

NMF based Speech Enhancement using CNN

Dr.D. Jessintha¹, Ms.P. Bini Palas², Dr. Berakhah. F Stanley³ and Dr. R. Jeen Retna Kumar⁴

¹⁻²Easwari Engineering College / Department of Electronics and Communication Engineering, Chennai, India

Email: jessintha.d@eec.srmmp.edu.in, binipalas16@gmail.com

³Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology / Department of Computer Science Engineering, Chennai, India

Email: berakhahfstanley@gmail.com

⁴Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology / Department of Electronics and Communication Engineering, Chennai, India

Email: jejinrsrch@gmail.com

Abstract—Convolutional Neural Networks (CNNs), are adept at learning complex patterns in data and can effectively denoise speech signals by mapping noisy spectrograms to clean ones. NMF, on the other hand, is a signal processing technique that can decompose a spectrogram into additive components, including a noise profile. By estimating the noise profile from the noisy speech spectrogram, NMF enables the separation of noise from the desired signal. Combining deep learning with NMF leverages the strengths of both approaches, allowing for robust and efficient speech enhancement even in challenging noisy environments. First, noisy and clean speech samples are loaded and pre-processed, including trimming or padding to a fixed length and computing magnitude spectrograms. A CNN-based deep learning model is then defined and compiled, with two convolutional layers for feature extraction from input spectrograms followed by dropout layers for regularization. This model is trained using the noisy spectrograms as input and their corresponding clean spectrograms as target output. Meanwhile, NMF is applied to estimate the noise profile from the magnitude spectrogram of the noisy speech signal, separating noise from the signal. Finally, the trained model is used to denoise the spectrogram, and the enhanced audio is reconstructed and saved. This approach combines the strengths of CNNs for learning complex mappings and NMF for noise estimation, resulting in improved speech enhancement performance.

Index Terms— Convolutional Neural Networks (CNNs), Non-Negative Matrix Factorization (NMF).

I. INTRODUCTION

Convolutional Neural Networks (CNNs) and Nonnegative Matrix Factorization (NMF) play distinct yet complementary roles in enhancement of speech. CNNs, renowned for their ability to learn complex patterns in data, are instrumental in denoising speech signals. Their involvement in the proposed framework begins with the preprocessing stage, where noisy and clean speech samples are loaded and transformed into magnitude spectrograms. These spectrograms serve as input data for the CNN-based deep learning model, which is designed to map noisy spectrograms to their corresponding clean counterparts. The model architecture includes convolutional layers for feature extraction, followed by dropout layers for regularization, effectively capturing both local and global dependencies in the input spectrograms. During training, the CNN is optimized to

minimize the mean squared error between predicted and target clean spectrograms, thus learning to denoise speech signals effectively. In contrast, NMF specializes in decomposing spectrograms into additive components, including noise profiles, which are crucial for separating noise from the desired speech signal. In the proposed framework, NMF's role is to estimate the noise profile from the magnitude spectrogram of the noisy speech signal. This involves applying NMF to the noisy spectrogram with additional constraints, such as sparsity constraints, to enhance noise suppression while preserving speech components.

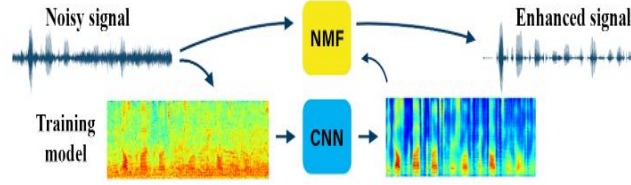


Figure 1. Proposed Method

By effectively estimating the noise profile, NMF facilitates the separation of noise from the speech signal, contributing to the overall speech enhancement process. The combined approach leveraging both CNNs and NMF synergistically integrates their respective strengths for enhanced speech enhancement. By integrating NMF's capability to estimate noise profiles with CNN's proficiency in learning complex mappings, the combined approach achieves robust and efficient denoising of speech signals, even in challenging noisy environments. This integration involves incorporating the noise profiles estimated by NMF into the CNN model, enabling direct optimization of the denoising objective. Unlike previous methods that predict activation coefficients inferred by NMF as intermediate targets, the combined approach directly optimizes the separation objective, alleviating accumulated errors and improving denoising performance. Despite the promising strides made by the integrated CNN-NMF approach in enhancing speech quality, several hurdles and constraints necessitate careful consideration. The process of integrating CNNs and NMF entails navigating through the intricate web of implementation complexities, demanding meticulous attention to diverse parameters and constraints. Notably, the incorporation of sparsity constraints within NMF stands as a critical juncture, exerting a profound influence on the precision and efficacy of noise separation. Concurrently, the architectural design of CNNs wields a significant impact on the overall performance of the denoising endeavor, underscoring the importance of strategic design decisions. Moreover, the pursuit of optimal performance through fine-tuning the combined approach often leads to an arduous journey rife with extensive experimentation and parameter adjustments, consequently amplifying computational overheads and prolonging training durations.

Furthermore, while the combined CNN-NMF approach showcases competitive prowess in controlled environments, its real-world efficacy and adaptability remain subject to scrutiny. The dynamic nature of real-world noise conditions poses a formidable challenge, necessitating robustness testing under diverse scenarios. Consequently, further research endeavours are imperative to ascertain the applicability and resilience of the combined approach in practical settings. By addressing these inherent challenges and limitations head-on, we can unravel the full potential of the integrated CNN-NMF approach, thereby ensuring its efficacy and viability in real-world speech enhancement applications. Our contributions can be succinctly outlined as follows: 1) We introduce a novel combinatorial strategy, effectively merging the strengths of CNN and NMF for speech separation. 2) We delve into a discriminative training objective, integrating sparsity constraints, thereby bolstering the combined model's capacity to attenuate noise while retaining essential speech elements.

The related works of speech enhancement using NMF and deep learning are presented in Section II. In Section III, we give system overview. Section IV is the performance evaluation showing individual and combining both DNN and NMF. Finally, we conclude this study in Section V.

II. RELATED WORKS

Prediction of NMF-based Wiener Filter for Speech Enhancement Using Deep Neural Networks states that the combination of DNNs with NMF for speech improvement, providing a possible way to lessen the difficulties caused by noisy surroundings. The method avoids the difficulties involved in encoding vectors and simplifies the improvement process by directly anticipating the NMF-based Wiener filter. This tailored prediction leads to better performance than traditional approaches by minimizing intermediate errors and enabling more accurate voice enhancement. The thorough objective evaluations highlight the suggested algorithm's effectiveness and resilience, making it a top choice for enhancing voice quality in a variety of SNR settings. In conclusion, this

novel combination of DNNs and NMF offers a convincing solution for practical applications, holding great promise for furthering the field of voice enhancement.

This combining [2] non-negative matrix factorization and deep neural networks for speech enhancement [24] and automatic speech recognition presents a unique method for robust Automatic audio Recognition (ASR) and audio enhancement using a combination of Deep Neural Networks (DNN) and Sparse Non-negative Matrix Factorization (SNMF). The study presents a full-stack solution that integrates DNN and SNMF components and improves the cost function of the DNN to spread the enhanced speech reconstruction error. Analysis of evaluations performed on CHiME-3 challenge data shows notable gains of 30% in speech quality and 10% reduction in word error rate comparing to state-of-the-art systems in both objective quality speech enhancement tests and ASR performance.

This Speech Separation [2] Based on Improved Deep Neural Networks with Dual Outputs of Speech Features [7] for Both Target and Interfering Speakers presents a novel architecture for a deep neural network (DNN) intended for speech separation, with the goal of separating the speech characteristics of a target speaker from those of an interfering speaker in mixed signals. In contrast to traditional approaches, which generally depend on manually generated features or sophisticated signal processing methods, the suggested DNN accurately represents the complicated, nonlinear relationship between the features of the two competing speakers and mixed speech signals. The DNN enhances the target speech's separation performance by improving its capacity to generalize to unseen interfering speakers by utilizing updated output speech features during training. Interestingly, this is accomplished by the suggested method without the need for any prior knowledge about the interfering speaker. The new DNN architecture's effectiveness is demonstrated by experimental data, which exhibit enhanced separation performance across a variety of objective measures. Its performance in semi-supervised scenarios, where the target speaker's training data is the sole information provided and the interfering speakers are predicted in the separation stage using several interfering speakers mixed with the target speaker [4] during training, is very remarkable.

A multi -pitch tracking algorithm for noisy speech [3] states an efficient algorithm for multipitch tracking of loud speech is described in this abstract. This algorithm integrates a number of essential elements to improve performance in demanding settings. First of all, it incorporates a better channel and peak selection technique that aids in finding important features in the midst of noise. Second, it presents a novel integration technique that enhances the algorithm's precision in pitch pattern detection by pulling periodicity information from several frequency channels. In order to create continuous pitch tracks, the method also includes a hidden Markov model (HMM). This allows for the reliable tracking of single and double pitch tracks, even in noisy environments. By incorporating these components, the suggested algorithm shows notable improvements over current techniques. Tested against a database of spoken utterances contaminated with different interferences, it performs better than other methods. The algorithm's capacity to precisely follow several pitches in noisy environments is highlighted, highlighting its potential for use in situations where reliable speech processing is needed.

DNN training based on classic gain function [3] for single-channel speech enhancement [7] and recognition states an innovative method for automatic speech recognition (ASR) and single-channel speech augmentation is described in this abstract. The method focuses on enhancing previous signal-to-noise ratio (SNR) estimation for precise noise suppression. The suggested strategy, known as GF-DNN-IRM, blends deep learning methodologies with traditional voice augmentation methods. During the training phase, the ideal ratio mask-based deep neural network (DNN-IRM) [3] is optimized using the gain function [7] of conventional speech enhancement. By using the estimated optimal ratio mask directly to generate enhanced speech, this optimization technique eliminates the need for intricate conventional speech enhancement procedures and speeds up the testing stage. In order to improve computing efficiency, the study also investigates the usage of DNNs with fewer parameters in causal processing mode. The effectiveness of the suggested approach is shown by experimental data from the CHiME-4 challenge test. On the Real Data test set, the GF-DNN-IRM technique achieves a significant relative word error rate decrease of 6.57% when compared to unprocessed speech without acoustic model retraining [3]. Remarkably, in comparable circumstances, conventional DNN-IRM techniques are unable to enhance ASR performance. This demonstrates how well the GF-DNN-IRM method works to improve speech quality and increase ASR accuracy, with potentially positive real-world applications.

A Regression Approach to Speech Enhancement Based on Deep Neural Networks [5] has supervised speech signal enhancement method which is presented in the abstract as an alternative to the traditional minimum mean square error (MMSE)-based noise reduction methods. The suggested technique makes use of deep neural networks (DNNs) to acquire the ability to map clean and noisy voice signals. An extensive training set encompassing a range of speech and noise combinations is created in order to guarantee efficacy in real-world situations with varied additive sounds. Because of its strong modelling capabilities, a DNN architecture is used

as a nonlinear regression function. Additional methods are suggested to improve the generalization capacity of DNNs across unknown noise settings, such as global variance equalization to address over-smoothing and dropout and noise-aware training methodologies. When compared to traditional MMSE-based methods, experimental results show notable gains in both objective and subjective metrics. Notably, very nonstationary noise is efficiently suppressed by the suggested DNN methodology, which is a difficult task for traditional methods. Furthermore, the resulting DNN model performs well when processing noisy speech data from real world circumstances without introducing the musical artifact typically associated with traditional augmentation methods. This model was trained using artificially produced data. This emphasizes how the suggested framework for speech enhancement in a variety of environmental circumstances is both reliable and practically applicable.

Two objective measures for speech distortion and noise reduction evaluation of enhanced speech signals [6] has two new objective measuring techniques which are presented which are intended to evaluate how well speech enhancement filters recover clean speech and reduce noise. It draws attention to the shortcomings of current objective measurements in terms of offering precise indicators of noise reduction or voice distortion—two critical criteria for assessing improved speech quality. The research establishes relationships between objective and subjective measurements through evaluations of many popular speech enhancement algorithms using both the provided objective measures and subjective listening tests. These correlations show how well the suggested objective metrics work to precisely assess the quality of enhanced speech signals, so addressing a shortcoming in the current evaluation techniques by offering more detailed information about speech distortion and noise reduction capabilities.

III. SYSTEM OVERVIEW

The speech enhancement system tackles noisy audio by combining the strengths of deep learning and signal processing techniques. First, clean and noisy speech data is loaded and transformed into spectrograms, highlighting the frequency content over time. A deep learning model, typically a Convolutional Neural Network (CNN) trained on pairs of noisy and clean spectrograms, is then used to learn the complex mapping between noisy and clean representations. The CNN employs convolutional layers to extract relevant features from the spectrograms and dropout layers to prevent overfitting during training. In parallel, Non-negative Matrix Factorization (NMF) decomposes the noisy spectrogram into components representing clean speech and noise. By incorporating sparsity constraints, NMF focuses on separating the speech from the noise profile. The system strategically leverages both approaches: the noise profiles from NMF provide valuable insights into the noise characteristics, which are then fed into the CNN model. Unlike prior methods relying on intermediate predictions, this approach directly optimizes the final denoising objective, minimizing accumulated errors. This combined CNN-NMF approach refines the denoised spectrogram, potentially through a deep learning step, and finally converts it back to the audio domain for saving the enhanced audio signal. This system offers robust and efficient speech enhancement, particularly in challenging environments with complex noise patterns.

A. Model training using CNN

The speech enhancement system utilizes a Convolutional Neural Network (CNN) for model training. This process begins with creating spectrograms from noisy and clean speech data. The CNN architecture, featuring convolutional layers to extract relevant features from the spectrograms, learns the complex mapping between noisy and clean representations. Dropout layers strategically placed within the network prevent overfitting. During training, the CNN is presented with pairs of noisy and clean spectrograms. It aims to minimize the Mean Squared Error (MSE) between its predicted clean spectrogram and the actual clean one. This minimization is achieved through an optimizer like Adam, which iteratively adjusts the network's weights and biases. Training involves feeding the CNN numerous noisy-clean spectrogram pairs in batches, backpropagating the error, and adjusting the network parameters until it converges and generalizes well to unseen noisy speech. The success of this training relies on factors like the CNN architecture, training data diversity (including potentially using data augmentation), and the chosen hyperparameters. Ultimately, CNNs excel in capturing local patterns within spectrograms, making them well-suited for noise reduction and leading to cleaner and more intelligible audio after enhancement.

B. Speech Enhancement using NMF

The noisy speech signal is first transformed into a spectrogram. This visual representation highlights the distribution of frequencies over time. NMF takes the noisy spectrogram as input and decomposes it into two non-

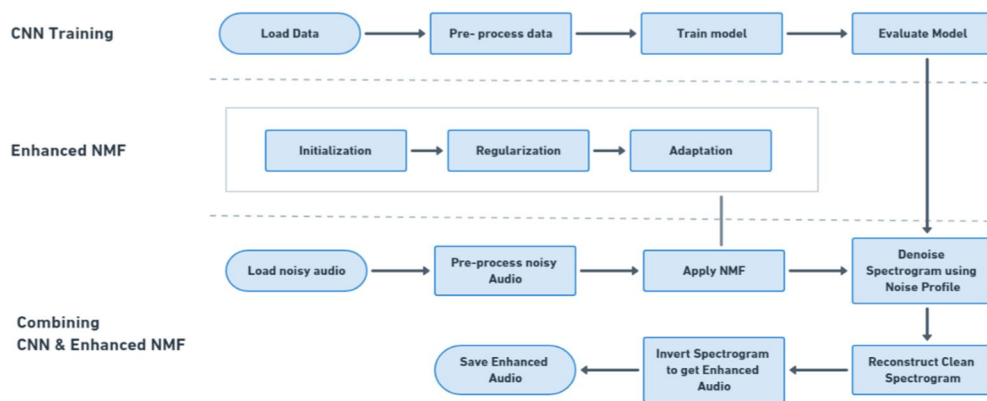


Figure 2. Flowchart of Proposed Method

negative matrices: Clean Speech Spectrogram matrix represents the underlying clean speech signal, free from noise contamination. Noise Profile Matrix captures the characteristics of the noise present in the original signal. NMF enforces a crucial constraint - all the resulting matrices must be non-negative. This reflects the inherent positivity of spectrograms, where each value represents the power (intensity) at a specific frequency. To enhance noise suppression during decomposition, sparsity constraints are often employed. These encourage the noise profile matrix to have many zero or near-zero values. This promotes a sparser representation of the noise, making it easier to distinguish from the speech components in the spectrogram. NMF achieves this by penalizing non-zero values in the noise profile matrix during the optimization process. The estimated noise profile by NMF provides valuable insights into the characteristics of the noise present in the speech signal.

Hierarchical Alternating Least Squares (HALS) is an algorithm used in NMF to factorization of a non-negative matrix into two matrices Q and R . It involves iteratively optimizing one matrix while fixing the other, often used in scenarios where sparsity and non-negativity need to be preserved.

In Non-Negative Matrix Factorization (NMF), L1-norm regularization promotes sparsity in the factor matrices (Q and R). This is achieved by adding a penalty term to the objective function during optimization. The L1-norm penalizes the magnitude of elements in Q and R , encouraging them to become exactly zero. This steers the optimization towards solutions with fewer non-zero elements, achieving sparsity. However, this comes at a cost - stronger L1 penalty might reduce reconstruction accuracy. To find the right balance, researchers tune the regularization parameters controlling the strength of this sparsity-enforcing effect.

RMSprop (Root Mean Square Prop) dynamically adjusts learning rates for each parameter. It tracks a moving average of squared gradients, essentially giving rapidly changing parameters larger updates compared to those with steadier gradients. This adaptability helps navigate the complexities of training neural networks by allowing RMSprop to tailor the learning rate for each parameter based on its recent gradient history, preventing overly aggressive updates and improving overall optimization.

C. Combining CNN with NMF

This code tackles the challenging task of speech enhancement, aiming to remove noise from a noisy audio sample. The process starts with pre-processing the audio, converting it to a spectrogram (a representation of the audio's frequency content) and normalizing it for consistency. Beforehand, a separate training phase equips a convolutional neural network (CNN) model to become an expert at differentiating noise from clean speech in spectrograms. Now, with a noisy audio loaded, its spectrogram is fed into Non-negative Matrix Factorization (NMF), a technique that decomposes the spectrogram into two key parts: a basis matrix potentially capturing the noise characteristics, and an activation matrix indicating the prominence of those noise components across frequencies. Essentially, NMF tries to separate the noisy spectrogram into components, potentially isolating the noise from the clean speech. The code then subtracts the reconstructed spectrogram obtained from NMF (thought to contain noise) from the original noisy spectrogram. This subtraction aims to remove the estimated noise component, leaving behind a spectrogram with hopefully less noise. However, there might still be some residual noise. To further refine the denoising process, the code feeds the remaining spectrogram into the trained CNN. Remember, the CNN was specifically trained to identify noise patterns in spectrograms. By analysing this spectrogram, the CNN makes a prediction, which is a denoised spectrogram (hopefully representing the clean

speech). Finally, the code retrieves the predicted spectrogram from the CNN output and inverts it back to a time-domain signal, essentially reconstructing the audio waveform. This enhanced audio waveform, hopefully free from most noise, is then saved as the final output. In essence, this approach leverages the strengths of both NMF and a CNN. NMF provides an initial separation of noise, and the trained CNN refines the denoising process using its knowledge of noise patterns learned from training data. This two-step approach has the potential to achieve robust and effective noise reduction, leading to a cleaner speech output.

IV. PERFORMANCE EVALUATION

The proposed method defines a code to load and preprocess noisy and clean speech samples, where the file paths to the noisy and clean audio files are specified along with the sampling rate set to 16 kHz. The deep learning model architecture is then defined using TensorFlow, with parameters such as input shape and number of components for Non-negative Matrix Factorization (NMF) set accordingly. The model is compiled with Adam optimizer and mean squared error loss. Training data is loaded and pre-processed, and the model is trained for 10 epochs with a batch size of 32 and a validation split of 0.1. The training data comprise clean speech and corresponding noisy speech samples recorded under different environmental settings, encompassing scenarios with background noise and reverberation. We preprocess the data by applying Short-Time Fourier Transform (STFT) to obtain magnitude spectrograms. Subsequently, a deep learning model, trained on clean and noisy speech pairs, is employed for speech enhancement. After training, the model is saved. A noisy speech signal is then loaded, and its magnitude spectrogram is calculated using Short-Time Fourier Transform (STFT). NMF is applied to estimate the noise profile, and a denoised spectrogram is obtained using the trained model. Finally, the enhanced audio signal is reconstructed and saved as a WAV file.

Later we conduct a comprehensive evaluation of a speech enhancement system utilizing various datasets with diverse conditions. To assess the system's performance, we select five input samples and obtain the enhanced spectrograms using the trained model. We then compute Signal-to-Noise Ratio (SNR), Short-Time Objective Intelligibility (STOI), and Perceptual Evaluation of Speech Quality (PESQ) metrics for both the input and output signals. The obtained results are tabulated and analysed to evaluate the effectiveness of the speech enhancement system across different conditions.

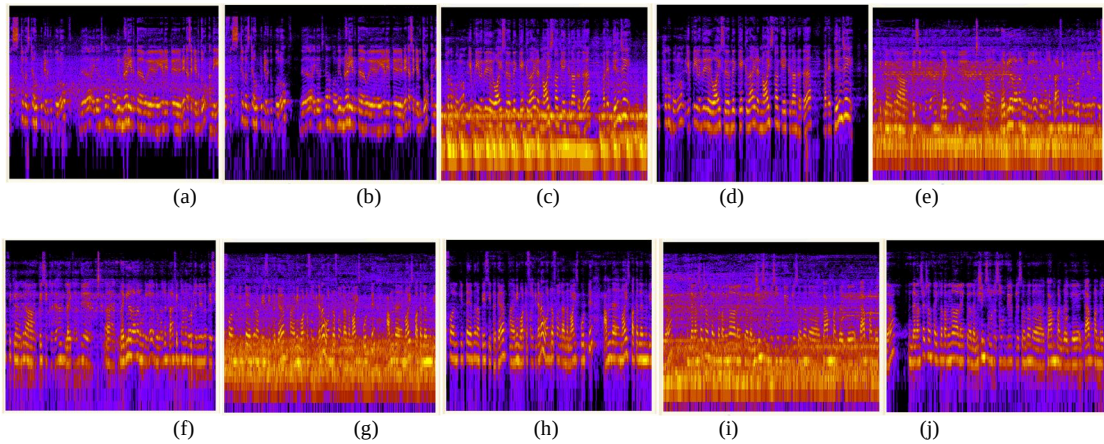


Figure 3. Spectrograms of noisy speech (a)INPUT-1, (c)INPUT-2, (e)INPUT-3, (g)INPUT-4, (i)INPUT-5 and enhanced speech (b)INPUT-1, (d)INPUT-2, (f)INPUT-3, (h)INPUT-4, (j)INPUT-5

TABLE I. EVALUATION RESULT OF PESQ

INPUT	NMF [1]	PRO
INPUT-1	1.187	4.64
INPUT-2	1.102	1.29
INPUT-3	1.101	1.54
INPUT-4	1.388	1.448
INPUT-5	1.113	4.643

TABLE II. EVALUATION RESULT OF STOI

INPUT	NMF [1]	PRO
INPUT-1	0.525	0.999
INPUT-2	0.046	0.104
INPUT-3	0.030	0.062
INPUT-4	0.028	0.078
INPUT-5	0.315	0.998

TABLE III. EVALUATION RESULT OF SNR

INPUT	NMF [1]	PRO
INPUT-1	0.103	0.128
INPUT-2	5.513	12.014
INPUT-3	3.241	5.812
INPUT-4	2.89	5.456
INPUT-5	1.064	4.982

The evaluation results across PESQ, STOI, and SNR metrics for the input and outputs obtained from the NMF-based method using mask [1] and proposed (PRO) speech enhancement method are presented in Tables 1, 2, and

3, respectively. In terms of PESQ, the proposed method consistently outperforms the NMF-based method [1] across all input references, exhibiting significantly higher scores, indicating superior perceived speech quality. Similarly, for STOI, the proposed method achieves notably higher scores compared to the NMF-based method, indicating improved speech intelligibility. Additionally, concerning SNR, the proposed method demonstrates enhanced noise reduction capabilities, as evidenced by higher SNR values across most input references compared to the NMF based method [1]. These findings suggest that the proposed method effectively addresses noise interference and improves speech quality and intelligibility, making it a promising approach for speech enhancement applications.

V. CONCLUSIONS

In conclusion, the integration of Convolutional Neural Networks (CNNs) with Non-Negative Matrix Factorization (NMF) in speech enhancement offers a promising approach that capitalizes on the strengths of both techniques. By combining CNNs for feature extraction and NMF for noise separation, the method achieves robust denoising of noisy speech signals with adaptability to diverse noise conditions, real-time processing capabilities, and seamless integration with other signal processing techniques. Despite its potential, challenges such as limited generalization, computational complexity, performance degradation in non-stationary noise, artifact generation, and integration with real-time systems remain. Addressing these limitations through innovative research in noise robustness, computational efficiency, adaptive learning, artifact mitigation, and real-time optimization is crucial for realizing the full potential of the CNN-based NMF approach in enhancing speech quality across various applications.

REFERENCES

- [1] E. M. Grais and H. Erdogan, "Single Channel Speech Music Separation Using Nonnegative Matrix Factorization and Spectral Masks," *2011 17th International Conference on Digital Signal Processing (DSP)*, Corfu, Greece, 2011, Pp. 1-6, DOI: 10.1109/Icdsp.2011.6004924
- [2] Vu, Thanh T., Benjamin Bigot, and Eng Siong Chng. "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition", 2016 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2016.
- [3] M. Wu, D. Wang and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 2002, pp. I-369-I-372, doi: 10.1109/ICASSP.2002.5743731.
- [4] Yanhui Tu, Jun Du, Yong Xu, Lirong Dai, Chin-Hui Lee. "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers", The 9th International Symposium on Chinese Spoken Language Processing, 2014
- [5] Gao, Tian, Jun Du, Li Xu, Cong Liu, Li-Rong Dai, and Chin-Hui Lee. "A unified speaker-dependent speech separation and enhancement system based on deep neural networks", 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), 2015.
- [6] Naixue Xiong, Guoxiang Tong, Wenzhong Guo, Jian Tan, Guanning Wu. "A Distributed Speech Algorithm for Large Scale Data Communication Systems", 2014 IEEE International Parallel & Distributed Processing Symposium Workshops, 2014
- [7] Yan-Hui Tu, Jun Du, Chin-Hui Lee. "DNN Training Based on Classic Gain Function for Single-channel Speech Enhancement and Recognition", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.