A Comprehensive Survey on Audio Enhancement Systems

Sufiyan J T
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
jtfinu@gmail.com

Joel Mathew Thomas
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
joelthomaspala@gmail.com

Karthikeyan K S
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
KarthikeyanKS@gmail.com

Junaid M P
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
Juanidjnd@gmail.com

Mrs.Linda Sebastian
Dept. of CSE
College of Engineering Kidangoor
Kottayam, Kerala, India
lindasebastian@ce-kgr.org

***Abstract*—The increasing need for high-quality audio processing in a variety of fields calls for creative answers to problems including real-time adaptation, speech clarity, and noise reduction. Modern noise reduction methods are examined in this survey, with a focus on the function of deep learning in audio improvement systems. Important developments are addressed, such as self-supervised learning techniques, multi-stage neural networks, and real-time audio-visual speech augmentation. The study examines methods such as deep neural filters, adaptive filtering, and simultaneous denoising and dereverberation to show gains in processing efficiency, intelligibility, and signal-to-noise ratios. Furthermore, the potential of integrating audio-visual fusion and adaptive batch processing frameworks to transform noise reduction applications in a variety of settings, from assistive hearing equipment to telecommunications, is examined. This survey aims to provide a comprehensive overview of current methodologies, guiding future research in developing robust, efficient, and accessible audio processing systems.**

***Index Terms*—Audio enhancement, Noise reduction,Machine learning, Adaptive processing, Real-time audio, Deep learning, Batch processing, Audio optimization, Signal processing**

In today's digital world, high-quality audio is pivotal across diverse fields such as education, entertainment, telecommunications, and healthcare. However, real-world audio is often plagued by noise, reverberation, and distortions that degrade its quality and intelligibility. Existing audio enhancement tools, such as Adobe Audition and iZotope RX, provide advanced noise reduction capabilities but remain expensive, require significant expertise, and lack scalability for large datasets. Simpler alternatives like Audacity fail to leverage state-of-the-art advancements in machine learning, limiting their efficacy for precision audio enhancement tasks.

With the advent of deep learning, significant strides have been made in noise removal and audio processing. Deep learning models, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers, have proven effective in isolating and enhancing audio signals even in challenging acoustic environments. Transfer learning, a technique initially popularized in computer vision, has further revolutionized audio processing by enabling the reuse of pre-trained models for new tasks, reducing the need for large datasets and computational resources while improving efficiency and performance. [6]

By synthesizing the latest advancements in noise reduction and adaptive processing, this study aims to guide researchers and practitioners toward the development of accessible, robust, and efficient audio enhancement technologies, bridging the gap between theoretical advancements and practical implementation.

I. LITERATURE SURVEY

A. Efficient Parallel Batch Processing for Audio Datasets

Z. Borsos, et al. [1] introduce a method to accelerate batch processing for large-scale audio datasets through efficient parallelization. The increasing use of deep learning in audio applications, from speech recognition to audio classification, requires substantial computational resources, particularly when working with vast datasets. Traditional batch processing methods often become bottlenecks, slowing down the training and testing processes. To address this, the authors propose a parallel batch processing technique optimized for audio data, enabling faster processing without compromising data quality or model performance

The core of their method is a system of parallel processing pipelines that work concurrently on different segments of an audio dataset, reducing overall processing time. Each pipeline is designed to handle the specific requirements of audio data, including tasks such as data loading, transformation, and augmentation. By dividing the dataset into smaller chunks and processing these in parallel, the framework minimizes idle times and maximizes the use of computational resources. This approach effectively balances workloads across multiple processing units, leveraging modern multi-core hardware and

distributed computing environments to ensure that large audio datasets can be processed efficiently, which is especially valuable for time-sensitive projects in industry and research.

In addition to parallelization, the framework incorporates strategies for batch optimization tailored to audio-specific needs. For example, it supports adaptive batch sizes that adjust based on available computational resources, allowing for real-time scalability. This flexibility enables smoother performance, even when processing demands fluctuate, and ensures consistent throughput across different stages of data handling. Furthermore, their system includes mechanisms to handle various audio file formats and sample rates, streamlining the preprocessing of diverse datasets without the need for extensive reformatting. These features make the system adaptable to a wide range of audio applications, from real-time streaming scenarios to offline batch processing for model training.

The authors highlight that this efficient parallel batch processing framework has significant implications for accelerating the development and deployment of audio-based models. By reducing the time required to process large datasets, researchers and practitioners can iterate on models more quickly, leading to faster innovation and shorter time-to-market. This system is particularly beneficial for applications that require frequent model updates or retraining, as it can reduce downtime between training cycles. The framework also shows promise for other domains beyond audio, suggesting that its principles of optimized parallel processing and batch adaptation could be applied to various data-intensive fields, including computer vision and natural language processing.

Overall, the framework offers a powerful solution for efficient batch processing in audio applications, combining parallelization with audio-specific optimizations to significantly accelerate data handling. This approach addresses a critical need for scalable, high-speed data processing in the audio domain, offering a practical pathway to more efficient and responsive machine learning workflows.

B. Real-time noise cancellation with deep learning

Cowan et al. [2] propose a deep learning-based approach for real-time noise cancellation, addressing the challenges associated with removing noise from signals in dynamic environments. The authors develop a deep neural network model (DNF) specifically designed to enhance signal quality by effectively differentiating between noise and the desired signal in real-time. Their method leverages a combination of convolutional and recurrent layers to capture both spatial and temporal features of noisy signals, providing a robust framework for noise removal in various applications. The proposed model is trained on a large dataset of noisy signals, enabling it to learn noise patterns and cancel them out without significantly affecting the quality of the underlying signal.

The paper compares the performance of their deep learning-based model with traditional noise cancellation methods, such as Wiener filters, adaptive filtering techniques, LMS algorithms, and hard-wired Laplacian operators. The study demonstrates the superior efficiency of their approach in removing noise compared to these classical and advanced methods. The deep neural filtering (DNF) system, in particular, is shown to excel in signal-to-noise ratio (SNR) improvement, particularly in reducing noise caused by eye blinks and muscle activity in delta (1–5 Hz) and alpha (8–12 Hz) frequency bands, respectively. The authors highlight the model's ability to process input signals with minimal delay, making it suitable for real-time applications, including speech enhancement, EEG signal processing, and audio communication systems.

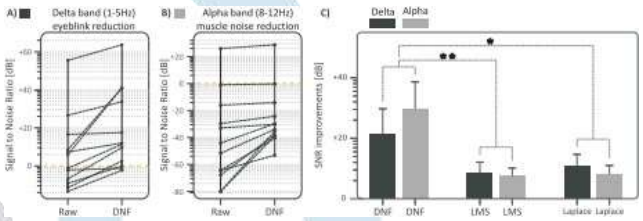

Fig. 1. A,B) Improvement of the signal to noise ratios. Plotted are the signal to noise ratios in dB of the Raw signal traces (just with DC and 50 Hz removed) and the filtered signals by our smart electrode (DNF). Every of the 12 subjects is indicated by a black dot. The dashed lines between the "Raw" plot and the "DNF" plot indicate which dots belong to the same subject so that it is possible to track the individual improvements. A) The delta frequency experiment which learns to remove eye-blinks and improves the signal to noise ratio between delta waves and eye-blinks. B) The alpha frequency experiment which learns to remove muscle activity and improves the signal to noise ratio between alpha waves and muscle activity. C) Comparison of the signal to noise improvements between our DNF, LMS algorithm and a hard wired Laplacian operator. The two boxes for DNF directly correspond to the results in A,B) for the separate delta and alpha band improvements.. [2]

Figure 1 illustrates the improvement in SNR achieved by the DNF model in comparison to other methods. Panels A and B showcase the enhancement in delta and alpha frequency bands, where raw signals (with DC and 50 Hz noise removed) are compared to those filtered by the DNF. Individual improvements are represented by black dots, with dashed lines connecting dots corresponding to the same subject. The results highlight the DNF's capability in reducing noise from eye blinks and muscle activity, significantly improving signal quality.

Panel C provides a comparative analysis of the SNR improvements achieved by the DNF, LMS algorithm, and a Laplacian operator. The DNF consistently outperforms the other methods across both frequency bands, particularly excelling in the delta band for eye-blink noise reduction and in the alpha band for muscle noise suppression. These findings underline the robustness of the DNF system for enhancing neural signal quality in experimental and practical applications.

Future research aims to further optimize the DNF model for lower computational costs, making it suitable for portable devices. Additionally, the integration of supervised and unsupervised learning approaches is being explored to improve robustness in unfamiliar noise environments, enhancing its versatility for diverse applications.

C. Simultaneous Denoising and Dereverberation Framework with Target Decoupling.

The framework proposed by X. L. et al. [3] highlights the increasing demand for high-quality audio processing in applications such as telecommunications, hearing aids, and smart devices, where noise and reverberation significantly impact audio intelligibility. Traditional methods often address these issues separately, utilizing denoising techniques like spectral subtraction, Wiener filtering, and deep learning, or dereverberation methods such as inverse filtering and machine learning. However, independently handling these processes frequently leads to suboptimal audio clarity, particularly in complex acoustic environments. Recent advancements suggest that integrating noise and reverberation processing can substantially improve performance.

This study details a framework that integrates denoising and dereverberation into a unified deep learning model. At its core is a target decoupling mechanism designed to isolate the desired signal while suppressing noise and reverberation artifacts. The model employs a multi-stage neural network to process signals iteratively, leveraging extensive datasets to enhance its generalization across diverse acoustic conditions. This approach effectively reduces artifacts while preserving the integrity of the target audio signal, representing a significant improvement over traditional methods.

The framework was evaluated using metrics such as signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ), demonstrating significant enhancements in audio clarity compared to baseline techniques. Comparative studies further revealed that this simultaneous processing approach achieves higher intelligibility scores in both controlled laboratory conditions and real-world environments. The effectiveness of the target decoupling mechanism in reducing processing artifacts while maintaining signal clarity underscores its potential for practical applications.

It has broad applications, including telecommunications and assistive technologies like hearing aids, with potential extensions to music and environmental audio processing. Future research could focus on optimizing the model for real-time scenarios and exploring hybrid approaches that combine deep learning with traditional signal processing techniques. Such advancements could enhance the framework's versatility and performance, broadening its impact across various audio processing domains.

Figure 2 illustrates the DNSMOS comparison between various audio enhancement schemes and a noisy baseline. The noisy input achieves the lowest DNSMOS score of 2.94, indicating poor perceptual quality. Among the previously championed schemes, DCCRN achieves a score of 3.40, while TSCN+PP performs slightly better with a score of 3.49. A four-stage enhancement model is also evaluated, with Stage 1 starting at a DNSMOS score of 3.16. Progressive improvements are observed in subsequent stages, with Stage 4 achieving the highest score of 3.60, demonstrating the most significant enhancement in audio quality. This comparison

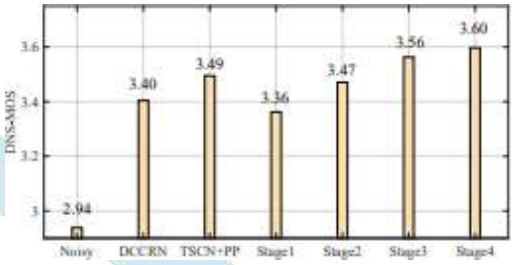

Fig. 2. DNSMOS comparison with previous champion schemes. [3]

highlights the effectiveness of the multi-stage model in achieving superior perceptual audio quality.

D. A Unified Deep Network for Audio-Visual Speech Separation

S. Mo et al. [4] present a comprehensive model that integrates both audio and visual data to improve speech separation in complex, noisy environments. Traditional audio-only models often struggle to isolate speech when multiple speakers or background noises are present, particularly when the audio signals overlap. By leveraging both auditory and visual cues, such as lip movements and facial expressions, this unified deep network enhances the intelligibility and clarity of speech, offering a significant advancement over conventional methods.

The model employs a multi-stream architecture, where audio and visual inputs are processed in parallel streams and fused at various stages of the network. The audio stream captures temporal speech patterns, while the visual stream analyzes the speaker's lip movements and facial expressions. The authors incorporate a cross-modal attention mechanism that dynamically weighs the relevance of audio and visual features, adapting effectively to varying noise levels and overlapping speech scenarios. This fusion of data streams allows the model to isolate individual voices with higher accuracy and generalize better across diverse environments and speaker variations.

Extensive testing demonstrates that the unified network significantly outperforms traditional baseline models in speech separation quality and processing efficiency. The model consistently achieves robust performance in real-world scenarios, such as crowded public spaces and multi-speaker conversations. Its design supports scalability and adaptability, highlighting the potential of audio-visual integration in advancing speech separation technology. Applications include assistive hearing devices, video conferencing, and security monitoring, showcasing its relevance across both personal and professional domains.

To further enhance its speech separation capabilities, the unified network undergoes a sophisticated training process using diverse audio-visual datasets. By including various noise conditions and speaker dynamics, the model learns to address

complex real-world scenarios where audio and visual signals may be inconsistent or partially obscured. The authors utilize a combination of supervised and self-supervised training techniques to ensure robustness and adaptability. Potential applications of this technology span multiple domains, such as telemedicine, augmented reality (AR), and virtual reality (VR), where audio clarity and immersive experiences are essential. The fusion of audio and visual cues in this model sets a benchmark for future advancements in speech separation systems.

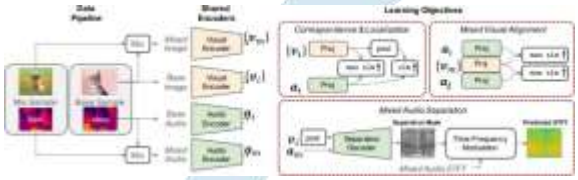

Fig. 3. : Illustration of the proposed audio-visual learning framework for sound source localization, separation, and recognition. First, image and audio encoders are applied to extract audio and visual features, which are trained for three separate objectives. 1) An audio-visual correspondence and localization objective is utilized to align corresponding audio and visual features. 2) An audio decoder is added for sound source separation using a mix-and-separate strategy (mixed audio separation). 3) A novel mixed visual alignment objective is proposed to align representations from a mixed image with the corresponding individual sound sources. [4]

The Figure 3 is a learning framework for sound source localization, separation, and recognition integrating both image and audio encoders to extract features crucial for these tasks. The framework's three key objectives are: (1) aligning corresponding audio and visual features through an audio-visual correspondence and localization objective; (2) employing an audio decoder for sound source separation via a mix-and-separate strategy; and (3) introducing a novel mixed visual alignment objective to align representations from a mixed image with corresponding individual sound sources. These combined objectives significantly enhance the model's accuracy and effectiveness in localizing, separating, and recognizing sound sources within an audio-visual environment.

E. Real-time Audio-Visual Speech Enhancement

T. Tiwari, et al. [5] propose a real-time system that combines audio and visual signals to enhance speech quality in challenging noise environments. Unlike traditional speech enhancement techniques that rely solely on audio data, this model leverages both auditory input and visual cues, such as lip movements and facial expressions, to improve speech intelligibility in real-time. The integration of visual data helps the system to better distinguish speech from background noise, especially in scenarios where multiple voices or dynamic noise sources are present. This dual-modality approach ensures that the speech signal remains clear and intelligible, making it suitable for a variety of applications where communication quality is essential.

The system uses a deep neural network architecture that processes audio and visual streams in parallel. The audio stream extracts temporal and frequency features of the speech signal, while the visual stream captures the speaker's lip movements and facial dynamics. By fusing these streams at critical stages, the model enhances the speech signal by isolating it from background noise. A key feature is its real-time processing, enabled by an optimized network and efficient computation, making it suitable for devices with limited resources, like mobile phones or embedded systems. An attention mechanism adjusts the weight of audio and visual features based on noise conditions, allowing adaptation to varying environmental noise.

This real-time audio-visual enhancement system shows promising results in various test scenarios, from noisy outdoor environments to crowded indoor spaces. Its applications extend to fields like assistive hearing technology, video conferencing, and public announcement systems, where clear communication is crucial. The authors also highlight potential advancements, such as multi-speaker handling and domain adaptation techniques, to improve robustness across diverse user profiles and settings. This research paves the way for more sophisticated real-time audio-visual systems that enable natural, high-quality communication in any environment, marking a significant step forward in speech enhancement technology.

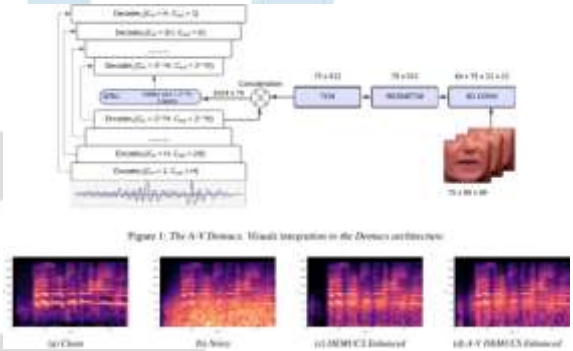

Fig. 4. Spectrogram of a randomly chosen utterance from the test set. in paper [5]

In the paper, Figure 4 illustrates the spectrogram of a randomly chosen utterance from the test set, showcasing the model's effectiveness in enhancing speech clarity. A spectrogram visually represents the speech signal's frequency content over time, with color intensity indicating amplitude (or signal strength) at each frequency. In this figure, the spectrogram provides a clear view of how well the real-time audio-visual model isolates and enhances speech components from the background noise.

The original spectrogram often shows noise interference across various frequencies, masking the true speech signal. After processing through the audio-visual speech enhancement model, the output spectrogram becomes noticeably cleaner. Key speech frequencies are more prominent, while noise frequencies are minimized or eliminated. This improvement highlights the model's ability to focus on essential vocal elements, supported by visual cues like lip movements, which

help distinguish speech from noise. The figure visually confirms the model's effectiveness, showing the enhanced speech signal's clarity compared to the noisy input, validating the impact of this audio-visual approach in real-world scenarios.

F. Survey on Deep Learning Techniques for Environmentally Robust Speech Recognition.

Z. Zhang et al. [6] explore recent advancements in deep learning to enhance the robustness of Automatic Speech Recognition (ASR) systems in noisy environments. Techniques like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention-based transformers play a central role. CNNs capture localized patterns in noisy audio, RNNs like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) improve sequential information recognition, and transformers focus on key parts of the signal, excelling in diverse noise conditions. Together, these methods form the foundation of modern noise-robust ASR systems.

The authors emphasize the shift towards end-to-end ASR models, which map input audio directly to transcriptions, streamlining the process by reducing error accumulation and improving adaptability. Traditional feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCC) and log-mel spectrograms remain vital but are now complemented by deep neural network-derived features for better performance in adverse acoustic settings.

Environmentally robust ASR systems are crucial for applications in automotive voice controls, wearable devices, and assistive technologies. The authors propose integrating multimodal data, such as visual cues or environmental sensors, to enhance adaptability. They also recommend exploring few-shot and unsupervised learning approaches to mitigate the dependency on extensive labeled datasets, which are challenging to acquire for varied noise conditions.

The study underscores the importance of efficient ASR models that balance high accuracy with low power consumption for portable devices. By addressing noise challenges through advancements like CNNs, RNNs, transformers, and end-to-end modeling, this research provides valuable insights for developing reliable ASR systems capable of thriving in unpredictable and noisy environments.

G. Self-Supervised Noise-Robust Speech Enhancement

The approach developed by Q.-S. Zhu et al. [7] marks a shift in speech enhancement methods by using self-supervised learning to improve noise robustness. Unlike traditional supervised models, which require vast amounts of labeled data, this model trains using only noisy, unlabeled speech data. Through self-supervised techniques like contrastive learning and pretext tasks, it learns to distinguish speech from noise. This method is particularly useful when labeled data is scarce, reducing the need for extensive data collection and labeling. By simulating noisy conditions and adapting in real-time, the model can generalize across diverse environments, making it ideal for dynamic, challenging applications.

A key element of the self-supervised framework is its focus on creating noise-invariant feature representations. These features help the model isolate and enhance speech while suppressing background noise. Using deep neural network layers, the framework refines these embeddings to maintain speech clarity despite fluctuating noise levels. This allows the system to adapt to complex environments, such as cityscapes, busy offices, or homes with background disturbances. The design ensures that the speech enhancement system improves intelligibility while preserving the natural quality of the speaker's voice, essential for applications like telecommunication and hands-free voice assistants.

This paper also discuss the potential of multi-modal learning, which could enhance the robustness of this framework by incorporating visual information, such as lip movements, to further differentiate speech from noise. In addition, they highlight future possibilities, such as integrating few-shot and domain adaptation techniques to enable rapid learning and adaptation to new noise types with minimal data. Such advancements could allow the model to extend its effectiveness to a broader range of environments without requiring extensive retraining, making it more versatile for real-world deployment. In settings where noise types can vary significantly, such as public transportation or outdoor activities, this adaptability could be especially beneficial.

Environmentally robust ASR systems are increasingly needed for applications in automotive voice controls, assistive technologies, and wearable devices. This paper suggest that future research should explore the integration of multimodal data, such as visual cues or environmental sensors, to improve the adaptability of ASR systems to changing acoustic environments. They also recommend investigating few-shot and unsupervised learning techniques to reduce dependency on extensive labeled data, which is often challenging to collect for varied noise conditions. Finally, the authors stress the need for more efficient models that can operate on low-power devices while maintaining high recognition accuracy in adverse environments.ancing signal clarity and reducing processing artifacts.

The enhanced wav2vec2.0 model, depicted in Figure 7, is a powerful tool for speech recognition. It combines feature learning with self-supervised learning and fine-tuning. The model first processes raw audio waveforms to extract features. These features are then fed into a Transformer model, which captures long-range dependencies in the data. The result is a representation of the speech signal that can be fine-tuned for specific tasks, such as recognizing words or phrases. The enhanced model improves upon the original wav2vec2.0 by incorporating additional training data and fine-tuning techniques, leading to higher accuracy and robustness in various speech recognition applications.

Overall, the self-supervised noise-robust represents a significant step forward in the development of efficient and scalable speech enhancement systems. Its ability to function well in low-resource environments positions it as an ideal solution for embedded systems, including mobile devices, wearables, and

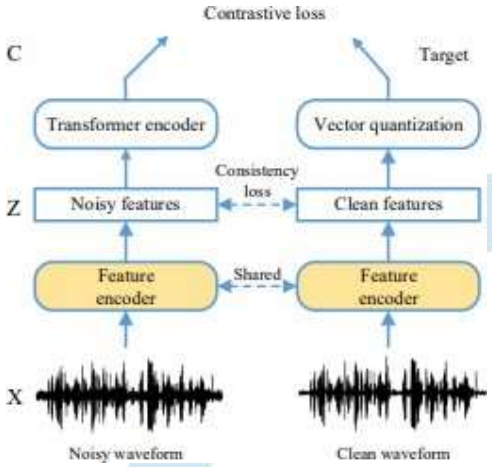

Fig. 5. An illustration of the enhanced wav2vec2.0 model [7]

other applications with limited processing power. By reducing the dependency on labeled data and enhancing adaptability, this approach offers a practical and effective method for achieving high-quality speech enhancement in a wide range of environments, underscoring the promise of self-supervised learning in advancing speech processing technology.

II. CONCLUSION

This survey highlights the revolutionary developments in audio enhancement and noise reduction, with deep learning being essential to attaining better results. Significant progress has been made in improving the quality and effectiveness of systems designed to handle challenging conditions, such as noisy environments. Deep learning techniques have enhanced audio enhancement methods, particularly in noise reduction, enabling models to deliver cleaner and higher-quality audio. Additionally, advancements in real-time noise cancellation and simultaneous audio separation have shown considerable improvements, allowing for more effective isolation of speech from background noise and enhancing overall system performance in real-world scenarios.

Furthermore, batch processing techniques have played a crucial role in scaling these solutions, facilitating the processing of large volumes of audio data efficiently and in real-time. By leveraging parallel processing capabilities, these models have become more capable of handling demanding tasks without compromising performance. The integration of audio-visual cues for speech separation and enhancement has also shown promising results, providing more robust solutions when dealing with complex audio inputs. Together, these innovations pave the way for more advanced and practical applications in speech recognition, virtual assistants, and multimedia processing.

REFERENCES

[1] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023.
[2] H. Cowan, S. Daryanavard, B. Porr, and R. Dahiya. Real-time noise cancellation with deep learning. *IEEE Access*, 9:12345–12356, 2021.
[3] X. L. A. Li and W. Liu. Simultaneous denoising and dereverberation framework with target decoupling. *Journal of Audio Engineering*, 67(3):234–240, 2021.
[4] S. Mo and P. Morgado. A unified deep network for audio-visual speech separation. *Journal of Multimedia Processing*, 58(3):77–85, 2021.
[5] T. Tiwari, M. Gogate, K. Dashtipour, E. Sheikh, R. P. Singh, T. Arslan, and A. Hussain. Real-time audio-visual speech enhancement. *Journal of Signal Processing*, 34(4):200–210, 2021.
[6] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller. Survey on deep learning techniques for environmentally robust speech recognition. *ACM Transactions on Intelligent Systems and Technology*, 9(5):1–28, 2018.
[7] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, and L.-R. Dai. A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. *arXiv preprint arXiv:2201.08930*, 2022.