Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230083

Comparative analysis of voice denoising using machine learning and traditional denoising

Ke Tang

Faculty of Engineering, University of New South Wales, Sydney, New South Wales, NSW 2052, Australia

1812231121@mail.sit.edu.cn

Abstract. Noise often affects the content of an audio signal, and noise reduction techniques can help retrieve the original speech content. In recent years, AI-based noise reduction has witnessed rapid development. This article provides a brief introduction to the background and principles of several AI-based noise reduction methods. One of the mentioned methods is an end-to-end time-domain deep learning speech division algorithm, which utilizes a multi-layer CNN network framework. Due to the need for deep network architectures to extract features, it involves a higher computational load. Traditional noise reduction algorithms, on the other hand, are based on researchers' understanding of noise patterns and modeling. Traditional methods may not perform well on non-stationary noise, but they are relatively simple in terms of algorithmic implementation. Through a comparison from various perspectives, AI-based noise reduction demonstrates superior performance in known environments compared to traditional methods. However, in unknown environments, AI-based noise reduction may encounter performance anomalies. Combining AI-based and traditional noise reduction techniques can provide better stability and higher performance in certain scenarios.

Keywords: AI-based denoising, traditional denoising, robustness.

1. Introduction

Noise is actually a relative concept, and the definition of useful sound and noise can vary across different scenarios. For example, in a conversation taking place in an environment with background music, the background music is considered noise and needs to be removed using denoising techniques. However, in a live broadcast where the host is singing accompanied by background music, the background music becomes a useful signal that needs to be preserved without distortion. Therefore, it is necessary to design denoising solutions tailored to different scenarios. In a conference scenario, common noise sources include keyboard and mouse typing sounds, background discussion noise, remote reverberation introduced when the person is far away from the microphone, notification sounds, door opening/closing, and construction noise. In entertainment scenarios such as singing, when the mouth is close to the microphone, there may be plosive sounds that need to be considered as interference. In a home environment, noise may include children crying, dog barking, cat meowing, and TV noise. In outdoor live-streaming scenarios, there may be wind noise, road noise, subway noise, and so on. In gaming scenarios, a common issue is the occurrence of feedback interference when multiple players are in the same room and have their speakers and microphones turned on simultaneously. When players are

DOI: 10.54254/2755-2721/30/20230083

gaming remotely, local sound effects, finger tapping on the screen, and microphone rubbing can also be considered noise for online teammates.

Denoising technology has evolved over the years, with various algorithms and significant technical breakthroughs at each stage. Early methods included linear filtering and spectral subtraction, followed by statistical model algorithms and subspace algorithms, which are commonly referred to as traditional denoising algorithms [1-4]. AI-based denoising techniques have advanced quickly in recent years. These include deep learning algorithms based on magnitude spectrum, complex spectrum, and later, algorithms based on time-domain signals. This paper briefly introduces the principles and background of several AI-based denoising algorithms and provides examples of traditional non-AI denoising techniques, which are referred to as traditional denoising algorithms. The paper will analyze and compare AI-based denoising with traditional denoising from different perspectives.

2. The principles and background of AI-based audio denoising

The main difference between traditional denoising and AI-based denoising lies in their modeling targets. Traditional denoising models focus on the noise itself, while AI-based denoising models focus on the speech itself. As a result, they have different technical approaches. Traditional denoising involves estimating the characteristics of noise within the noisy speech, and once estimated, the noise is removed from the speech to obtain denoised audio. On the other hand, AI-based denoising follows a data-driven approach. It requires preparing a large dataset consisting of noisy and clean speech pairs and iteratively optimizing the model using loss functions defined for speech denoising and reverberation removal. Due to different technical implementations and modeling targets, traditional denoising methods often focus on eliminating stationary noise but struggle to effectively remove non-stationary noise. The main reason behind this is that traditional denoising algorithms make strict expectations when estimating noise, such as assuming that the noise follows a normal distribution. However, in real-life scenarios, we encounter mostly non-stationary noise, which refers to transient noise. For example, the characteristics of noise produced by opening and closing doors do not adhere to a normal distribution. As a result, traditional denoising algorithms struggle to achieve effective suppression when faced with non-stationary noise. On the other hand, AI-based denoising encompasses both stationary and non-stationary noise types because it operates in a data-driven manner and does not rely on any specific assumptions.

The earliest deep learning-based method for speech denoising was published jointly by the Georgia Institute of Technology and the University of Science and Technology of China in 2014 [5]. It was an algorithm based on amplitude spectrum estimation. This study was the first to demonstrate the effectiveness of machine learning in the field of voice denoising.

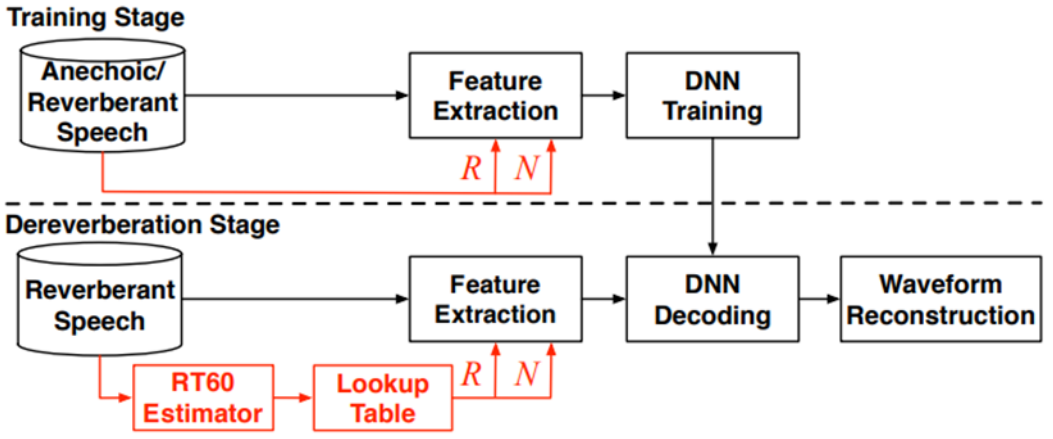

Figure 1. A block diagram of the proposed RTA-DNN dereverberation system [6].

DOI: 10.54254/2755-2721/30/20230083

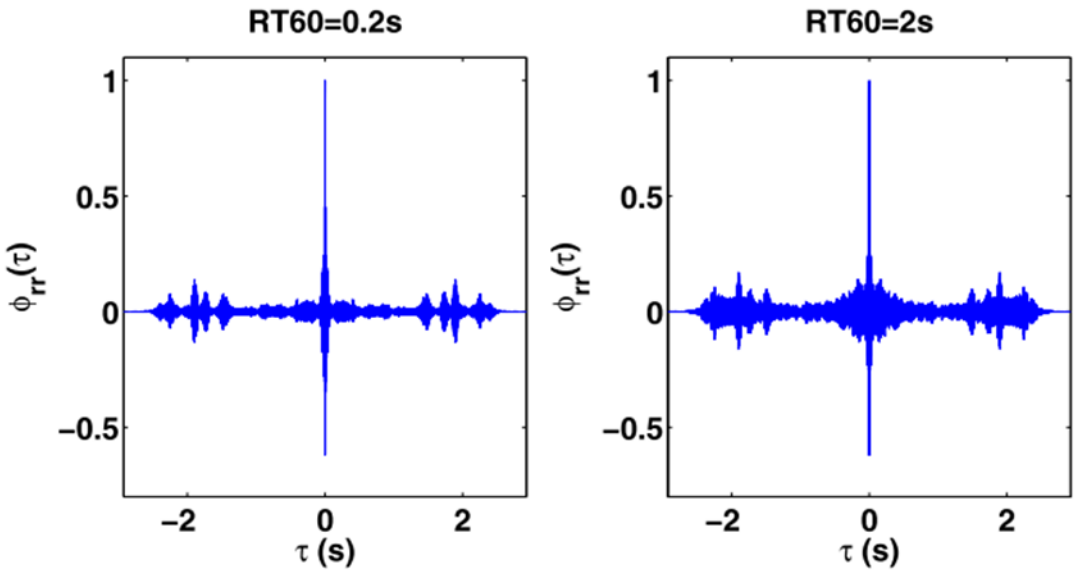

Figure 2. A TIMIT dataset utterance's temporal auto-correlation function that has been damaged by reverberation at RT60 = 0.2 and 2 s, respectively [7,8].

In the years 2015 to 2016, Ohio State University and the Georgia Institute of Technology independently published deep learning-based algorithms for speech dereverberation using amplitude spectrum estimation [6,9]. Figure 1 illustrates the process of the RTA-DNN dereverberation system. The main reason for proposing such an algorithm is that denoising and dereverberation are different tasks. To achieve optimal dereverberation, it is crucial not to directly apply denoising algorithms but to deeply understand the physical characteristics of dereverberation. Reverberation has an important physical characteristic, which is a temporal correlation. Figure 2 mentioned above depicts the autocorrelation under a strong reverberation environment with an RT60 of 2s and a weak reverberation environment with an RT60 of 0.2s. The horizontal direction indicates time, and the vertical dimension reflects autocorrelation strength. It is evident that under a strong reverberation environment, the autocorrelation is significantly stronger than in a weak reverberation environment. This indicates that when performing dereverberation, it is possible to combine the temporal correlation features. The use of the temporal correlation concept in dereverberation is the study's main accomplishment. It proposes a reverberation time-aware algorithm that is robust to different reverberation environments.

In 2016, Ohio State University published a deep learning-based speech separation algorithm using complex domain masking. This algorithm can be extended to both denoising and dereverberation tasks. In 2019, Columbia University published an end-to-end time-domain deep learning algorithm for speech division [10]. Similarly, it can be extended to denoising and dereverberation tasks, but its approach differs. Unlike the previous methods, this methodology operates directly on time-domain characteristics without first transforming them to frequency-domain information via the STFT. Indeed, this algorithm directly takes time-domain samples as input and predicts time-domain samples as output. In other words, it performs a time-domain-to-time-domain mapping. One of its key characteristics is that it can achieve very low latency, typically as low as a frame length of 5 milliseconds, which corresponds to a few tens of sample points. The entire network structure of this algorithm utilizes multiple layers of CNN (Convolutional Neural Network). CNN has the characteristic of having a relatively small number of parameters. However, because it performs a time-domain-to-time-domain mapping, it requires a deeper network structure to extract higher-dimensional and more meaningful features. As a result, the computational complexity of this algorithm is relatively high.

DOI: 10.54254/2755-2721/30/20230083

3. The Principles of Traditional Denoising

Traditional algorithms for noise reduction in speech are based on researchers' understanding and modeling of noise patterns. These algorithms include linear filtering, spectral subtraction, statistical modeling, and subspace methods [1-4]. The linear filtering method involves using filters such as high-pass filters to remove known frequency components from a signal [1]. For example, if there is interference at 50 Hz, using a high-pass filter with a cutoff frequency above 50 Hz can effectively eliminate the 50 Hz interference signal. To obtain a clean voice, spectral subtraction calculates the noise spectrum from non-speech sections and eliminates it from the noisy speech spectrum of the noisy speech [2]. Statistical modeling algorithms calculate the speech and noise components at various frequency points based on statistical methods [3]. Subspace algorithms map the noisy speech to signal subspace and noise subspace [4]. By eliminating the noise subspace components and retaining the useful signal subspace components, these algorithms estimate the clean speech signal. These traditional methods rely on manual modeling and assumptions about the noise characteristics and are often effective for stationary noise but struggle with non-stationary noise and complex acoustic environments.

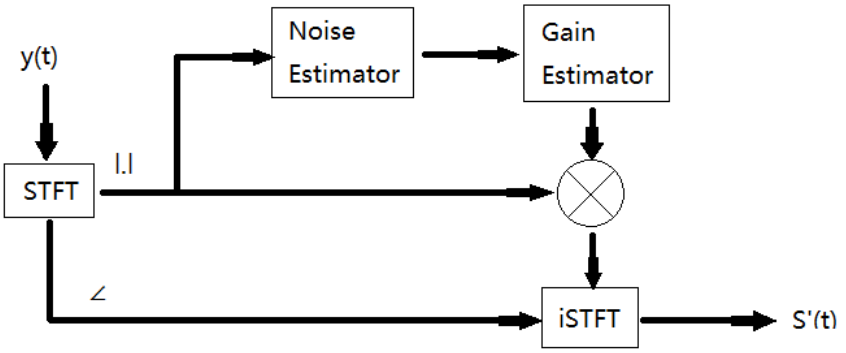

Figure 3. The principles of traditional noise reduction algorithms.

Figure 3 illustrates the typical principles of traditional denoising methods. After performing the short-time Fourier transform (STFT) on the signal y(t), the amplitude spectrum and phase spectrum of the noisy speech are obtained. In traditional methods, the amplitude spectrum of the spoken signal is typically emphasized. The amplitude spectrum information is estimated using a Noise Estimator module to estimate the noise. Then, the final gain value is calculated using a Gain Estimator. The increased speech amplitude spectrum is obtained by multiplying the noisy speech amplitude spectrum by the gain value. Next, the enhanced speech is obtained by combining the enhanced speech amplitude spectrum with the phase spectrum of the noisy speech and performing inverse STFT (iSTFT). However, due to the use of smoothing and recursive methods in the noise estimation module, it becomes challenging to accurately estimate non-stationary noise.

4. Comparison of advantages and disadvantages between traditional denoising and AI-based denoising

4.1. Noise suppression level

The various perspectives of traditional noise reduction and AI-based noise reduction are summarized in Table 1. For stationary noise, both traditional noise reduction algorithms and AI-based noise reduction algorithms can achieve good performance. However, for non-stationary noise, whether it is continuous non-stationary or transient non-stationary noise, the effectiveness of traditional methods is not very good, especially in handling transient noise where performance is the poorest. This is because non-stationary noise comes in various forms, making it difficult to summarize their patterns and challenging for traditional methods to model non-stationary noise. In this regard, AI-based noise reduction methods can

DOI: 10.54254/2755-2721/30/20230083

introduce a large amount of non-stationary noise to allow the model to learn its characteristics, thereby achieving good results.

Table 1. Hard thresholding function.

Indicator Items	Traditional noise reduction	AI-base noise reduction
Noise suppression level	High	High
Speech distortion	High	Low
Algorithm robustness	The performance remains stable across different environmental conditions.	The algorithm performs well in known environments but may exhibit performance anomalies in positional environments.
Music scene	Poor performance	Good performance
Low signal-to-noise ratio (SNR)	Poor performance	Good performance

4.2. Speech distortion

Traditional noise reduction methods struggle to accurately estimate the amount of noise, and excessive estimation can lead to speech distortion. In contrast, AI-based noise reduction algorithms primarily rely on introducing various types of noise in the training dataset to enable the model to estimate the speech and noise relatively accurately. As a result, speech distortion is generally smaller in AI-based methods.

4.3. Robustness of the algorithm

Traditional methods exhibit relatively stable performance in both new and old environments, and their algorithm complexity is not very high. Therefore, classical traditional denoising methods are still used in some scenarios. AI denoising algorithms excel in known environments, surpassing traditional methods. However, in unknown environments, AI denoising may occasionally yield suboptimal results. Nevertheless, with the advancement of AI denoising technology, its algorithm robustness is expected to improve over time.

4.4. Music scene

Using traditional noise reduction algorithms directly can cause serious damage to music signals because their noise tracking principle cannot effectively distinguish between music signals and background noise. On the other hand, AI-based noise reduction techniques can handle music noise in the model by expanding the training data, enabling them to differentiate between music and noise and achieve good results.

4.5. Low signal-to-noise ratio (SNR)

Traditional noise reduction algorithms struggle to accurately estimate the noise level, leading to higher speech distortion and more residual noise. In contrast, AI-based noise reduction can enhance the model's performance in low SNR scenarios by introducing various SNR data, including low SNR data.

5. Combining traditional denoising with AI algorithms

In a scenario where a segment of audio needs to be denoised while preserving the music content, it can be challenging for traditional noise reduction methods to achieve this effect. In such cases, incorporating AI-based denoising algorithms can be a viable solution. Figure 4 shows the principle.

DOI: 10.54254/2755-2721/30/20230083

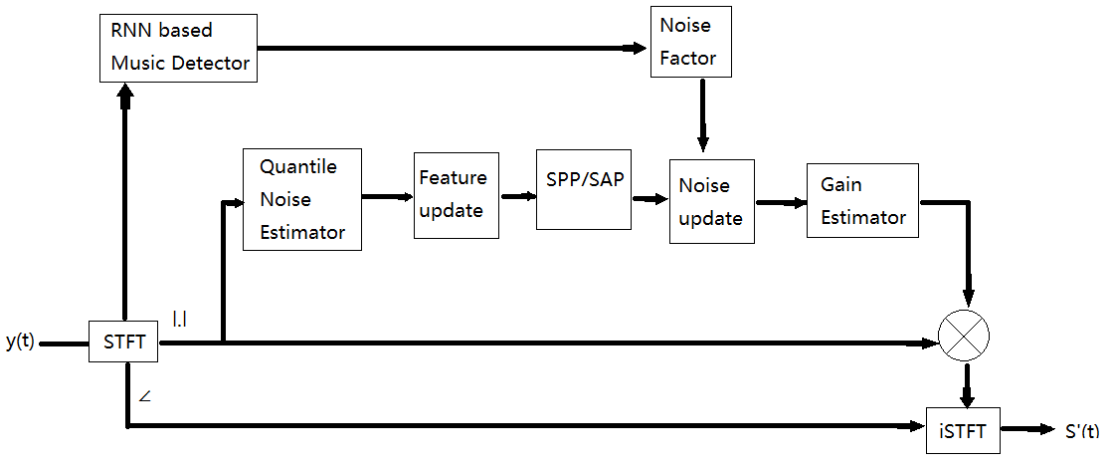

Figure 4. The principle of combining AI algorithms with traditional noise reduction.

The core idea is to combine traditional noise reduction approaches with AI music detection algorithms, leveraging the strengths of both methods. This combination aims to achieve stable performance while benefiting from the advancements of deep learning techniques, resulting in a significant improvement in algorithm performance. The blue box in the figure represents the principal diagram of the traditional noise reduction algorithm. Specifically, the noisy speech signal y(t) undergoes STFT processing to obtain the magnitude spectrum. Through quantile noise estimation, feature updating, and speech presence probability modules, the updated value of the noise is obtained. The Gain value is then calculated, and finally, the enhanced audio is obtained through iSTFT transformation. The yellow module corresponds to the AI music detection module. The input speech signal undergoes STFT processing and is fed into the RNN-based music detection module. The detection results are then used in the Noise Factor module to calculate the factor that guides the noise update. This factor aims to effectively preserve the music signal by providing an accurate estimation of the noise, thus protecting the music signal. This approach effectively improves the fidelity of the music signal without significantly increasing computational complexity. The training dataset for the network consists of speech and music signals as the target signals, while various background noise signals from multiple scenarios are used as the background noise dataset.

6. Conclusion

The article provides a brief introduction and comparison of several AI-based voice-denoising algorithms and traditional voice-denoising algorithms. Undoubtedly, AI-based noise reduction algorithms require greater computational power, but it is precisely this increased computational power that allows AI-based techniques to outperform traditional methods in known environments. However, in unknown environments, AI-based noise reduction algorithms may encounter exceptional cases. AI-based noise reduction techniques exhibit stronger adaptability and learning capabilities, enabling them to automatically learn complex audio features and effectively suppress noise. Traditional noise reduction algorithms typically rely on pre-set rules and models, making them less adaptable to different types of noise and environmental changes. In contrast, AI-based algorithms can dynamically adjust their parameters based on real-time signal characteristics, providing more accurate noise reduction. However, AI-based noise reduction algorithms have higher computational complexity, requiring more computational resources and time. This can pose limitations in real-time applications or resource-constrained devices. Additionally, AI-based algorithms are highly dependent on training data, and insufficient or unrepresentative training data may result in decreased performance.

In summary, AI-based noise reduction technology demonstrates excellent performance in known environments but may encounter exceptional cases in unknown environments. In some scenarios, a

DOI: 10.54254/2755-2721/30/20230083

combination of AI-based algorithms and traditional noise reduction techniques can achieve stable performance with performance improvements. With the continuous development of AI technology and the availability of richer datasets, the robustness of AI-based noise reduction algorithms is expected to improve, providing high-quality noise reduction experiences in various scenarios. The development of AI-based speech denoising technology has been rapidly advancing, and the performance of everyday hardware is also improving. With the progress in algorithms and hardware, the handling of non-stationary noise in unknown environments is expected to become more reliable. This advancement will greatly enhance the quality of human voice in scenarios such as voice calls and online streaming, leading to a significant improvement in overall voice quality.

References

[1] Chen J D, Benesty J and Huang Y T 2007 On the optimal linear filtering techniques for noise reduction *Speech Commun.* **49(4)** p 305-316 doi: 10.1016/j.specom.2007.02.002.

[2] Boll S 1979 Suppression of Acoustic Noise in Speech Using Spectral Subtraction *IEEE Trans. Acoust., Speech Signal Process.* **27(2)** p 113-120 doi: 10.1109/TASSP.1979.1163209.

[3] Xia B Y and Bao C C 2013 Speech Enhancement with Weighted Denoising Auto-Encoder *INTERSPEECH* p 3444-3448

[4] Hermus K, Wambacq P and Van hamme H 2007 A Review of Signal Subspace Speech Enhancement and Its Application to Noise Robust Speech Recognition *EURASIP J. Adv. Signal Process.* p 15 doi: 10.1155/2007/45821.

[5] Xu Y, Du J, Dai L D and Lee C H 2015 A Regression Approach to Speech Enhancement Based on Deep Neural Networks *IEEE/ACM Trans. Audio Speech Lang. Process.* **23(1)** p 7-19 doi: 10.1109/TASLP.2014.2364452.

[6] Wu B, Li K H, Yang M L and Lee C H 2016 A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks *IEEE/ACM Trans. Audio Speech Lang Process.* **25(1)** p 102-111 doi: 10.1109/TASLP.2016.2623559.

[7] Garofolo J S 1988 DARPA TIMIT acoustic phonetic continuous speech corpus CD-ROM *Technical report. NIST*

[8] Wu B, Yang M L, Li K H, Zhen H, Siniscalchi S M, Wang T and Lee C H 2017 A reverberation-time-aware DNN approach leveraging spatial information for microphone array dereverberation *EURASIP J. Adv. Signal Process.* doi: 10.1186/s13634-017-0516-6.

[9] Han K, Wang Y, Wang D L, Woods W S, Merks I and Zhang T 2015 Learning Spectral Mapping for Speech Dereverberation and Denoising *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.* **23(6)** p 982-992 doi: 10.1109/TASLP.2015.2416653.

[10] Luo Y and Mesgarani N 2019 Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation *IEEE/ACM Trans. Audio Speech Lang. Process.* **27(8)** p 1256-1266 doi: 10.1109/TASLP.2019.2915167.