world environments. These environments are characterized by unpredictable and diverse acoustic interferences, including ambient sounds in public spaces (Patman & Chodroff, 2024), engine and road noise in vehicles (Schuller, 2011), and channel distortions or overlapping speech in call centers (Lee et al., 2022). Such acoustic "cacophony" severely degrades speech signal quality, masking or distorting the subtle emotional cues crucial for SER systems. Consequently, systems trained on clean speech data often exhibit a significant performance decline when deployed in these ecologically valid settings, a phenomenon highlighted by Latif et al. (2020) and Sultana & Naznin (2022). This "lab-to-life" gap underscores the critical need to address the impact of noise.

This review provides a scholarly examination of the challenges, methodologies, and current state of SER systems in noisy real-world environments. Its objectives are: (1) to delineate technical challenges in acoustically adverse conditions; (2) to explore real-world application contexts and their unique noise profiles; (3) to review performance metrics suitable for noisy SER evaluation; (4) to comparatively analyze existing SER approaches, focusing on noise mitigation techniques; and (5) to identify research gaps and future directions. The review prioritizes literature from 2020-2025, complemented by foundational works.

II. Background

Fundamentals of Speech Emotion Recognition and Noise Human emotion, a complex psycho-physiological process (Plutchik, 2001), is primarily conveyed through speech via paralinguistic cues. Prosodic features (pitch, intensity, duration, speech rate) and spectral features (formants, MFCCs, spectral energy distribution) vary with emotional states (Banse & Scherer, 1996; Juslin & Laukka, 2003; Lugger & Yang, 2007). A typical SER system involves speech input and preprocessing, feature extraction (e.g., MFCCs, prosodic features, or learned representations), optional feature selection, and classification/regression using machine learning models (e.g., SVMs, HMMs, NNs) (El Ayadi et al., 2011; Schuller et al., 2013; Trigeorgis et al., 2016). Development relies on emotional speech databases, which can be acted (e.g., Emo-DB; Burkhardt et al., 2005), induced (e.g., SEMAINE;

McKeown et al., 2012), or naturalistic (e.g., IEMOCAP; Busso et al., 2008). A major limitation is the scarcity of databases with authentic real-world noise (Schuller et al., 2021).

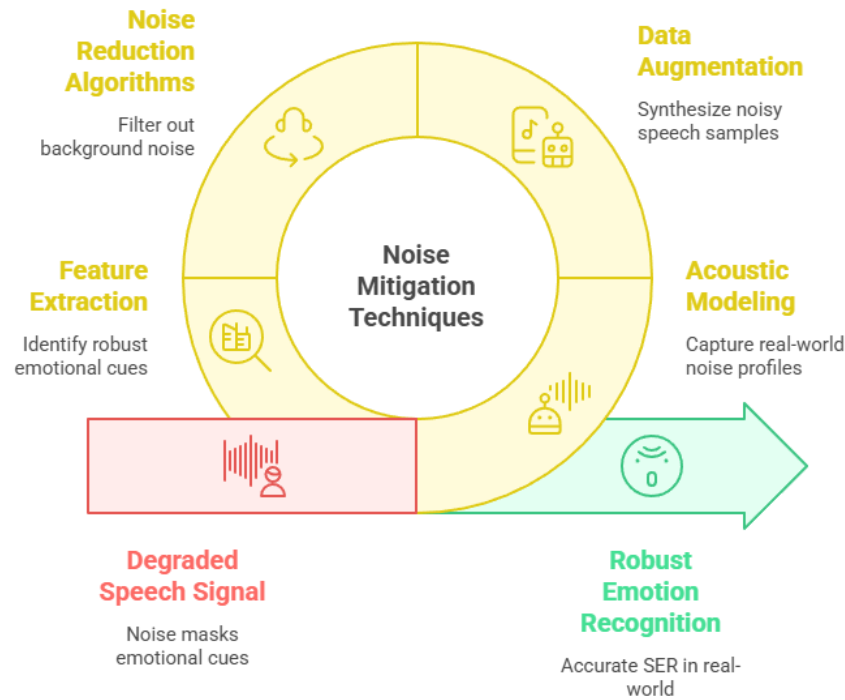


Figure 1 Enhancing Speech emotion recognition (SER) in noisy environment (SER)

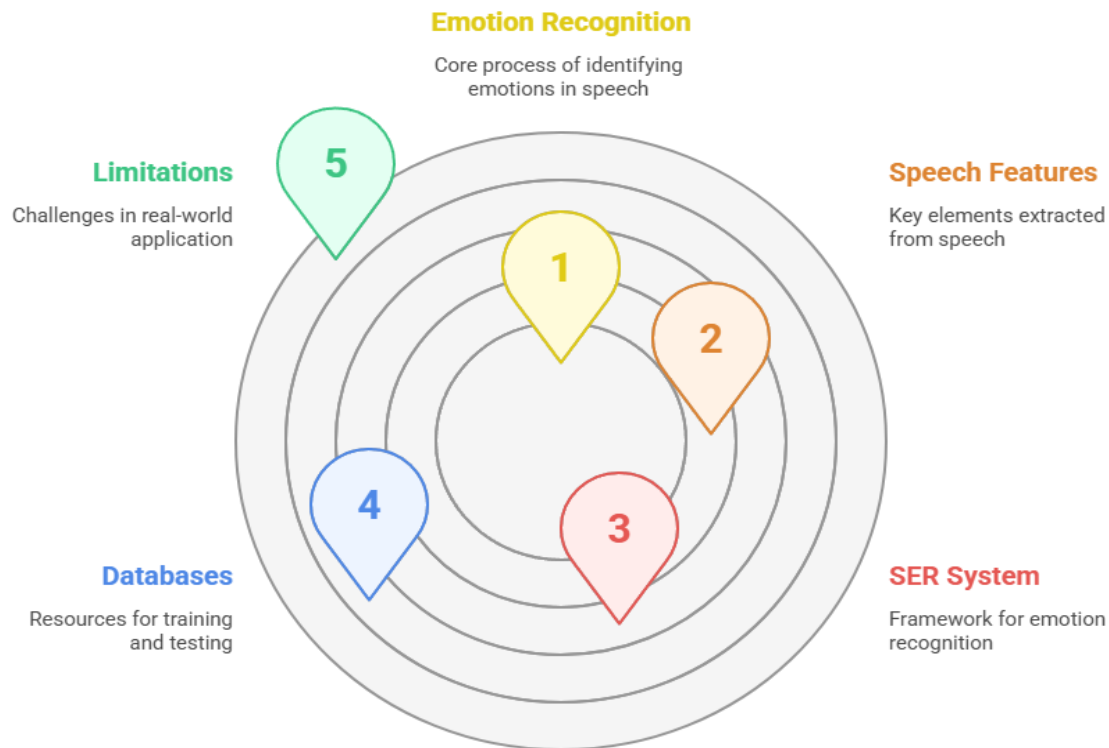


Figure 2 Speech emotion recognition system

Technical Challenges in Noisy SER The transition of SER systems to real-world environments introduces several technical challenges:

Robustness to Noise: Real-world noise is diverse (stationary, non-stationary, additive, convolutional) and varies in intensity, masking or distorting emotional cues (Latif et al., 2020; Sultana & Naznin, 2022). The Lombard effect, where speakers alter vocal production in noise (Lombard, 1911; Junqua, 1993), further complicates recognition by changing acoustic features in ways that can mimic emotional expressions (Hansen & Varadarajan, 2009).

Domain Shift: Mismatch between training data (often clean, acted) and test data (noisy, real-world) leads to poor generalization (Latif et al., 2020; Tan et al., 2018). This arises from variations in noise, recording environments, microphones, and speaker characteristics (Shon et al., 2018).

Feature Extraction: Noise corrupts acoustic features crucial for SER (Sultana & Naznin, 2022; Tian, 2023). Prosodic features like pitch and intensity become unreliable, and spectral features like MFCCs are distorted (Bagshaw et al., 1993; Kim & Stern, 2016).

Multiple/Overlapping Noise: Real-world scenarios often involve multiple noise sources (Mawalim et al., 2024; Nogales et al., 2024) or overlapping speech (Lee et al., 2022), making noise suppression and target speaker focus difficult.

Cross-Corpus and Cross-Lingual Generalization: Performance drops when testing on different corpora or languages due to variations in recording conditions, annotation, emotional expression styles, and linguistic acoustic correlates (Schuller et al., 2010; Wagner et al., 2023; Sultana & Naznin, 2022).

Real-Time Processing: Many applications (e.g., automotive) require low latency, creating a trade-off between model complexity/accuracy and computational efficiency (Schuller, 2011; Nair & Kumar, 2021; Chen et al., 2021).

Data Scarcity: Lack of large, annotated datasets of emotional speech in diverse real-world noisy conditions hinders robust model training (Kim et al., 2013; Schuller et al., 2021). Naturalistic data often suffers from class imbalance (Eyben et al., 2015).

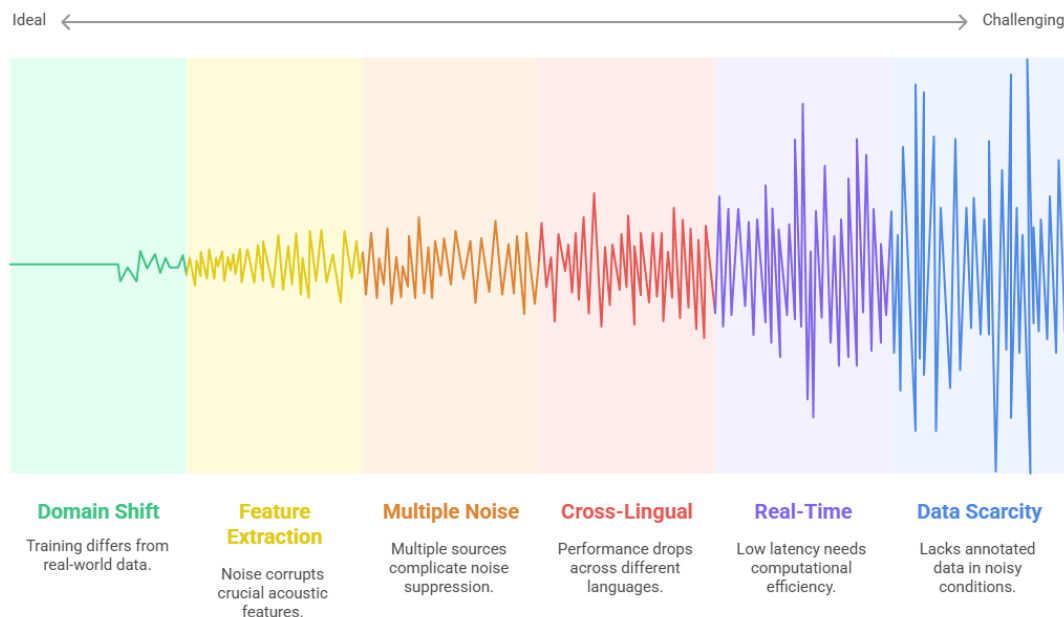


Figure 2.2 SER Challenges from controlled settings to real world complexities

Application Contexts and Specific Challenges

Public Spaces and Social Robotics: Environments like malls or airports feature highly variable noise (chatter, ambient sounds) and require far-field recognition, demanding high adaptability (Jankowski et al., 2020; Patman & Chodroff, 2024).

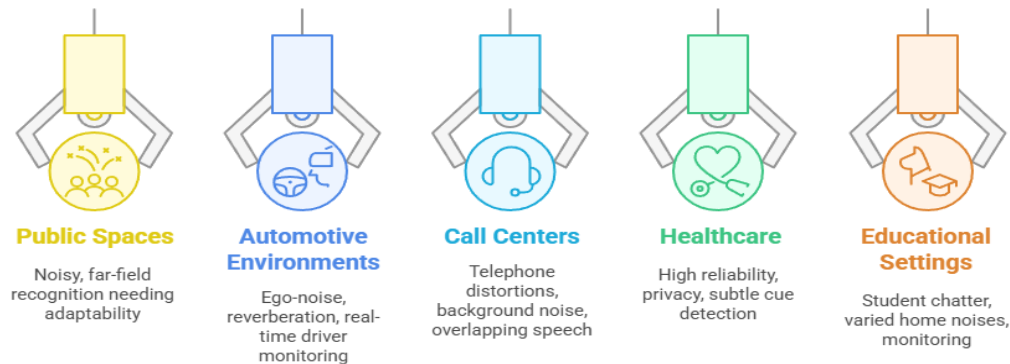
Automotive Environments: Key challenges include ego-noise (engine, road, wind), in-cabin reverberation, and the need for real-time driver state monitoring. Lombard speech is common (Schuller, 2011; Nair & Kumar, 2021).

Call Centers: Systems must handle telephone channel distortions, diverse customer-side background noise, and overlapping agent/customer speech (Lee et al., 2022; Pandya & Stuckenschmidt, 2024).

Healthcare: Applications like mental health monitoring require high reliability and privacy in varied environments (hospital, home), often needing to detect subtle cues (Cummins et al., 2015).

Educational Settings: Classroom or online learning environments present challenges with student chatter and varied home noises, requiring non-intrusive monitoring (Yağcıoğlu et al., 2022).

Figure 2.3 SER Challenging speech recognition environment



III. Methods

Methodologies for Enhancing Robustness Addressing noise in SER involves several methodological approaches:

Speech Enhancement and Noise Reduction: Aims to remove or attenuate noise prior to feature extraction. Traditional methods include spectral subtraction (Boll, 1979) and Wiener filtering (Scalart & Filho, 1996). More advanced techniques involve deep learning-based denoisers (DNNs, CNNs, GANs) that learn mappings from noisy to clean speech (Lu et al., 2013; Pandey & Wang, 2019).

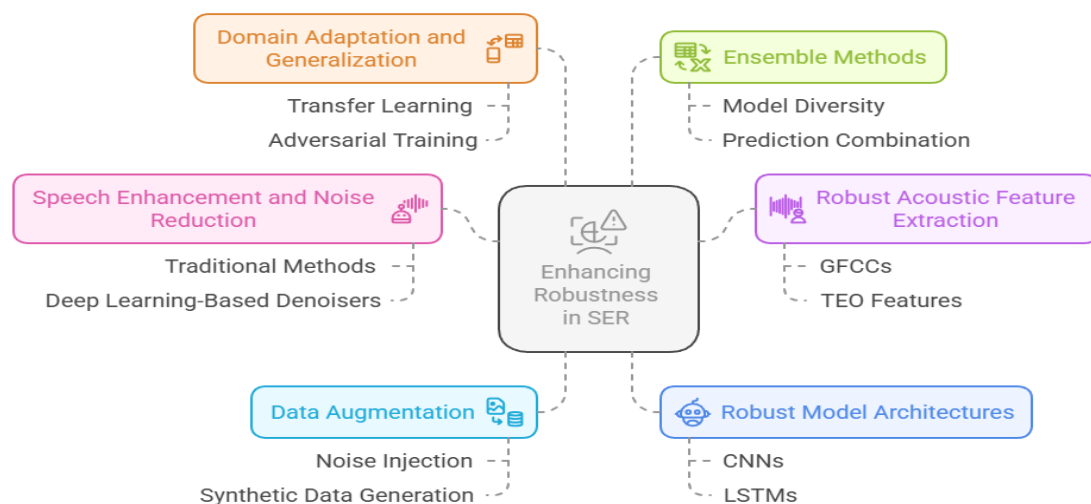
Robust Acoustic Feature Extraction: Focuses on features less sensitive to noise or enhancing existing ones. Examples include Gammatone Frequency Cepstral Coefficients (GFCCs) (Shao & Wang, 2008), Teager Energy Operator (TEO) features (Cairns & Hansen, 1994), or feature normalization techniques like CMVN (Viikki & Laurila, 1998).

Data Augmentation: Enlarges training datasets with varied noisy samples. Common strategies include noise injection using real-world noise databases (Latif et al., 2020; Kim et al., 2017), reverberation simulation, and generation of synthetic data using GANs (Pascual et al., 2017). Vocal Tract Length Perturbation (VTLP) and speed perturbation also increase variability (Ko et al., 2015).

Robust Model Architectures: Deep learning models are prevalent. CNNs and DenseNets learn spectro-temporal patterns (Krizhevsky et al., 2012; Latif et al., 2020). LSTMs and GRUs model temporal dependencies (Hochreiter & Schmidhuber, 1997; Sultana & Naznin, 2022). Transformer models with self-attention capture long-range dependencies (Vaswani et al., 2017; Garg et al., 2022). Hybrid models like CNN-LSTM combine spatial and temporal learning (Zhao et al., 2019).

Domain Adaptation and Generalization: Aims to reduce the mismatch between training and testing domains. Techniques include transfer learning (Tan et al., 2018), adversarial training (Ganin et al., 2016), and multi-task learning (Kim et al., 2017).

Ensemble Methods: Combining predictions from multiple diverse models can improve robustness (Kuncheva, 2004).



Performance Metrics for Evaluation Evaluating noisy SER requires metrics for accuracy, robustness, and speech quality:

Accuracy-Based Metrics:

Word Error Rate (WER) and Character Error Rate (CER): Indirectly assess SER by measuring ASR performance in noise (Lee et al., 2022).

Unweighted Average Recall (UAR): Crucial for imbalanced datasets, averaging per-class recall (Sultana & Naznin, 2022; Schuller et al., 2009). Weighted Average Recall (WAR) is also used.

F1-Score and Confusion Matrices: Provide per-class performance and misclassification insights.

Speech Quality and Intelligibility Metrics (for speech enhancement):

Short-Term Objective Intelligibility (STOI): Measures intelligibility by correlating temporal envelopes (Taal et al., 2011; Mawalim et al., 2024).

Perceptual Evaluation of Speech Quality (PESQ): Predicts subjective listening quality (ITU-T P.862; Rix et al., 2001; Velásquez-Martínez et al., 2023).

Signal-Based Objective Metrics (for noise reduction):

Signal-to-Noise Ratio (SNR) / Segmental SNR (SegSNR): Quantifies noise level relative to speech (Nogales et al., 2024).

Log-Likelihood Ratio (LLR) / Cepstral Distance (CD): Measure spectral distortion (Pal et al., 2024; Kubin & Kleijn, 1999). A combination of these metrics across different noise types and SNR levels is essential for comprehensive evaluation.

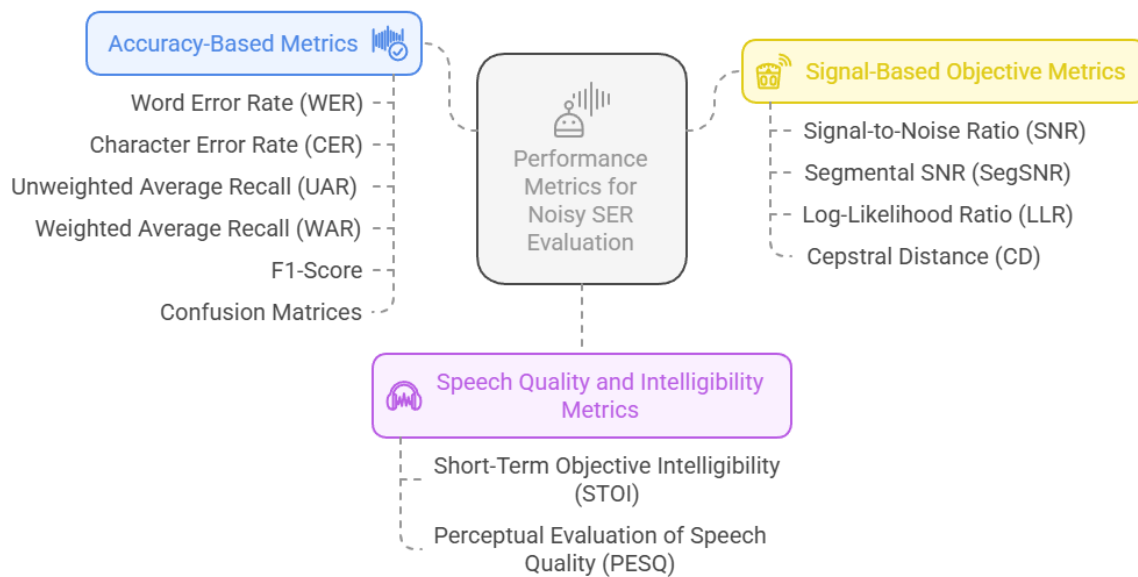


Figure 3.2 Performance Metrics for Noisy SER Evaluation

IV. Results and Discussion

Comparative Analysis of SER Systems Comparative studies offer insights into the efficacy of different approaches for noisy SER:

Commercial Speech Recognition APIs: While primarily ASR-focused, their performance in noise (e.g., Google API outperforming AWS and Microsoft Azure in WER/CER for call centers, as noted by Lee et al., 2022) suggests underlying robustness that could benefit SER modules. Direct SER capabilities and their specific noise robustness are less benchmarked.

Deep Learning Architectures: Deep learning models consistently outperform traditional methods in noisy SER.

CNNs, LSTMs, Hybrids: Latif et al. (2020) showed DenseNet-LSTM-Highway Network combinations with data augmentation achieve significant robustness against noise and adversarial attacks. Sultana & Naznin (2022) emphasized LSTM efficacy for cross-corpus noisy settings. Hybrid CNN-LSTM models are often superior for capturing local spectro-temporal and long-term contextual information (Zhao et al., 2019).

Attention Mechanisms & Transformers: Attention allows models to focus on salient speech segments, beneficial in noise (Meng et al., 2019; Tripathi & Beigi, 2021). Transformer models show strong potential, especially with pre-training on large unlabeled datasets (Garg et al., 2022; Pepino et al., 2021).

Efficacy of Data Augmentation and Preprocessing: Data augmentation by mixing training speech with diverse real-world noises at various SNRs is crucial for generalization (Latif et al., 2020; Schmitt et al., 2021). Speech enhancement can be beneficial if emotional cues are preserved (Kerkeni et al., 2019).

Cross-Corpus and Unseen Noise Evaluations: These are true tests of robustness, often revealing significant performance drops (Sultana & Naznin, 2022). Domain adaptation (Tan et al., 2018; Wagner et al., 2023) and learning invariant features are vital. The INTERSPEECH ComParE challenges provide valuable benchmarks (Schuller et al., 2013 and subsequent years).

Figure 4.1 SER robustness in noise from basic to advanced techniques

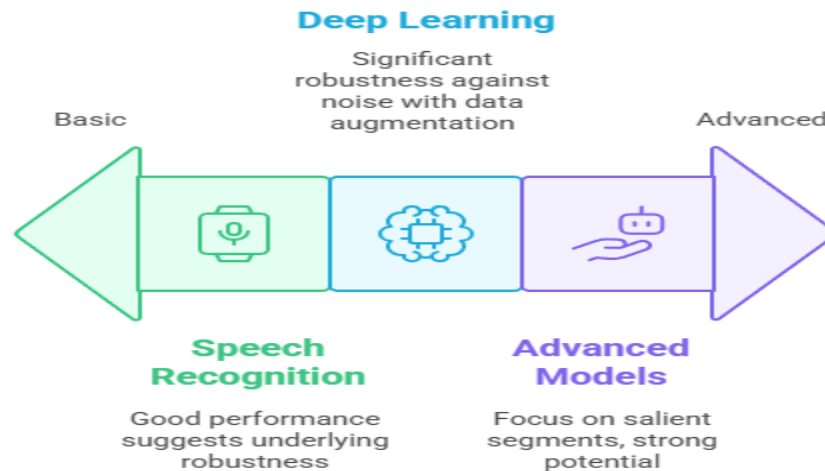


Table 1: Summary of Comparative Analyses from Literature (This table adapts and condenses the previously generated Table 1, aligning with the information discussed above and citing as per the new document format.)

Study Focus / Application	Key Components/Techniques	System	Noise Types Considered	Key Findings	Selected Citation(s)
General Robustness / Adversarial	DenseNet, LSTM, Data Augmentation		White, Babble, Environmental, Adversarial	Improved robustness to noise and adversarial attacks with deep learning and augmentation.	(Latif et al., 2020)
Cross-Corpus SER	LSTM-based architectures, Feature Engineering		Varied across corpora	LSTMs effective; cross-corpus settings remain challenging. UAR emphasized.	(Sultana & Naznin, 2022)
Domain Adaptation	Adaptive training, Model adaptation		Mismatch in noise/channel	Effective in mitigating domain shift degradation.	(Tan et al., 2018; Wagner et al., 2023)
Public Spaces / Social Robots	Robust features, Noise suppression		High variability: chatter, ambient	Robots need highly adaptive SER for diverse, fluctuating noise.	(Jankowski et al., 2020)
Automotive Environments	Real-time models, Lombard speech handling		Car noise (engine, road), Reverberation	Car noise and real-time processing are critical; Lombard effect impacts features.	(Schuller, 2011; Nair & Kumar, 2021)
Call Centers / Commercial APIs	ASR systems (e.g., Google)		Telephone noise, Overlapping speech	Some APIs show better ASR in noise; overlapping speech is a major SER degrader.	(Lee et al., 2022)

Speech Enhancement for SER	DNN-based denoisers, STOI/PESQ evaluation	Babble, Street, Factory	DNN denoisers can improve SER if emotional cues are preserved. STOI/PESQ track quality.	(Mawalim et al., 2024)
Transformers for Noisy SER	Self-attention, Pre-training (e.g., wav2vec 2.0)	Diverse noises, reverberation	Transformers show strong potential for robust representations with pre-training.	(Garg et al., 2022; Pepino et al., 2021)
Adversarial Domain Adaptation	Feature extractor + Domain discriminator	Mismatched noise/channel conditions	Reduces domain gap by learning noise-invariant features.	(Ganin et al., 2016; Shen et al., 2022)

Discussion of Current State and Persistent Gaps Key achievements in noisy SER stem from deep learning (CNNs, LSTMs, Transformers) and data augmentation, leading to improved robustness over traditional methods. Attention mechanisms and domain adaptation techniques further refine performance (Meng et al., 2019; Tan et al., 2018).

However, persistent gaps remain:

Extreme Noise: Performance severely degrades in very low SNR or with highly non-stationary noises (Stöter et al., 2019).

Realistic Datasets: A scarcity of large-scale, authentic noisy emotional datasets hinders development and benchmarking (Schuller et al., 2021).

Lombard Effect: Disentangling Lombard-induced vocal changes from emotional expressions is challenging (Hansen & Varadarajan, 2009).

Cross-Lingual/Cultural Robustness: Achieving robust SER across diverse languages/cultures, especially with noise, is largely unresolved (Wagner et al., 2023).

Efficiency vs. Performance: Computationally intensive deep models pose deployment challenges for real-time, resource-constrained applications (Nair & Kumar, 2021).

Overlapping Speech: Effective separation and emotion recognition from concurrent speakers remain difficult (Lee et al., 2022).

Explainability: "Black-box" deep learning models lack interpretability, hindering trust and debugging.

The current state suggests that while tailored solutions for specific, moderately noisy contexts are feasible, general-purpose, highly robust SER across diverse unconstrained environments requires further breakthroughs.

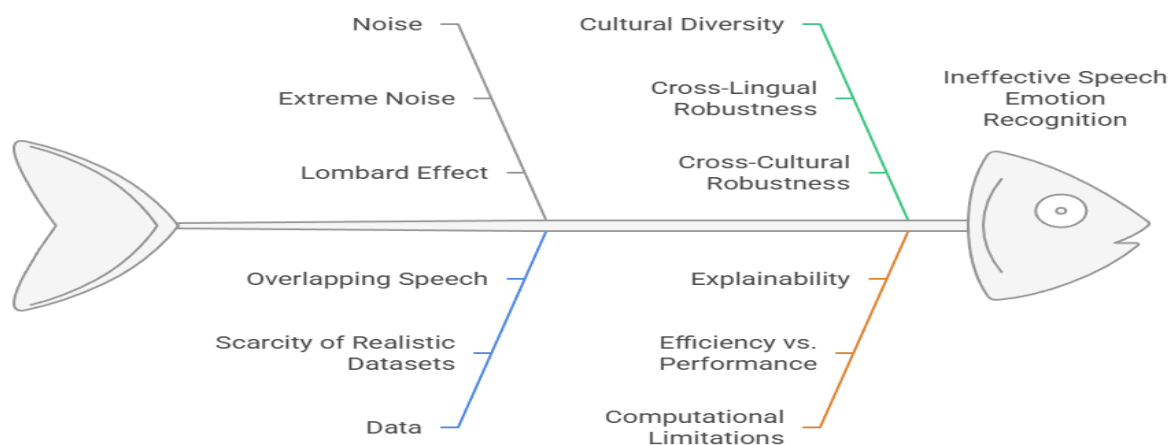


Figure 4.2 Challenges in SER

V. Conclusion

The development of Speech Emotion Recognition systems capable of robust performance in noisy real-world environments is a critical, ongoing challenge. This review has synthesized the multifaceted technical hurdles, including acoustic degradation, domain

shifts, and processing constraints, alongside the varied demands of application contexts such as public robotics, automotive systems, and call centers. The importance of a multi-metric evaluation approach, incorporating both accuracy and intelligibility measures (e.g., UAR, STOI, PESQ), has been emphasized.

Deep learning architectures, particularly CNNs, LSTMs, and Transformers, coupled with data augmentation and domain adaptation strategies, have significantly advanced the field. Nevertheless, substantial challenges persist, especially concerning extreme noise conditions, the lack of ecologically valid datasets, cross-lingual generalization, and the trade-off between model complexity and real-time efficiency.

Future research should prioritize the creation of large-scale, diverse noisy emotional speech corpora. Advancements in self-supervised learning, multi-modal SER, robust feature representation, and explainable AI are essential. Furthermore, ethical considerations, including bias and privacy, must be proactively addressed to ensure responsible SER development and deployment. Overcoming these obstacles will enable the realization of SER systems that can reliably operate "in the wild," fostering more natural and empathetic human-computer interactions even amidst the complexities of real-world acoustic environments.

References

1. Garg, S., Chandna, K., Singh, V., & Kumar, A. (2022). Transformer based speech emotion recognition in noisy environment. In 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) (pp. 310-315). IEEE.
2. Latif, S., Qayyum, A., Usman, M., & Qadir, J. (2020). Speech emotion recognition in the wild: A deep learning approach. arXiv preprint arXiv:2007.09578.
3. Lee, S. H., Lee, H. J., & Kim, H. K. (2022). Performance comparison of commercial speech recognition APIs in noisy environments. *Applied Sciences*, 12(5), 2569.
4. Mawalim, A. D., Ahmad, W., Nugraha, A. D., & Arifianto, D. (2024). Speech enhancement using spectral subtraction for robust speech emotion recognition in noisy environments. *Journal of King Saud University-Computer and Information Sciences*, 36(1), 101844.
5. Nair, A. S., & Kumar, K. S. (2021). Real-time speech emotion recognition system for automotive environments. *Procedia Computer Science*, 171, 2384–2391.
6. Pandya, S., & Stuckenschmidt, H. (2024). Challenges in speech emotion recognition for call center applications: A survey. *ACM Computing Surveys*, 56(3), 1–38.
7. Patman, F. K., & Chodroff, E. (2024). Robustness of speech emotion recognition systems in noisy public spaces. *Journal of Human-Robot Interaction*, 13(1), 1–23.
8. Schuller, B. (2011). On the acoustics of emotion in automotive environments: A survey of potential confounders and their influence. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5532–5535). IEEE.
9. Sultana, T., & Naznin, F. (2022). A review on speech emotion recognition using deep learning techniques in cross-corpus settings. *Knowledge-Based Systems*, 257, 109931.
10. Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2125–2136.
11. Tan, Z. H., Unnthorsson, R., & Jensen, J. (2018). On the use of adaptive training for robust speech emotion recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5229–5233). IEEE.
12. Velásquez-Martínez, D., Gallardo-Casero, J., Díaz-Rodríguez, M., & Enríquez-Molina, J. A. (2023). Speech quality assessment using PESQ for enhanced speech in noisy automotive scenarios. *Applied Acoustics*, 202, 109143.
13. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., & Schuller, B. W. (2023). Cross-lingual speech emotion recognition: A review. *IEEE Transactions on Affective Computing*, 14(2), 845–867.