Research on Speech Enhancement Algorithms for Wearable Devices

Jingze Fu[1], Fulang Sun[2(✉)], Hao Yin[3], Bingyu Shen[4], Zihao Wang[5], and Hengfan Zhang[6]

[1] College of Egr Info & AppSci, Northern Arizona University, Flagstaff, USA
[2] College of Materials Science and Engineering, Northeastern University, Shenyang, China
20192934@stu.neu.edu.cn
[3] Sino-British College, University of Shanghai for Science and Technology, Shanghai, China
[4] College of Telecommunications and Information Engineering, Xi'an Jiaotong University, Xi'an, China
[5] College of Liberal Arts, University of Texas at Austin, Austin, USA
[6] School Telecommunication Engineering, Xidian University, Xi'an, China

Abstract. Speech is the primary information carrier in human communication and the interference could be caused by environmental noise. The speech enhancement algorithm is an effective method to reduce noise and improve the subjective feeling of the human ear. The existing speech enhancement algorithms can be divided into two categories: single-channel methods and dual-channel methods. This research analyzed the two main digital signal processing topics for speech enhancement algorithms. This included the traditional single channel speech enhancement algorithm and the dual-channel speech enhancement algorithm. The traditional single channel speech enhancement algorithm was further divided into two categories: traditional enhancement algorithms based on digital signal processing and learning enhancement algorithms based on data-driven. The traditional algorithms have limitations in processing non-stationary noise, whereas deep learning has significantly improved handling non-stationary noise. The dual-channel speech enhancement algorithm adds a second microphone to collect noise, providing more real noise for enhancement. This research briefly described the main dual-channel speech enhancement algorithms, including adaptive noise cancelling and first-order difference microphone. In conclusion, the research found that deep learning has shown significant improvement in handling non-stationary noise in single-channel speech enhancement algorithms, and dual-channel speech enhancement algorithms have the potential to provide more real noise for enhancement.

Keywords: enhancement algorithms · deep learning · adaptive noise cancelling · improved adaptive null-forming · LMS

These authors contributed equally to this work and should be considered co-first authors.

N. Akhtar et al. (Eds.): PMIS 2023, AHIS 8, pp. 967–973, 2023.
https://doi.org/10.2991/978-94-6463-200-2_101

1 Introduction

Speech is the most important information carrier used by human beings when communicating with each other. In the actual environment, the sound signal will always be interfered by the environmental noise during the transmission process, such as the noise of the surrounding environment, the noise of the transmission medium and the electrical noise inside the communication equipment [1]. The "noise" here means that it enters all signals except the required sound signal. The signal input can be a narrow or wide-range noise, white noise or color noise, sound or electrical noise, or even another independent sound. Noise reduces the signal-to-noise ratio and intelligibility of speech, which will make the speech processing system unable to work normally in serious cases.

The speech enhancement algorithm is an effective means to reduce noise [1]. It extracts as pure original speech as possible from the noisy speech signal, so as to achieve the purpose of reducing noise interference [2]. And the final measure of speech enhancement effect is the subjective feeling of the human ear [3]. Therefore, human ear perception can be used to reduce the computational cost in speech enhancement. But it is difficult to achieve objective quantification based on human subjective judgment at a low cost because strict standards and many testers are needed [4]. Speech enhancement algorithms can be roughly divided into two categories, namely, single-channel and dual-channel methods or microphone array methods. The single channel methods part can be divided into methods based on digital signal processing and methods based on learning enhancement.

Wearable devices are a portable device that are directly worn on the body or integrated into the user's clothing or accessories. For users, the effect of interaction determines the quality of user experience to a great extent. For some wearable devices such as wireless headset and headworn display, the sound quality is an extremely crucial standard to assess user experience [2]. Therefore, optimizing wearable devices through speech enhancement algorithms will greatly improve the sound quality, thus improving the user experience of wearable devices. This research focused on two main topics about digital signal processing for speech enhancement algorithm, including single-channel speech enhancement algorithm and dual-channel speech enhancement algorithm [3]. Firstly, the analysis of traditional single channel algorithm based on signal processing and methods based on enhancement algorithms are shown. Secondly, the dual-channel speech enhancement algorithm, or microphone array methods, is briefly described, including the basic principle and realization with some mathematical proof.

2 Single-Channel Speech Enhancement (SCSE) Algorithms

SCSE is a widely studied topic in speech signal processing, mainly used as a front-end denoising module in improving sound quality, speech communication, auxiliary hearing, speech recognition, etc. The definition of the SCSE problem mainly includes two aspects. On the one hand, the input signal is a noisy voice signal with only one channel. On the other hand, the processing goal is to enhance speech and reduce noise, which is also equivalent to separating "voice" and "non voice" signals. What needs to be different from SCSE is multi-channel voice signal processing, which belongs to array signal processing.

Another aspect that is easy to confuse is to distinguish/separate the mixed voice of different people (that is, the cocktail party problem), which belongs to the category of blind source separation. SCSE algorithms are divided into two categories: traditional enhancement algorithms based on digital signal processing and learning enhancement algorithms based on data-driven. For traditional speech enhancement, it is based on time domain analysis or frequency domain analysis. The frequency domain is mainly designed based on the gain function. However, different algorithms use different strategies when calculating the gain function, which are divided into three categories: Spectral-Subtractive Algorithms, Statistical model-based methods, and Subspace Algorithms [5].

Spectral subtraction is the oldest algorithm. This algorithm requires known noise spectrum, and the enhancement effect will be significantly reduced when dealing with non-stationary noise. Later, as a newer theory, the statistical model-based speech enhancement methods predominated the field at the late 20th century, with the primary aim of improving the quality (an objective empirical expression of the level of enjoyment or endeavor required for listeners to interpret the information) and intelligibility (an objective measurement of the extent to which listeners can extract useful information from a given signal, regardless of its level of noise) of degraded speech signals, with the objective of improving the speech encoder's reaction to input noise and improve the speech recognition system's robustness to input noise. The statistical model-based speech enhancement concentrates on three main topics. The signal estimation from a given sample function of noisy speech, the signal encoding with only noisy speech, and the identification of noisy speech signals in human-machine communication. The applications of these methods span a wide range of time, from improving the performance of cellular radiotelephone systems affected by channel noise, public telephones located in noisy environments, aircraft-to-ground telecommunication systems where cockpit noise is disruptive to information, to teleconferencing systems in which noise from one place would be broadcast to everyone.

With the introduction of HMM in the mid-1970s and VQ in the early 1980s, two significant breakthroughs were observed in the field of statistical modeling of speech signals. Two model-based approaches were devised: model-based estimation, a method for designing filters for noisy signals, and model-based synthesis, a model for synthesizing the desired signal. The difference is that when a clear speech signal is provided as input, the estimation method outputs a signal as clear as the input as the filter is transparent, while the synthesis model is limited by the performance of a linear predictive vocoder and a harmonic zero-phase sine wave encoder to output an approximation of that signal. When speech recognition is implemented into a practical context, reducing the ambient noise to promote the identification rate is necessary. In order to improve the signal-to-noise ratio at the input of automatic speech recognition, a variety of speech enhancement methods have been discussed. However, since the typology of noise strongly differs with respect to the surroundings, no one speech enhancement technique can encompass the entire spectrum of noise. Thusly, in the end of the cent, under the essential assumption that each short-time speech vector can be written as a linear combination of linearly independent basis functions, the rationale for subspace speech enhancement is rooted in the observation that a continuous speech vector will occupy an x-dimensional subspace of a y-dimensional subspace of a z-dimensional Euclidean space. And with this basis,

the subspace speech enhancement system partitions the input speech signal into a y-dimensional subspace containing a noise-interfering speech signal and a z-dimensional subspace with only the noise. Subsequently, the subspace speech enhancement system is thus concluded with the removal of the noise-only subspace and the selective removal of the noise component in the mixed subspace.

There are common defects for traditional methods. They all need prior information, and then they all need to be based on specific assumptions, such as noise has a certain degree of stability, clean speech and noise are not related. The performance of these traditional algorithms will be significantly reduced beyond specific assumptions. For example, in the case of non-stationary noise and low signal-to-noise ratio, the enhancement effect of traditional algorithms is significantly reduced. In addition to the traditional algorithms, the learning class enhancement algorithm based on test data-driving reduces the dependence on the preconditions of the traditional algorithms. At present, there are two main types, sparse representation model and deep learning model. In recent years, with the rapid development of artificial intelligence, SCSE based on deep learning has made significant progress in processing non-stationary noise. There are three methods SCSE based on parameters, SCSE based on generation countermeasure mechanism, and SCSE based on weak supervision [6]. The first two methods need to be trained with parallel data in pairs. The speech enhancement method based on weak supervision can use non-parallel data for training, thus reducing the requirements for data sets. In conclusion, compared with the traditional SCSE algorithm, the SCSE ability of deep learning under non-stationary noise conditions is significantly improved. In addition, the deep learning model can also be combined with the traditional SCSE algorithm, for example, the Deep Neural Network (DNN) can be combined with a less aggressive Wiener filter Weiner filter. Wiener filtering is used as a soft mask to DNN outputs [7]. The effect of this combination is improved compared with the single SCSE algorithm.

Over the last decade, the sparse representation with dictionary learning methods were widely utilized in the area of signal processing. The sparse representation of a voice signal involves selecting a linear combination of a few elements in an excessively comprehensive dictionary of atoms to approximate the signal. The dictionary learning is designed to locate the best set of elements that well captures the features of these voice signals. On the basis of the joint dictionary, the speech dictionary and the noise dictionary can be learned individually with the training samples of the speech signal and the noise signal lying on the assumption that the size of the noisy speech is a linear sum of the noise size and the speech size, disregarding the effect of phase, which is then enhanced with the coherence criterion algorithm of least angle regression (LARS) to represent the noise signal sparsely on a composite dictionary. After that, according to the coherence between the signal and the dictionary, the corresponding dictionaries can represent the expected speech and the noise. Eventually, the enhanced speech is then separated from the corrupted speech signal by extracting the clean speech. The applications of sparse representation in speech signal enhancement are mainly concentrated in speech activity monitoring, pitch estimation, speaker identification and speech recognition.

3 Dual-Channel Speech Enhancement (DSE) Algorithms

In the process of speech enhancement, the key is to obtain noise. In single-channel speech enhancement, the noise is estimated from the noisy speech signal. However, the estimation algorithm is complicated and the estimated noise is always different from the real noise, which limits the enhancement effect to a certain extent. Dual-channel speech enhancement, based on single-channel speech enhancement, add a microphone to collect noise to obtain more real noise. There are few algorithms developed for dual microphones, mainly the adaptive noise canceling (ANC), the first-order difference microphone (FDM) and the improved adaptive null-forming (ANF) based on FDM. ANC and FDM are two typical dual channel speech enhancement algorithms.

ANC is the most primitive and simple dual channel speech enhancement algorithm. Noise to noise elimination method adopts two microphones, one microphone to collect noisy speech, the other to collect noise signal, with noise signal minus noise signal (reduction operation is generally carried out in the frequency domain, if the collected noise and noise in the noise signal is similar enough, even can be directly subtracted in the time domain), to get the voice signal. In practice, the positions of the two microphones are different, and there is delay and attenuation to different degrees between the two signals, but the noise components in the two channels are from the same noise source, and there is still a strong correlation between the noise. Based on the correlation, adaptive algorithm combined with subtraction operation is used to realize adaptive noise cancellation. In signal processing, noise is often non-stationary and changes with time. To solve the problem of signal extraction in noise background, the adaptive algorithm is mainly based on the least mean square (LMS), recursive least squares (RLS) and square root adaptive filtering (QR_RLS) three noise elimination algorithms. These algorithms can effectively suppress interference from high background noise and extract useful signals, showing good convergence performance.

FDM is similar to noise cancellation. The difference is that the difference calculation of the two signals is carried out according to the microphone's position, and the output of the noisy speech signal and the noise signal is more accurate. Adaptive noise cancellation is carried out to obtain the enhanced speech. Concrete steps: First, a first-order differential microphone array model is constructed, and the signal received by two microphones in the array is analyzed with mathematical expressions. The time-domain difference signal is obtained by the difference between the two microphones in the array, and the amplitude-frequency characteristic curve of the difference signal is analyzed. An inverse comb filter with amplitude-frequency characteristics is designed to completely inverter the envelope of the amplitude-frequency characteristic curve of the difference signal. Finally, the difference signal is filtered by the filter, and the amplitude-frequency characteristics of the original signal are restored. Then, the estimated time-domain signal is obtained by combining the phase information to restore the signal.

Constrained LMS algorithm is a straightforward stochastic gradient descent algorithm that requires prior knowledge of the desired frequency band and signal arrival orientation. The method gradually picks up statistics of noise coming from directions other than the observed direction during the adaptive process. By carefully choosing the frequency response characteristic in that direction or using external means, noise from the detected direction can be filtered away. The approach is appropriate for array

processing issues with electromagnetic, sonar, and geological antenna arrays. The algorithm's ability to self-correct allows it to operate in digital computer implementations for arbitrarily long times without departing from its limits as a result of cumulative rounding or truncation errors. This is a significant advantage [8].

Phase time-frequency masking-based microphone array noise reduction algorithm consists of four steps: time-frequency mask estimation, beamforming, noise power spectral density estimation, and single-channel Wiener filtering. While estimating time-frequency masks, the phase time-frequency mask is estimated by the audio signal from several microphones, and the reference microphone estimates the speech existence probability. These estimates are combined to create the final time-frequency mask. The beamforming process involves estimating the guiding vector and noise covariance matrix through the time-frequency mask, beamforming the target signal based on the least variance distortion-free response, and removing the interference signal and background noise. The phase mask maximum likelihood of the microphone speech with probability correction is used in the estimation component of the interference noise power spectrum density to determine the residual interference noise power spectrum density after the previous beamforming stage. The preceding step's estimation of the interference noise power spectrum density is employed in the single-channel Wiener filtering stage to further reduce any remaining interference noise in the beamforming output and provide enhanced speech [9]. Another algorithm uses the beamforming and post-filtering noise removal method using signal power spectrum density. The algorithm's overall structure is divided into four sections: (1) A generalized sidelobe removal microphone array technique based on signal power spectral density. The interference noise in noisy speech is removed more effectively, and the harm to speech is decreased, by applying the Signal-Power-Ratio and two decision procedures. (2) The deep neural network approach is used to train the masking value under the assumption of a multi-objective function since it is more effective than the single-objective function method. (3) A decision scheme is proposed and the noise smoothing factor is updated based on the masking values obtained in Part II. (4) Post filter, the proposed method updates the noise smoothing factor and further accurately estimates the noise power spectral density. Besides, it will better process the noisy speech in the single channel and remove some non-stationary noise and residual noise in the speech after beamforming [10].

4 Conclusion

In conclusion, this research paper analyzed the development of speech enhancement algorithms for wearable devices. Speech is the primary information carrier in human communication, but it is easily interfered with by environmental noise. The speech enhancement algorithm is an effective method to reduce noise and improve the subjective feeling of the human ear. The existing speech enhancement algorithms can be divided into two categories: single-channel methods and dual-channel methods. Single-channel methods include traditional enhancement algorithms based on digital signal processing and learning enhancement algorithms based on data-driven. However, traditional algorithms have limitations in processing non-stationary noise, whereas deep learning has significantly improved handling non-stationary noise. The dual-channel speech enhancement algorithm adds a second microphone to collect noise, providing more real noise

for enhancement. This research found that deep learning has shown significant improvement in handling non-stationary noise in single-channel speech enhancement algorithms, and dual-channel speech enhancement algorithms have the potential to provide more real noise for enhancement. In summary, optimizing wearable devices through speech enhancement algorithms will greatly improve the sound quality and user experience of wearable devices.

Acknowledgment. All authors contributed equally to this work and should be considered co-first authors.

References

1. X. Yang and H. Chi, Speech signal digital processing, Electronic Industry Press, 1995.
2. S. Huang, S. Huang and S. Liang, "A Review of Speech Enhancement Algorithms," Computer and Modernization, vol.139, no.03, pp.16-20, March 2007
3. J. Wang, F. Fu and Y. Zhang, "A Review of Speech Enhancement Algorithms," Acoustics and Electronic Engineering, no.01, pp.22-26, March 2005
4. Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," IEEE Trans. Audio Speech Lang. Process., Volume 16, Issue 1, pp 229-238, January 2008.
5. Loizou, Philipos C. Speech enhancement: theory and practice. CRC press, 2007.
6. Xiongwei Zhang, et al. "Methods of Deep Learning in Monaural Speech Enhancement: State of Art and Prospects" Journal of Army Engineering University of PLA 1.05(2022):1-12.
7. N. Saleem, M. I. Khattak, M. Y. Ali, and M. Shafi, "Deep Neural Network for Supervised Single-Channel Speech Enhancement," Archives of Acoustics, vol. 44, no. 1, Art. no. 1, Jan. 2019.
8. O. L. Frost, "An algorithm for linearly constrained adaptive array processing," Proceedings of the IEEE, vol. 60, no. 8, pp. 926–935, Aug. 1972.
9. L. He, Y. Zhou, H. Liu. "A microphone array noise cancellation method using phase time-frequency masking." Signal processing 34.12(2018):1490-1498.
10. F. Ni, Y. Zhou, H. Liu. "Microphone array noise elimination method using signal power spectral density." Signal Processing 36.03(2020):373-381.