A Survey on Enhancing Deep Learning based Speech Recognition Systems using Noise Suppression Techniques

Aryan Dande[1], Ayush Gala[2], Omkar Bhosale[3], Sagar Abhyankar[4]
Department of Computer Engineering
Pune Institute of Computer Technology, Pune 411043, India.
{[1]aryan.dande, [2]ayushgala2, [3]omkarbhosale277, [4]sagarabhyankar18}@gmail.com

***Abstract*—The field of speech enhancement has witnessed significant advancements in recent years, driven by the pursuit of improving speech recognition systems' robustness in noisy environments. Various sophisticated approaches have emerged, leveraging deep learning techniques to mitigate the detrimental effects of noise on speech signals. This research paper presents a comprehensive survey of diverse noise suppression techniques in deep learning for improving the performance of speech recognition systems in challenging environments. The survey analyzes the methodologies implemented by Deep Neural Networks (DNNs), Deep Filter Nets, Multi-branched Encoders, Long Short-term Memory (LSTM) Networks, Gated Recurrent Unit (GRU) Networks, Deep Complex Convolutional Recurrent Networks (DCCRNs), and the use of MFNet. Each technique's architecture, key findings, training approach, and performance against non-stationary noise are thoroughly examined. The significance of noise suppression techniques is emphasized, addressing their role in enhancing speech signal quality and system accuracy. The paper aims to provide valuable insights into practical applications and potential solutions for real-world challenges in speech recognition systems.**

***Index Terms*—Noise Suppression, Speech enhancement, Deep Neural Networks, Deep Filter Nets, Autoencoders, LSTM networks, GRU networks, CNNs, MFNet.**

I. Introduction

There are many obstacles that make real-world speech recognition of a challenging problem. [1] Accents, speaking styles, different possible pronunciations, various languages, and noise – are some of these obstacles. [2] There's a substantial loss in accuracy when we move from a controlled experimental setup to real-life situations due to the presence of different overlapping noise streams. Despite this, automatic speech recognition has abundant usage in dictation, human-machine interfaces, and control of machines among others. [3] Speech recognition systems have gained significant traction in various applications, ranging from remote work scenarios to voice-activated technologies. However, one of the key challenges that limit their performance in real-world settings is the presence of background noise, which can significantly degrade the accuracy and robustness of these systems. [4] To address this critical issue, researchers have extensively explored various noise suppression techniques, leveraging the power of deep learning. The utilization of deep neural networks (DNNs) has emerged as a promising approach, offering the potential to effectively enhance speech quality in challenging and noisy environments.

This paper presents a detailed survey of diverse noise suppression techniques in deep learning, highlighting their significant contributions to improving the performance of speech recognition systems. The survey encompasses a comprehensive analysis of various methodologies, including the application of deep neural networks, deep filter nets, autoencoders, long short-term memory (LSTM) networks, gated recurrent unit (GRU) networks, convolutional neural networks (CNNs), as well as the use of MFNet. Each technique is dissected to understand its underlying architecture, key findings, approach to training, and performance evaluation against non-stationary noise.

The significance of noise suppression techniques in the context of speech recognition is underscored, emphasizing the vital role they play in enhancing the accuracy, intelligibility, and overall quality of speech signals. The complexities associated with noise environments necessitate robust methods that can effectively separate speech from various types of interfering noise, thereby ensuring optimal performance of speech recognition systems. Consequently, this survey not only highlights the advancements in noise suppression techniques but also provides valuable insights into their practical applications and their potential for addressing real-world challenges. This survey aims to provide a comprehensive overview of the latest methodologies in noise reduction for speech recognition, highlighting the key findings and performance analyses from notable research works in the domain. By comprehensively exploring the intricacies of each approach, this survey aims to provide a holistic understanding of the advancements in noise suppression techniques and their transformative impact on enhancing the capabilities of speech recognition systems.

II. Deep Neural Networks

A. Architecture

The architecture and training approach for the Deep Neural Network (DNN) model used for speech enhancement generally involves a feed-forward neural network with multiple layers of non-linear activation functions, specifically sigmoid for

hidden units and linear for the output unit. [5] The DNN aims to model intricate, non-linear relationships between noisy speech features and clean speech features. Prior to supervised fine-tuning, the DNN undergoes a crucial unsupervised pre-training phase, which mitigates the risk of getting stuck in local minima. [6]

This pre-training phase employs Restricted Boltzmann Machines (RBMs) in a layer-by-layer fashion, with the first RBM handling real-valued input features in a Gaussian-visible layer and a binary hidden layer, followed by Bernoulli-Bernoulli RBMs. These RBMs are trained greedily, one layer at a time, using contrastive divergence (CD) as the objective criterion. Subsequently, the DNN is fine-tuned in a supervised manner, with the training objective based on the Mean Squared Error (MSE) between the DNN's estimated output and the reference clean speech features. The fine-tuning phase utilizes the back-propagation algorithm, and a mini-batch stochastic gradient descent approach optimizes the network. Two critical hyperparameters, learning rate, and momentum, guide the training process, controlling parameter update sizes and aiding convergence. The DNN's capability lies in its ability to automatically learn complex relationships between noisy and clean speech without imposing specific assumptions about their relationship. This is achieved by combining temporal information (acoustic context over time) and spectral information (full-band spectrum details) into a single extended input feature vector. Notably, the DNN does not assume independence among different dimensions, as is common in Gaussian mixture models (GMMs), allowing it to model more intricate data relationships.

B. Key Findings

The experimental results of the Deep Neural Network (DNN) for speech enhancement revealed several crucial findings. First, training the DNN with a diverse set of 104 noise types significantly improved its performance, especially in challenging noise environments like the HF channel. [6] This diversity helped the DNN generalize better to various noisy conditions. Second, deeper DNN architectures with multiple hidden layers outperformed shallow networks, emphasizing the effectiveness of deeper models for speech enhancement. Third, the length of the acoustic context window had an impact on DNN performance, with longer context windows (up to 11 frames) providing better results by offering more information about the speech signal's characteristics, although using too many frames introduced irrelevant information. Moreover, increasing the training data size led to performance improvements, with a notable jump in quality when the training set reached 5 hours, but further increases beyond 100 hours did not yield significant gains.

Additionally, the choice of training targets played a critical role, with the proposed normalized clean log-power spectra outperforming ideal ratio masks and short-time Fourier transform spectral masks. Lastly, comparing DNNs trained with Restricted Boltzmann Machine (RBM) pre-training against those with random initialization demonstrated that RBM pre-training had advantages, particularly in low Signal-to-Noise Ratio (SNR) conditions and with limited training data. These findings provide valuable insights into optimizing DNN-based speech enhancement systems, showing the importance of diverse training data, deep network architectures, appropriate context length, training dataset size, target selection, and the role of RBM pre-training in achieving superior performance, especially in challenging and real-world noisy environments.

C. Performance against non-stationary noise

The DNN outperformed conventional methods, like LogMMSE, even when trained with limited noise types, demonstrating its ability to significantly reduce non-stationary noise and enhance speech quality. Moreover, the DNN exhibited adaptability to changing noise environments, maintaining consistent speech enhancement as noise types shifted. When comparing different training data sizes and configurations, a larger dataset (625 hours) slightly improved performance, with the best DNN system achieving an average Perceptual Evaluation of Speech Quality (PESQ) of 3.15. It offered substantial PESQ gains over LogMMSE, akin to the difference between LogMMSE and unprocessed noisy speech.

III. Deep Filter Nets

A. Introduction

Monaural speech enhancement is essential in various applications. [7] Most state-of-the-art methods work in the short-time Fourier transform (STFT) domain and use deep neural networks to estimate time-frequency (TF) masks. Some use real-valued masks, while others use complex masks.

The paper by H. Schröter et al. [8] introduces DeepFilterNet as an open-source speech enhancement framework based on deep filtering. Deep filtering is a complex filter applied in the TF domain to enhance audio. The deep filter incorporates information from multiple time steps and exploits local correlations within each frequency band. It can recover signal degradations like notch-filters or time-frame zeroing, making it suitable for speech enhancement. The motivation behind using deep filtering to enhance speech, especially for lower frequencies where periodic components have the most energy is also explored subsequently to facilitate real-time speech enhancements on embedded devices. [9]

B. Framework Overview

The DeepFilterNet framework processes audio signals using the short-time Fourier transform (STFT) to work in the frequency domain. It supports various sampling rates and STFT window sizes. The framework consists of two stages: the first stage enhances the spectral envelope, and the second stage focuses on enhancing the periodic components of speech. Deep filtering is used in the second stage. The framework also introduces sparsity and low complexity in its design.

The paper describes the deep neural network (DNN) model used in DeepFilterNet. It uses standard DNN layers such as convolutions, batch normalization, ReLU activation, and more. The architecture follows a UNet-like structure. Grouped

linear and GRU layers are utilized to introduce sparsity. The model includes an encoder/decoder architecture for predicting ERB-scaled gains and deep filter coefficients. The DeepFilterNet framework uses on-the-fly data augmentation to enhance the model's robustness. It mixes clean speech signals with various noise types at different signal-to-noise ratios (SNR). Second-order filters, EQs, random gains, resampling, and room impulse responses are used to augment speech and noise signals. The framework also supports training attenuation-limited models to maintain environmental awareness.

The loss function mentioned uses a compressed spectral loss to implicitly learn ERB gains and filter coefficients. The loss function models the perceived loudness. A local signal-to-noise ratio (LSNR) loss term is used to ensure that deep filtering is only applied to relevant segments of the audio.

C. Conclusion

The paper [8] evaluates the DeepFilterNet framework with various experiments, comparing it to related work. The results show that DeepFilterNet outperforms complex ratio masks (CRMs) over different FFT window sizes. The performance of CRMs decreases for smaller FFT window sizes. DeepFilterNet is also compared to other models based on objective metrics such as WB-PESQ, SI-SDR, and computational complexity (MACS). The paper concludes that DeepFilterNet is a low-complexity speech enhancement framework that performs on par with other algorithms. It highlights the computational efficiency of DeepFilterNet. Future work may focus on improving the perceptual approach by better applying deep filtering to periodic speech components.

IV. Use of Autoencoders

A. Approach

Making speech sound better in noisy environments is important for applications like speech recognition and hearing aids. One approach to enhance speech is to use autoencoders that can specifically focus on denoising the speech. The paper [10] suggests the use of a combination of mathematical models: one for separating clean speech from noise and another for representing speech variations. The authors call this combination a Variational Autoencoder (VAE) and Non-negative Matrix Factorization (NMF). The VAE helps understand the speech, while NMF helps with noise. They also deal with variations in how loud the speech is. The goal is to remove noise and make the speech clearer. However, making the VAE work well with noise is tricky because it's usually trained only on clean speech data.

The research done by Cheng et al. [12] suggests the use of multibranched encoders for performing the denoising. The Speech Enhancement model suggested here is divided into 3 parts.

1) DSDT: DSDT stands for Dynamically Sized Decision Tree. This tree is constructed using prior knowledge about speech and noise characteristics obtained from a training dataset. The paper takes into account attributes at the utterance level, such as the speaker's gender and signal-to-noise ratio, as well as signal-level attributes like low and high-frequency components to build the DSDT.

2) Training of Multiple Encoders: The DSDT divides the training dataset of clean speech into different clusters. Each specific cluster is then used to train an encoder.

3) Training of Decoder: A single decoder is constructed using multiple encoders. Various decoder architectures are explored and evaluated, including Linear Regression, Convolutional Neural Network, and Best-First based decoders.

B. Training

For the models suggested in [10], a VAE is trained on clean speech data, and then a noise-aware encoder is introduced. This new encoder is trained to make the latent variables estimated from noisy mixtures as close as possible to those inferred from clean speech. The goal is to minimize the divergence between these latent variables in the presence of noise. The research evaluates this approach on various datasets, including seen and unseen noise conditions, and shows that the noise-aware VAE consistently outperforms standard VAE and a fully-connected DNN model in terms of speech enhancement, especially in low signal-to-noise ratio (SNR) scenarios.

In the case of the multi-encoder approach, [12] the noisy speech undergoes processing by individual component models. The multiple outputs generated by these models are then combined within the decoder to produce the final enhanced speech. Essentially, the goal is to create several "matched" SE component models, enabling the decoder to incorporate both "local" and "matched" information from these component models to enhance speech under specific noise conditions. The authors are leveraging the concept of ensemble learning to ensure that the model doesn't underperform. By training each encoder on a specific cluster, each encoder becomes highly proficient at cleaning a particular type of signal. This approach enhances the model's capability to effectively clean a wide range of signals with high accuracy. Due to its flexible design, the model can be customized and trained to handle various types of noises and attributes, resulting in an exceptionally accurate model. Therefore, selecting the right hyperparameters for the DSDT is crucial for improving its accuracy.

C. Conclusion

The results show that the noise-aware VAE consistently outperforms the standard VAE in various signal-to-noise ratio (SNR) scenarios, with particularly noticeable improvements at low SNR levels. It also outperforms a fully connected DNN model. [11] Furthermore, the study demonstrates that even with a small amount of labeled data, the noise-aware VAE can significantly enhance performance, suggesting its potential to address overfitting issues in supervised training strategies.

Consequently, thanks to the design, the multi-encoder model achieves higher PESQ scores than BLSTM. Moreover, it demonstrates robustness against various types of noise, with minimal fluctuations in PESQ scores when cleaning out these noises, outperforming BLSTM in this regard. The architecture

also offers flexibility in terms of using various decoder designs and training attributes. This adaptability makes it feasible to fine-tune the model to meet specific use cases effectively.

V. Long Short Term Memory

Deep neural networks have offered new approaches to audio processing. Many models developed for offline processing are not suitable for real-time applications, as they often rely on non-causal processing. Bidirectional RNNs are commonly used for real-time applications but require a full sequence as input.

The most widely used recurrent network in speech processing applications is the long short-term memory (LSTM) architecture. Because the LSTM addresses the vanishing gradient problem of the standard RNN, it is easier to train. In [13], an LSTM was employed to model the F0 contour. In [14], a bidirectional LSTM was employed to map a sequence of linguistic features to the corresponding sequence of acoustic features. In [15], an LSTM with a recurrent output layer was proposed to perform sequence mapping from linguistic to acoustic representations. These studies all formulate SPSS as sequence-to-sequence mapping and all demonstrate the effectiveness of LSTMs.

A. Introduction

Noise suppression is crucial for speech enhancement in various applications, including remote work scenarios. The paper [16] introduces the Dual-Signal Transformation LSTM Network (DTLN), a model for real-time speech enhancement. It combines Short-Time Fourier Transform (STFT) and a learned analysis and synthesis basis in a stacked network with less than one million parameters. The model is trained on 500 hours of noisy speech and shows competitive results. DTLN extracts information from magnitude spectra and incorporates phase information from the learned feature basis. It outperforms the DNS-Challenge baseline by 0.24 points in terms of Mean Opinion Score (MOS).

B. Approach

DTLN combines STFT-based signal transformation and a learned signal representation. In noise suppression, the goal is to separate speech and noise signals. The STFT of the noisy signal is used to predict a clean speech signal. The learned signal transformation multiplies time-domain frames with learned basis functions and is used to reconstruct the speech signal. Combining these approaches provides a robust magnitude estimation and incorporates phase information. DTLN consists of two separation cores, each with two LSTM layers followed by a fully connected layer and a sigmoid activation for mask output. The first core uses STFT-based analysis and synthesis and combines the predicted mask with the noisy signal's magnitude to generate the estimated speech. The second core uses a learned feature representation to further enhance the signal with phase information. Instant Layer Normalization (iLN) is applied to account for real-time processing.

Training data is created from Librispeech speech data and noise signals from various sources, resulting in a 500-hour dataset. The DNS-Challenge provides both known and blind test sets. The training setup uses the scale-sensitive negative Signal-to-Noise Ratio (SNR) as the training objective.

C. Performance

DTLN is compared to the DNS-Challenge baseline (NSNet) [17] and four additional models with different architectures and feature representations. Objective evaluation uses measures like PESQ, SI-SDR, and STOI. Results indicate improvements with DTLN over the noisy condition, especially in non-reverberant environments. Subjective evaluation, involving human judges, aligns with the objective results, showing better quality with DTLN over the noisy and baseline conditions. DTLN's real-time processing performance is evaluated. The execution time for one 32 ms frame is measured. Sequence processing proves significantly faster than frame-wise processing. The paper also discusses differences between baseline systems, particularly the impact of STFT features versus learned feature representations. Stacking networks using both STFT and learned feature transformations is shown to slightly improve performance.

VI. Gated Recurrent Unit Networks

A. Approach

As an alternative to the LSTM, the Gated Recurrent Unit (GRU) architecture was proposed in [18]. In [19], the GRU was found to achieve better performance than the LSTM on some tasks. The simplified architecture proposed by the authors is a variant of the Long Short-Term Memory (LSTM) network that aims to reduce computational complexity without compromising the quality of the synthesized speech. It retains only the forget gate, discarding other components like the input gate, output gate, and peep-hole connections. The forget gate determines how much of the previous memory should be forgotten and how much of the current input should be stored in the memory cell, and is responsible for controlling the flow of information within the memory cell.

The paper [20] suggests the use of fused features from both DNN and GRU models for the enhancement of speech. For doing so, the speech signal is first decomposed into 25 milliseconds windows with each Window smoothened by applying the Hamming Window clause. The Discrete Fourier Transform is obtained for each frequency point upon which log transformations are performed on each frame to get the LPS feature.

B. Training

A DNN with three hidden layers is typically used to learn the mapping between the local LPS features of noisy speech and clean speech to estimate the clean LPS features from the noisy ones in the first stage. The DNN will then give its own predicted clean frame. This makes sure that the frames are cleaned on a local level. To capture the effective contextual information in features, the layer of feature fusion is adopted.

DNN-GRU has a cascade architecture consisting of a prior NN (DNN) and a posterior NN (GRU-NN) for the first and second stages of DNN-GRU. The cleaned frame and the original frames are fused together either by union operation or addition operation. This is done to ensure that the time series information between noisy speech frames is not lost after cleaning at the local level. This fused frame is then fed to the GRU.

The first GRU layer has 1024 cells, which encode the input and pass its hidden state to the second GRU layer, which has 512 cells. The two GRU layers are used to establish the mapping from the new fused features to the training target features to achieve the whole frames speech enhancement, and meanwhile preserving the contextual information of speech.

VII. Complex Convolutional Recurrent Networks

A. Modelling using CNNs

After the success of DBN-HMM hybrid models for speech recognition, work was carried out using CNN based acoustic models. CNN for speech data with convolution along the time axis was first proposed by LeCun et al. [21], but no validation was carried out at the time. It was theorized that convolution along time will help obtain features robust to small temporal shifts.

This was confirmed by Lee et al. [22] and, Hau and Chen [23]. In these works, convolution was applied over windows of acoustic frames that overlap in time. This resulted in learning acoustic features that were relatively more stable with respect to variations arising from speakers and genders. Abdel-Hamid et al. [24] achieved significant improvements by applying convolution and max-pooling along frequency axis rather than the time axis. Applying such a convolution was found to generate features robust to small frequency shifts, which often happens because of different speakers and even different moods. More researchers explored convolution over both time and frequency axes simultaneously [25], [26].

B. Need For Neural Network with Complex Weights

For a long time, it was believed that estimating phase was a challenging task. As a result, the earlier studies primarily concentrated on training targets related to magnitude while overlooking phase information. The approach involved synthesizing estimated speech by applying the estimated magnitude to the noisy speech phase. However, this approach limited the potential performance, as the phase of the estimated speech could deviate significantly in the presence of strong interference.

To address this issue, a solution emerged in the form of neural networks designed to handle complex inputs and complex weights. These neural networks incorporate phase information during forward propagation, as the complex weights also affect the phase of the inputs. Furthermore, activation functions have been designed with phase considerations in mind, further enhancing the ability to estimate speech accurately.

C. Architecture

The Convolutional Recurrent Network (CRN) can be thought of as a Convolutional Encoder-Decoder Network with an important addition: it includes two Long Short-Term Memory (LSTM) layers positioned between the encoder and decoder. These LSTM layers are crucial for capturing and modeling temporal dependencies between the different frames of a speech signal. The CRNs originally described in [27], is an essentially causal CED architecture with two LSTM layers between the encoder and the decoder. Here, LSTM is specifically used to model the temporal dependencies. The encoder is composed of five Conv2d blocks, and its primary function is to extract high-level features from the input features while also reducing the resolution. On the other hand, the decoder is responsible for reconstructing the low-resolution features back to their original size, effectively reverting the changes made by the encoder. This design choice results in a symmetrical structure for the encoder-decoder, creating a balanced architecture.

These models however, treat real and imaginary parts as two input channels, only applying a real-valued convolution operation with one shared real-valued convolution filter, which is not confined with the complex multiply rules. Hence the networks may learn the real and imaginary parts without prior knowledge. To address this issue, in the paper by [28], the proposed DCCRN modifies CRN substantially with complex CNN and complex batch normalization layer in encoder/decoder, and complex LSTM is also considered to replace the traditional LSTM. Specifically, the complex module models the correlation between magnitude and phase with the simulation of complex multiplication. When training, DCCRN estimates CRM and is optimized by signal approximation (SA)

D. Performance

On the simulated WSJ0 test set, the four DCCRNs outperform the baseline LSTM and CRN, which indicates the effectiveness of complex convolution. DCCRN-CL achieves better performance than other DCCRNs with a PESQ score of 2.972, 3.301, 3.559 on the 0dB, 5dB, and 10dB models respectively. This further shows that complex LSTM is also beneficial to complex target training. Moreover, we can see that full-complex-value network DCCRN and DCUNET are similar in PESQ given that the computational complexity of DCUNET is almost 6 times that of DCCRN-CL, according to run-time tests.

VIII. Use of MFNet

A. Introduction

The current state-of-the-art on Deep Noise Suppression (DNS) Challenge is MFNET. The paper [29] begins by highlighting the significant advancements in speech enhancement techniques, particularly with the rise of deep learning. It categorizes these techniques into two main groups: time domain methods and T-F domain methods, focusing on the latter due to its success in DNS Challenge competitions. The paper's main objective is to develop an effective T-F domain system for

single-channel speech enhancement. The introduction points out the challenges in speech enhancement. Current approaches often require complex and resource-intensive network architectures. Achieving competitive performance using basic network structures and straightforward methods is challenging. A debate exists on the effectiveness of masking and mapping methods in T-F domain speech enhancement.

B. Proposed solution:

The paper [29] introduces MFNet as a novel approach to address the challenges. MFNet is described as a direct and simple network designed to map both speech and reverse noise. It is constructed by stacking Global Local Former Blocks (GLFBs) to balance global processing and local interaction. The network is designed to use real-valued Short-Time Discrete Cosine Transform (STDCT) features, avoiding the complexities of phase information estimation. The model's structure includes an encoder, decoder, and bottleneck, with jump layer connections for direct summation. Importantly, the network adopts the mapping method, which eliminates the need for masks. An ablation study is conducted to compare the performance of masking and mapping methods: The results favor the mapping approach, particularly when mapping reverse noise.

C. Performance

When MFNet is evaluated on the DNS 2020 test set, it achieves competitive results in PESQ (Perceptual Evaluation of Speech Quality), STOI (Short-Time Objective Intelligibility), and SNR (Signal-to-Noise Ratio) metrics. The model exhibits a low computational complexity of 6.09 GMACs/s. A real-time factor (RTF) test shows a value of 0.236. MFNet is a promising solution for speech enhancement, emphasizing its simplicity and direct mapping. The proposed network architecture, based on GLFB modules, demonstrates strong performance when compared to other state-of-the-art models. Future work is underway to make the model causal for practical real-world applications.

IX. CONCLUSION

The area of deep learning has seen rapid progress and has led to significant improvements in various fields. Through this survey, we have unveiled crucial insights into the advancements in enhancing the performance of speech recognition systems, particularly in challenging real-world environments. The research highlights the efficacy of methodologies such as deep neural networks, deep filter nets, autoencoders, LSTM networks, GRU networks, CNNs, and MFNet in significantly improving the accuracy, intelligibility, and overall quality of speech signals. These techniques play a vital role in addressing the complexities associated with diverse noise environments, ensuring optimal performance and robustness of speech recognition systems.

The findings emphasize the significance of advanced noise suppression techniques in facilitating the widespread adoption of speech recognition technology across multiple societal domains. Improved speech recognition systems can revolutionize human-machine interactions [30], enhance accessibility for individuals with disabilities [31], and streamline various daily tasks through voice-activated technologies. In sectors such as healthcare, education, customer service, and communication, the integration of robust noise suppression techniques can foster greater efficiency, accuracy, and accessibility [32], [33]. Moreover, the implementation of these techniques in emerging technologies, including virtual assistants, smart devices, and automated systems, has the potential to transform the way individuals interact with their environments. Furthermore, the impact of better noise suppression techniques extends to remote work scenarios, where clear communication is essential for seamless collaboration and productivity. By mitigating the adverse effects of background noise, these advancements can significantly improve the quality of teleconferencing, virtual meetings, and remote communication, thereby fostering a more efficient and productive remote work culture.

Ultimately, the advancements in noise suppression techniques presented in this survey not only pave the way for improved speech recognition systems but also hold the promise of creating a more accessible, efficient, and interconnected society. Embracing these advancements can lead to transformative changes, empowering individuals and organizations to leverage the full potential of speech recognition technology in various aspects of daily life and professional endeavors.

REFERENCES

[1] Kumar, Akshi & Verma, Sukriti & Mangla, Himanshu. (2018). A Survey of Deep Learning Techniques in Speech Recognition. 179-185. 10.1109/ICACCCN.2018.8748399.

[2] Rabiner, L. R., Juang, B. H., & Lee, C. H. (1996). An overview of automatic speech recognition. Automatic speech and speaker recognition: advanced topics, 1-30.

[3] Furui, S. (1999, June). Automatic speech recognition and its application to information extraction. In Proceedings of the 37th annual meeting of the association for computational linguistics (pp. 11-20).

[4] Seltzer, M. L., Yu, D., & Wang, Y. (2013, May). An investigation of deep neural networks for noise robust speech recognition. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 7398-7402). IEEE.

[5] Y. Xu, J. Du, L. -R. Dai and C. -H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7-19, Jan. 2015, doi: 10.1109/TASLP.2014.2364452.

[6] Y. Xu, J. Du, L. -R. Dai and C. -H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," in IEEE Signal Processing Letters, vol. 21, no. 1, pp. 65-68, Jan. 2014, doi: 10.1109/LSP.2013.2291240.

[7] Ochieng, P. (2022). Deep neural network techniques for monaural speech enhancement: State of the art analysis. arXiv preprint arXiv:2212.00369.

[8] H. Schroter, A. N. Escalante-B, T. Rosenkranz and A. Maier, "Deepfilternet: A Low Complexity Speech Enhancement Framework for Full-Band Audio Based On Deep Filtering," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 7407-7411, doi: 10.1109/ICASSP43922.2022.9747055.

[9] H. Schröter, A. Maier, A. N. Escalante-B and T. Rosenkranz, "Deepfilternet2: Towards Real-Time Speech Enhancement on Embedded Devices for Full-Band Audio," 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), Bamberg, Germany, 2022, pp. 1-5, doi: 10.1109/IWAENC53105.2022.9914782.

[10] H. Fang, G. Carbajal, S. Wermter and T. Gerkmann, "Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 676-680, doi: 10.1109/ICASSP39728.2021.9414060.

[11] Leglaive, S., Alameda-Pineda, X., Girin, L., & Horaud, R. (2020, May). A recurrent variational autoencoder for speech enhancement. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 371-375). IEEE.

[12] Yu, C., Zezario, R. E., Wang, S. S., Sherman, J., Hsieh, Y. Y., Lu, X., ... & Tsao, Y. (2020). Speech enhancement based on denoising autoencoder with multi-branched encoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2756-2769.

[13] Raul Fernandez, Asaf Rendel, Bhuvana Ramabhadran, and Ron Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in Proc. Interspeech, 2014.

[14] Yuchen Fan, Yao Qian, Fenglong Xie, and Frank K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in Proc. Interspeech, 2014.

[15] Heiga Zen and Hasim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2015.

[16] Westhausen, N. L., & Meyer, B. T. (2020). Dual-signal transformation lstm network for real-time noise suppression. arXiv preprint arXiv:2005.07551.

[17] Reddy, C. K., Gopal, V., Cutler, R., Beyrami, E., Cheng, R., Dubey, H., Matusevych, S., Aichner, R., Aazami, A., Braun, S., Rana, P., Srinivasan, S., & Gehrke, J. (2020). The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results. ArXiv. /abs/2005.13981

[18] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, ¨ Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[19] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[20] Wu, Z., & King, S. (2016, March). Investigating gated recurrent networks for speech synthesis. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5140-5144). IEEE.

[21] LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10):1995.

[22] Lee, H., Pham, P., Largman, Y., and Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In Advances in Neural Information Processing Systems, pages 1096–1104.

[23] Hau, D. and Chen, K. (2011). Exploring hierarchical speech representations with a deep con- volutional neural network. In 11th UK workshop on computational intelligence (UKCI '11), page 37.

[24] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., and Penn, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pages 4277–4280.

[25] Chan, W. and Lane, I. (2015). Deep convolutional neural networks for acoustic modeling in low resource languages. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pages 2056–2060.

[26] Abdel-Hamid, O., Deng, L., and Yu, D. (2013). Exploring convolutional neural network structures and optimization techniques for speech recognition. In Interspeech, volume 2013, pages 1173–5.

[27] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in Interspeech, vol. 2018, 2018, pp. 3229–3233.

[28] Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., ... & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264.

[29] Liu, L., Guan, H., Ma, J., Dai, W., Wang, G., & Ding, S. (2023). A Mask Free Neural Network for Monaural Speech Enhancement. arXiv preprint arXiv:2306.04286.

[30] S. Braun, H. Gamper, C. K. A. Reddy and I. Tashev, "Towards Efficient Models for Real-Time Deep Noise Suppression," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 656-660, doi: 10.1109/ICASSP39728.2021.9413580.

[31] Noyes, J. M., Haigh, R., & Starr, A. F. (1989). Automatic speech recognition for disabled people. Applied Ergonomics, 20(4), 293-298.

[32] Keith Bain, Sara H. Basson, and Mike Wald. 2002. Speech recognition in university classrooms: liberated learning project. In Proceedings of the fifth international ACM conference on Assistive technologies (Assets '02). Association for Computing Machinery, New York, NY, USA, 192–196. https://doi.org/10.1145/638249.638284

[33] Y. Zhao, "Speech-recognition technology in health care and special-needs assistance [Life Sciences]," in IEEE Signal Processing Magazine, vol. 26, no. 3, pp. 87-90, May 2009, doi: 10.1109/MSP.2009.931993.