Proceedings of CONF-SPML 2026 Symposium: The Artificial Intelligence Tools & Applications
DOI: 10.54254/2755-2721/2026.CH32169

A Review on Intelligent Enhancement and Recognition Techniques of Speech and Image Multimodal Signals for Complex Environments

Jiayuan Luo

School of Computer and Communication Engineering, Northeastern University, Shenyang, China
972709632@qq.com

Abstract. In recent years, how to intelligently enhance as well as accurately recognize speech and image signals in complex scene environments is a major problem facing the field of Artificial Intelligence Single-modal signals can easily be affected by factors such as noise or occlusion leading to a significant drop in their performance, which becomes a key obstacle to the further development of the field. The multimodal fusion problem involves many disciplines, and its complexity and uncertainty bring great challenges. In this paper, it systematically reviews the research progress of intelligent enhancement and recognition techniques for speech and image multimodal signals oriented to complex environments, summarize the challenges faced by the current techniques, such as data alignment and modal missing, and provide an outlook on the future research direction. Based on the analysis of the basic principles and technical framework of multimodal signal processing, the key issues of related technologies such as intelligent enhancement of multimodal signals and multimodal feature fusion and recognition methods are elaborated.This article sorts out the literature and core challenges of multimodal signal processing in complex environments, provides new ideas for breaking through the complex interference bottlenecks of unimodal signal processing, and offers technical guidance for academic research and industrial implementation in this field.

Keywords: Multimodal signal processing, Speech enhancement, Image enhancement, Deep learning

1. Introduction

Accompanied by the rapid development of artificial intelligence technology, the processing of speech signals and image signals has been widely used in scenarios such as smart home, automatic driving, and human-computer interaction. However, practical application environments often have a variety of complex interference terms, such as background noise, reverberation, occlusion, light changes, etc., which can bring different degrees of impact on single-modal signal processing. On the one hand, speech signals are susceptible to environmental noise interference, especially under low signal-to-noise ratio conditions, and the speech enhancement and recognition performance decreases significantly, and on the other hand, image signals are sensitive to factors such as illumination

DOI: 10.54254/2755-2721/2026.CH32169

changes, occlusion, blurring, and so on, affecting the subsequent recognition and analysis. Traditional unimodal processing methods perform poorly under the above disturbances and are difficult to meet the demands of practical applications, so signal processing in complex environments faces many challenges.

Multimodal signal processing provides new ideas for improving signal processing in complex environments by integrating complementary information from different modalities such as speech and image. Although multimodal signal processing has made significant progress, there are still many problems that need to be solved. First, different modal signals differ in sampling rate, feature expression, etc. The key challenge is how to realize effective modal alignment and feature fusion. Second, modal absence or uneven quality often occurs in complex environments, which requires the system to have good fault tolerance. Third, multimodal models usually have high computational complexity and are difficult to run in real time on resource-constrained devices. This paper summarizes the relevant literature to provide scholars with a relatively clear technical development line to promote the development and application of multimodal signal processing techniques in complex environments. Section two outlines deep learning applications, related findings and multimodal fusion benefits for speech and image enhancement in intelligent multimodal signal processing. Section three outlines key methods and findings for multimodal feature fusion and recognition.

2. Intelligent multimodal signal enhancement

2.1. Speech enhancement technology

Targeted speech recovery for noisy speech in complex environments is known as speech enhancement, which is an effective means to improve speech quality as well as enhance speech recognition rate. Deep learning-based methods achieve technological breakthroughs through end-to-end modeling, such as the dual-stream gated audio-visual fusion architecture proposed by Peng et al. , which can make good use of dynamic and static audio-visual information, and achieve good speech enhancement under low signal-to-noise conditions [1]; Li et al. use generative adversarial networks combined with the attention mechanism and mask learning, which can provide good speech enhancement under low signal-to-noise conditions [2]; and because deep learning models have better nonlinear mapping characteristics, they are effective means to improve speech quality as well as enhance the speech recognition rate. speech enhancement: at the same time, due to the better nonlinear mapping characteristics of the deep learning model, the separation of speech and noise can be carried out better. In addition, the introduction of multimodality can provide more visual cues, for example, the target speaker can be accurately localized by lip movement information in a multi-person speaking scene, thus eliminating redundant speech information such as background noise and interference. Pan proposed an EMD-based forward and backward filtering speech enhancement algorithm using empirical modal decomposition to address non-smooth noise [3]. The adaptive multimodal flow shape learning method proposed by Dai is suitable for denoising rotating machinery signals, and this idea can also be referred to for speech enhancement [4].

2.2. Image enhancement technology

Image enhancement refers to the process of improving the visual effect of images in poor environments such as low light, blurring, degradation, etc., and is an important part of image processing work in complex environments. Deep learning has led to a new breakthrough in image enhancement, and image denoising, deblurring, and super-resolution reconstruction based on convolutional neural networks have been realized. Wang et al. systematically combed and summarized the current mainstream image defogging algorithms, specifically analyzed the advantages and shortcomings of various defogging methods and combined their use in different conditions [5]. Li proposed a multimodal EEG-fNIRS signal fusion method based on wavelet transform, which can effectively fuse EEG signals with functional brain imaging signals, and at the same time is suitable for research in image enhancement [6]. Multimodal image enhancement utilizes auxiliary information such as audio information to enhance visual perception, its advantages compared with other image enhancement methods can be seen from Table 1. For example, in video sequences, audio information can provide semantic contextual information to the scene, which in turn guides image enhancement; in low-light situations, the combination of audio event detection can accurately localize regions of interest, thus achieving local enhancement of key regions. As mentioned above, Zhang proposed a modal parameter identification method using an improved particle swarm optimization algorithm for identification, which can be fully used in the research of multimodal data co-processing [7].

Table 1. Comparison of multimodal signal enhancement techniques

Type of Technology	Representation Method	Advantage	Limitation	Applicable Scenarios
Conventional single-mode enhancement	Spectral subtraction, Wiener filtering	Simple calculation and good real-time performance	Poor effect of non-stationary noise	Noise stable environment
Deep Learning Single Modal Enhancement	Method based on CNN/LSTM	Strong nonlinear mapping ability	Great dependence on training data	Large data scenario
Multimodal fusion enhancement	Audio-visual fusion method	Strong robustness and anti-jamming ability	Multimodal data alignment required	Complex and changeable environment

3. Multimodal feature fusion and recognition

3.1. Feature extraction and representation learning

Effective feature extraction is the basis of multimodal recognition. Speech feature extraction focuses on time-frequency domain features, while image feature extraction uses deep convolutional neural networks (VGG, ResNet) to obtain spatial features. Representation learning refers to mapping different modal data into the same semantic space to solve the problem of modal heterogeneity. Multimodal representation learning includes two strategies, joint representation and collaborative representation, in which joint representation means that multimodal data are mapped into the same space, suitable for the scenario that all modalities exist; while collaborative representation is to ensure the independence of each of the multimodal modalities, but to make sure that these modalities maintain a certain connection with each other, and is applicable to scenarios in which a few modalities do not exist. He et al. summarized the multimodal fusion methods for deep learning and analyzed the advantages and disadvantages of different representation learning methods [8].

3.2. Multimodal fusion strategy

Multimodal fusion can effectively improve the recognition performance, which can be categorized into data-level, feature-level and decision-level fusion according to the fusion levels, the differences among different fusion levels can be seen from Table 2. Among them, data-level fusion is computationally efficient but requires strict alignment; feature-level fusion is good in encoder flexibility; and decision-level fusion is robust but ignores early interactions. Li et al. proposed a recognition algorithm based on matching layer fusion, which effectively improves the recognition rate through adaptive weighted fusion [9]. The vehicle network beam assignment method based on multimodal feature fusion studied by Nie et al. realizes the effective fusion of multimodal features through the attention mechanism [10]. Cheng and Zhang et al. explored the multimodal fusion strategy from the perspective of emotion recognition and modal parameter recognition, respectively [7,11]. Liang optimized the system architecture and algorithms to improve the practicality in the design of Chinese speech recognition system in complex environments [12].

Table 2. Comparison of multimodal fusion levels and methods

Fusion Level	Fusion Mode	Typical Method	Merit	Disadvantage
Data level fusion	Early fusion	Feature splicing	Information remains intact	Strict modal alignment required
Feature level fusion	Mid-stage fusion	Attention mechanism	High flexibility	Complex model
Decision-level fusion	late fusion	Weighted voting	Strong robustness	Ignore early interactions

4. Current technical challenges

Multimodal signal processing faces several problems in practical applications.

First, modal alignment, different modal signals need to be precisely synchronized in time and space. The difference in sampling rates between audio and video signals leads to alignment difficulties and affects the fusion effect. The application of particle swarm algorithm in modal parameter identification studied by Zhang et al. provides a new idea for multimodal alignment [7].

Second, modal missing, a modal data quality is often poor or completely missing in complex environments, which requires the system to have good fault tolerance. The adaptive multimodal manifold learning method studied by Dai, which deals with the modal missing problem through manifold learning, has important reference value [4].

Third, the high computational complexity restricts the application of multimodal techniques on resource-constrained devices. The beam assignment method for vehicular networks studied by Nie et al. optimizes the computational efficiency while guaranteeing the performance, which provides a reference for solving the complexity problem [10].

5. Conclusion

This paper systematically reviews the research progress of intelligent enhancement and recognition of speech and image multimodal signals in complex environments. Firstly, multimodal technology can significantly improve the robustness of signal processing in complex environments by fusing the complementary information of speech and image. The innovation is mainly reflected in the fact that the multi-modal enhancement technology breaks through the limitations of the single mode, and the fusion methods such as the attention mechanism realize the dynamic adaptation between the modes. The combination of deep learning and multi-modal fusion provides an effective way to solve signal processing problems in complex environments. Secondly, multi-modal signal processing technology has shown wide application prospects in smart home, autonomous driving, medical health, and other fields. Through multi-modal information collaboration, the system can understand the environment more comprehensively and make accurate decisions. Future research should focus on lightweight models, self-supervised learning, causal inference, and other directions to further improve technical performance and application scope. The intelligent enhancement and recognition technology of multi-modal signals has important research value and wide application prospects. With the advancement of technology and the growth of application requirements, multimodal technology will be paid more and more attention to in the perception and understanding of complex environments.

References

[1] Peng Minxuan, Liang Yan. (2025). Multimodal Speech Enhancement Based on Dual-Stream Gated Audio-Visual Fusion. Computer System Applications, 34 (11), 127-138.Retrieved from https: //doi.org/10.15888/j.cnki.csa.009996.

[2] Li Tongyan, Pei Haoyan, Pei Yan, Chen Xu, Wang Tao. (2025). GAN Speech Enhancement Algorithm Based on Attention Mechanism and Mask Learning. Journal of Chengdu University of Information Engineering, 40 (02), 137-142.Retrieved from https: //doi.org/10.16836/j.cnki.jcuit.2025.02.003.

[3] Pan Qing, Ran Fuxing, Li Yakun. (2018). EMD-based pre-and post-filtering speech enhancement algorithm. Journal of Henan Normal University (Natural Science Edition), 46 (03), 33-39. Retrieved from https: //doi.org/10.16366/j.cnki.1000-2367.2018.03.006.

[4] Dai Lei. (2021). Research on Rotating Machinery Signal Denoising and Feature Identification Methods Based on Adaptive Multimodal Manifold Learning (Master's Thesis, Chongqing University).Retrieved from https: //doi.org/10.27670/d.cnki.gcqdu.2021.001012.

[5] Wang Daolei, Zhang Tianyu. (2020). Review and Analysis of Image Defogging Algorithms. Journal of Graphics, 41 (06), 861-870.

[6] Li Lizhu, Meng Ming, Gao Yunyuan, Ma Yuliang. (2023). EEG-fNIRS Multimodal Data Fusion Method Based on Wavelet Transform. Journal of Sensing Technology, 36 (07), 1064-1072.

[7] Zhang Jindong, Guo Xiaonong, Luo Xiaoqun, Zhang Yujian, Xu Hongjun. (2022). Improved modal parameter identification method based on particle swarm optimization algorithm. Vibration and Shock, 41 (02), 255-264.Retrieved from https: //doi.org/10.13465/j.cnki.jvs.2022.02.031.

[8] He Jun, Zhang Caiqing, Li Xiaozhen, Zhang Dehai. (2020). Review of Multimodal Fusion Technology for Deep Learning. Computer Engineering, 46 (05), 1-11. Retrieved from https: //doi.org/10.19678/j.issn.1000-3428.0057370.

[9] Li Aomei, Hu Zhenghao, Zhou Chuanchuan. (2020). Research and Implementation of Recognition Algorithm Based on Matching Layer Fusion. Electronic Technology Application, 46 (07), 57-59.Retrieved from https: //doi.org/10.16157/j.issn.0258-7998.191426.

[10] Nie Jiali, Cui Yuanhao, Zhang Di, Zhang Ronghui, Mu Junsheng, Jing Xiaojun. (2025). Vehicle network beamforming method based on multi-modal feature fusion. Journal of Radar (Chinese and English), 14 (04), 994-1004.

[11] Cheng Dalei, Zhang Daiwei, Chen Yaqian. (2022). Review of Multimodal Emotion Recognition. Journal of Southwest University for Nationalities (Natural Science Edition), 48 (04), 440-447.

[12] Liang Tao. (2021). Design Optimization of Chinese Speech Recognition System in Complex Environment (Master's Thesis, Xidian University).Retrieved from https: //doi.org/10.27389/d.cnki.gxadu.2021.000611.