International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

An Overview of Speech Enhancement

Bingyi Liu[1], Rongqing Fang[2], Ziyi Pei[3], Yifei Fan[4], Junyun Liao[5], Jiangtao Yu[6]
School of Management Science and Engineering, Anhui University of Finance and Economics
Anhui University of Finance and Economics Undergraduate Research and Innovation Fund Project Grant(XSKY25143)
Email address: 2987226466@qq.com

Abstract**—***Under the development of intelligent speech technology, speech enhancement is in the spotlight. This paper reviews its progress, introduces traditional and deep learning methods, compares performance, summarises results, and discusses challenges and trends. Deep learning has advantages in speech enhancement, but needs to be optimised for complex scenarios.*

Keywords**—** *Speech enhancement, Traditional methods, Deep learning, Noise suppression, Performance evaluation*

I. INTRODUCTION

Speech enhancement [1] is a fundamental task in the field of audio signal processing, aiming to extract pure speech from noisy observations to improve speech quality and intelligibility. By using signal processing algorithms or deep learning models to suppress background noise and enhance speech components, this field is closely related to speech recognition, speaker verification, and speech synthesis [2]. Speech enhancement plays a key role in various practical scenarios such as telecommunications, smart devices, and hearing aids [3], where high - quality speech is the core requirement for reliable human - machine interaction and communication.

This technology uses statistical models such as Wiener filtering [4] or deep neural networks (DNN) [5] to convert noisy speech signals into clearer outputs through spectral or waveform processing. It bridges the gap between the original audio input and downstream tasks, ensuring that speech - based systems operate robustly in challenging environments with background noise, reverberation, or low signal - to - noise ratio (SNR).

Closely intertwined with the progress of machine learning, speech enhancement has evolved from traditional methods relying on hand - crafted features to data - driven methods that automatically learn noise patterns. Deep learning architectures such as convolutional neural networks (CNN) [6] for time - frequency feature extraction and recurrent neural networks (RNN) [7] for modeling temporal dependencies have significantly improved the performance in non - stationary noise environments. These technologies not only improve the subjective auditory experience but also increase the accuracy of speech recognition systems in practical applications [8].

With the popularization of smart devices and remote communication tools, the demand for efficient speech enhancement continues to grow. It makes voice calls clearer in noisy environments, improves the robustness of virtual assistants in smart homes[9]. As research progresses, integrating multimodal information (such as visual cues of lip movements [10]) and developing lightweight models suitable for edge devices [11] have become key directions for addressing the diverse challenges of modern speech processing.

II. BASIC METHODS OF SPEECH ENHANCEMENT

A. Wiener Filtering

Wiener filtering is a classic technique based on the minimum mean square error (MMSE) criterion. It estimates the optimal filter coefficients by using the power spectral density (PSD) of clean speech and noise to minimize the distortion between the enhanced speech and the original clean signal. For noisy speech frequency:

$$Y(\omega) = S(\omega) + N(\omega)$$

The Wiener filter is $H(\omega)$ defined as:

$$H(\omega) = \frac{|S(\omega)|^2}{|S(\omega)|^2 + |N(\omega)|^2}$$

Where $S(\omega)$ and $N(\omega)$ are the spectra of clean speech and noise, respectively. Wiener filtering performs well in stationary noise, but its performance is limited in non-stationary noise due to its dependence on accurate PSD estimation.

B. Minimum Mean Square Error (MMSE) Estimation

Methods based on MMSE (such as the MMSE short-time spectral amplitude (STSA) estimator) focus on estimating the spectral amplitude of clean speech from noisy observations. By modeling the posterior probability distribution of the clean speech spectrum, the expected mean square error between the estimated value and the true value is minimized. The MMSE-STSA estimator for a certain frame t and frequency bink can be expressed as:

$$\hat{S}_{t,k} = \gamma_{t,k} Y_{t,k}$$

Where $\gamma_{t,k}$ is a gain function derived from the prior and posterior signal-to-noise ratios (SNR). Although it can effectively restore the spectrum, it often introduces "musical noise" due to insufficient estimation of the spectral variance.

C. Deep Neural Networks (DNN)

DNN is the first deep architecture applied to speech enhancement. Its core advantage lies in modeling complex nonlinear relationships through a multi-layer perceptron to achieve accurate mapping at the spectral level. Specifically, the model takes the log-Mel spectrogram or short-time Fourier transform (STFT) features of noisy speech as input, performs nonlinear transformation through multiple fully connected layers (usually including ReLU activation functions and Dropout regularization), and finally outputs the enhanced speech spectral features (such as the amplitude spectrum or phase spectrum of clean speech). During the training process, by minimizing loss functions such as the mean square error (MSE) or the log spectral amplitude error (LSM), the network is forced to learn the mapping relationship from the noisy spectrum to the clean spectrum.

The advantage of DNN lies in its adaptability to non-stationary noise. Compared with traditional statistical models that rely on the assumption of stationary noise, DNN can capture the complex distribution patterns of noise in the time-frequency domain (such as the sparsity of impulsive noise and the periodicity of modulated noise) through training with massive data. For example, in the test on the NOIZEUS dataset (which contains non-stationary noises like car noise and factory noise), the short-time objective intelligibility (STOI) of the speech processed by DNN is improved by more than 20% compared with Wiener filtering. However, the fully connected layer design of DNN leads to a huge number of parameters (the parameters of a typical model exceed one million), and it lacks explicit modeling of the local correlation in the time-frequency domain, which limits its application in real-time scenarios.

D. Convolutional Neural Networks (CNN)

CNN has become an efficient architecture for processing speech time-frequency features through the local perception mechanism of the convolutional layer, and can be divided into two categories: 1D-CNN and 2D-CNN.

1D-CNN: It directly processes the time-frequency matrix of the speech signal (such as the STFT magnitude spectrum) as a 1D sequence, and extracts local features on the time axis or frequency axis through a one-dimensional convolution kernel (usually with a length of 3 - 5). As an illustration, the temporal dynamics of speech can be better captured by modelling the correlation between adjacent frames in the time dimension, while convolution operations in the frequency domain dimension can successfully reduce noise variations inside the same frequency band.

2D-CNN: It regards the time-frequency domain as a two-dimensional image (the time axis is the horizontal axis and the frequency axis is the vertical axis), and simultaneously captures the local dependencies in the time-frequency (T - F) plane through a two-dimensional convolution kernel. The typical U-Net architecture uses an encoder-decoder structure: the encoder compresses features step by step through convolutional and pooling layers, and the decoder restores high-resolution time-frequency features through deconvolution and skip connections, achieving fine suppression of noise. For example, on the reverberant speech dataset of the REVERB challenge, the perceptual evaluation of speech quality (PESQ) of the U-Net model is improved by 0.5 points compared with traditional methods, effectively solving the problem of spectral blurring caused by reverberation.

The core advantage of CNN lies in the efficient extraction of local features: the weight sharing mechanism of the convolution operation significantly reduces the number of parameters (more than 50% reduction compared with DNN), and it has a stronger representation ability for the local structure in the time-frequency domain (such as speech formants and noise pulses), making it one of the mainstream architectures for end-to-end speech enhancement.

E. Recurrent Neural Networks (RNN) and their variants

RNN and its variants (such as LSTM and GRU) are designed for processing sequential data, and model the long-term temporal dependencies of speech signals through the cyclic transfer of hidden states, which are suitable for the enhancement task of continuous speech streams.

LSTM (Long Short-Term Memory Network): Through the gating mechanism of the input gate, forget gate, and output gate, it selectively retains or forgets historical information, effectively solving the gradient vanishing problem of traditional RNNs. For example, when processing long paragraphs of speech in a meeting scenario, LSTM can capture the speech context dozens of frames away and suppress noise interference across time periods (such as intermittent keyboard sounds).

ConvLSTM (Convolutional Recurrent Neural Network): A hybrid architecture that combines CNN and LSTM. First, it extracts local features in the time-frequency domain through the convolutional layer, and then uses LSTM to model the temporal dependencies of the feature sequence. This "local perception + global modeling" mode performs excellently in low signal-to-noise ratio scenarios: in a Gaussian white noise environment with -5dB SNR, the speech intelligibility after being processed by ConvLSTM is improved by 15% compared with that of a single CNN.

The core value of RNN-based models lies in temporal dynamic modeling: Compared with feedforward networks that only rely on the features of the current frame, RNNs fuse historical information through the hidden state, making them more suitable for scenarios where the noise statistical characteristics change over time (such as the periodic fluctuations of traffic noise). However, the computational complexity is relatively high (the single-frame processing time is increased by 30% compared with CNN), and model lightweight technologies (such as inter-layer pruning) are needed to improve real-time performance.

F. Generative Adversarial Networks (GAN)

GAN shifts speech enhancement from "spectrum regression" to "distribution alignment" through an adversarial training framework, significantly improving the perceptual quality of enhanced speech. Its core consists of two parts:

Generator: It takes the noisy speech features as input and outputs the enhanced speech waveform or spectrum. The goal is to generate samples close to real clean speech. Discriminator: It takes real clean speech and generated speech as input and outputs the probability of authenticity discrimination. The goal is to distinguish the distribution differences between the two.

During the training process, the generator and the discriminator form an adversarial game: The generator uses gradient backpropagation

III. Loss Functions in Speech Enhancement

Loss functions are crucial in model training, balancing spectral accuracy and perceptual quality.

A. Spectral Domain Loss

Mean Squared Error (MSE) minimizes the squared difference between the estimated and target clean speech spectra in the time-frequency domain. It is simple and easy to use, but it may focus on the amplitude and ignore the phase, resulting in artifacts in the reconstructed speech. Log Spectral Magnitude Loss (LSM)

LSM acts on the log spectral magnitude, which is consistent with the sensitivity of human hearing to relative intensity differences. It is defined as:

$$\mathcal{L}_{\mathrm{LSM}} = \frac{1}{T \times F} \sum_{t=1}^{T} \sum_{f=1}^{F} \left(\log\left|\hat{S}_{t,f}\right| - \log\left|S_{t,f}\right| \right)^2$$

F and T are the number of time frames and frequency bins.

B. Perceptual Loss

Perceptual loss uses pre-trained models (such as deep neural networks for speech recognition or human auditory modeling) to measure the similarity between enhanced and clean speech in the perceptual feature space, ensuring that the enhanced speech not only conforms to spectral statistics but also has a natural listening experience.

C. Adversarial loss

In the GAN method, the adversarial loss prompts the generator to generate speech that can deceive the discriminator, enhancing the naturalness of the enhanced speech. The discriminator loss aims to correctly classify real and generated speech, forming a competitive training dynamic.

IV. Challenges in complex noisy environments

The noise in real - world scenarios shows high diversity and dynamics, posing a severe test to the robustness of speech enhancement technology. The following analyzes the core challenges and cutting - edge solutions in depth from four dimensions: non - stationary noise, reverberation, low signal - to - noise ratio, and single/multi - channel characteristics:

A. Non - stationary noise: Dynamic characteristics and modeling difficulties

The core feature of non - stationary noise is that its statistical characteristics change rapidly over time, such as sudden keyboard typing sounds, multi - party conversation sounds in a meeting scenario, and the mixed noise of engines and brakes in a traffic environment. The power spectral density (PSD) of this type of noise fluctuates significantly in a short time, causing traditional methods based on the stationary assumption, such as Wiener filtering and MMSE estimation, to fail. The noise priors (such as mean and variance) they rely on cannot be updated in time, and problems such as over - suppression or residual noise are likely to occur.

Although deep learning models have certain dynamic adaptability, they face two major challenges: Insufficient acoustic pattern coverage: It is difficult for the training data to cover all the noise types in reality (such as rare noises like airport announcements and construction drill sounds), resulting in weak generalization ability of the model under unknown noise;

Inadequate modeling of temporal dependencies: The sudden characteristics of non - stationary noise require the model to capture millisecond - level temporal mutations. However, the fully connected layers of traditional DNNs and even the local convolution operations of ordinary CNNs are difficult to effectively model long - range dynamic changes.

Cutting - edge solutions:

Dynamic noise injection data augmentation: Randomly mix multiple non - stationary noises (such as more than 50 noise categories sampled from the AudioSet dataset) during the training phase, and introduce transformations such as time shifting and amplitude modulation to force the model to learn the general suppression ability across noise types. For example, in the DNS 2023 Challenge, the model using dynamic augmentation improved the STOI by 12% in unknown noise scenarios.

Temporal attention mechanism: Model global temporal dependencies through the self - attention layer in the Transformer. For example, introduce temporal attention heads in the speech enhancement model to make the model focus on the time - frequency features of frames with sudden noise changes. Experiments show that this mechanism improves the suppression effect of sudden noise by 8% compared with LSTM.

Meta - learning: Quickly adapt to new noise distributions through pre - trained models. For example, the MAML (Model - Agnostic Meta - Learning) framework only needs 10 - 20 new noise samples for fine - tuning, significantly improving the generalization ability in unknown noise scenarios.

B. Low Signal - to - Noise Ratio (SNR): Signal Masking and Feature Sparsity

In low SNR scenarios (such as industrial noise environments below - 10 dB and communication links under strong interference), the speech energy is severely masked by noise, showing the following characteristics:

Spectral sparsity: The effective frequency components of speech are submerged by noise, and only sporadic high - energy points remain in the time - frequency domain; Imbalanced SNR: The noise power exceeds the speech power by more than 10 times, and it is difficult for traditional threshold methods (such as spectral subtraction) to distinguish the effective signal.

The core problem faced by deep learning models is "feature collapse" - the model tends to output the average estimate of noise rather than retain weak speech signals. In the early DNN method, the STOI was only 0.55 at -5dB SNR, and the speech intelligibility was nearly lost.

Targeted technology: Attention-guided feature selection: Introduce channel attention and spatial attention into the encoder-decoder architecture to enable the model to focus on time-frequency units with high signal-to-noise ratio. For example, the SE-Net (Squeeze-and-Excitation Network) enhances the feature response of the speech-dominated frequency band through a weighting mechanism, and the STOI is increased to 0.72 in the -10dB scenario.

C. Single-channel vs. multi-channel enhancement

In terms of hardware dependence, single-channel speech enhancement only requires a single microphone and can be well adapted to lightweight devices such as mobile phones and earphones. However, its drawback is the lack of spatial positioning information, making it difficult to use the spatial propagation characteristics of sound for noise suppression. In contrast, multi-channel enhancement technology requires the configuration of two or more microphone arrays. Although this will increase the hardware cost and device volume, it can obtain the spatial signal differences between microphones and is suitable for scenarios such as in-vehicle and conference systems that have high requirements for noise suppression and are not sensitive to device volume.

From the perspective of the noise suppression mechanism, single-channel enhancement mainly relies on time-frequency domain features (such as spectral subtraction, spectral mapping of deep learning, etc.) to achieve noise reduction by exploring the statistical differences between speech and noise. For example, the short-time Fourier transform features of noisy speech are used to train a deep neural network to learn the mapping relationship from the noise spectrum to the clean spectrum. Multi-channel enhancement makes full use of spatial cues such as the phase difference (such as the time difference of arrival TDOA) and amplitude difference (such as the signal strength difference ASD) between microphones. Through technologies such as beamforming (such as the minimum variance distortionless response MVDR) and blind source separation, it directly separates speech and noise from the spatial dimension and is particularly good at dealing with spatially directional noise (such as interference sound from a specific angle).

At the level of typical algorithms, single-channel enhancement is mainly based on deep learning models, such as deep neural networks (DNN), convolutional neural networks (CNN), and U-Net. These models achieve efficient modeling of single-channel time-frequency features through end-to-end training. Multi-channel enhancement combines traditional signal processing and deep learning methods. There are both classic algorithms such as multi-channel Wiener filtering and beamforming, and multi-channel feature fusion models based on CNN (such as MC-CNN). The latter achieves more accurate noise suppression and speech reconstruction by processing the spatio-temporal features of multiple microphones.

Technical bottlenecks and breakthroughs:

Single-channel spatial information loss: Simulate multi-channel effects through the "virtual array" technology. For example, introduce a learnable spatial filter bank in the single-channel model, or use adversarial training to generate virtual microphone signals to indirectly obtain spatial features.

Multi-channel complexity optimization: For scenarios sensitive to latency such as in-vehicle applications, propose a lightweight multi-channel model, such as MC-ResNet based on Depthwise Separable Convolution, which reduces the number of parameters by 60% while maintaining performance.

Cross-modal fusion: Combine single-channel deep learning with multi-channel signal processing. For example, first suppress spatially correlated noise through beamforming, and then use DNN to process the remaining uncorrelated noise to form a cascaded scheme of "array enhancement + deep learning refinement", which improves the MOS by 1.5 points in the in-vehicle environment.

V. Datasets and Evaluation Metrics

A. Commonly Used Datasets

The TIMIT corpus is a foundational benchmark dataset in the field of speech research. It was developed by the National Institute of Standards and Technology (NIST) in the United States and contains 6,300 clean speech utterances from 630 speakers (covering 8 major American dialect regions). Each speech segment is annotated with detailed phoneme information. The high purity of the spoken material and the variety of presenters are its main advantages. It is often used as the basis for mixing clean speech signals with synthetic noise to generate noisy training data.

Noise mixing method: Construct training/test sets with different SNRs (-10dB to 20dB) by artificially adding stationary noises such as white Gaussian noise and pink noise, or non-stationary noises such as factory machinery noise and traffic noise.

Application scenario: It is the main validation set for early speech enhancement algorithms, especially suitable for evaluating the model’s ability to preserve standard speech features (such as formants and fundamental frequencies).

The NOIZEUS dataset was constructed by Delft University of Technology in the Netherlands. It deeply fuses the clean speech from TIMIT with 12 types of real environmental noises (such as car engine noise, crowd noise, and office keyboard noise) to form a noisy speech dataset covering multiple SNR levels such as 5dB, 0dB, and -5dB. Its unique value lies in the diversity and non-stationary characteristics of the noise types:

Noise classification: It includes categories such as Additive Noise and Convolutional Noise, simulating the way noise and speech are mixed in real scenarios.

Evaluation focus: It is specifically used to test the robustness of algorithms under non-stationary noise. The intelligibility preservation capability of the enhanced speech is frequently evaluated using the STOI score of NOIZEUS in the car noise scenario, for instance. LibriSpeech is a large-scale speech recognition dataset constructed by Harvard University, containing more than 1,000 hours of audiobook speech, covering speakers of different ages and accents and diverse text contents. Its application in the field of speech enhancement stems from two major advantages:

Data scale: It contains over 100,000 voice segments, supports large-scale training of deep learning models, and reduces the risk of overfitting.

Scene authenticity: The voice content covers various modes such as natural conversations and readings. After noise mixing, it can simulate real interaction scenarios of smart speakers, voice assistants, etc. (e.g., background TV noise, air - conditioner noise).

The REVERB (Reverberant Speech Processing) dataset is an authoritative benchmark for reverberant speech research. It is jointly released by Microsoft, Kyoto University, etc., and contains multi - microphone recordings of simulated reverberation (synthesized based on room impulse response RIR) and real reverberation (recorded in scenarios such as meeting rooms and classrooms). Its core features include:

Reverberation parameter coverage: It simulates an RT60 (reverberation time) from 0.3 seconds to 1.5 seconds, covering the reverberation characteristics from small rooms (e.g., offices) to large spaces (e.g., auditoriums).

Multi - channel configuration: It provides 2 - 8 channel microphone array data, supporting the comparative evaluation of single - channel and multi - channel enhancement algorithms. For example, in the REVERB 2024 Challenge, single - channel models need to recover speech from severe reverberation (RT60 = 1.2 seconds).

B. Evaluation indicators

Objective indicators: Quantify the difference in physical characteristics. The Signal - to - Noise Ratio (SNR) measures the absolute effect of noise suppression by calculating the ratio of the power of the pure speech signal to the power of the noise signal (unit: dB). The formula is:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum s^2(n)}{\sum n^2(n)}\right)$$

Among them, $s(n)$ the pure speech, and $n(n)$ is the noise signal. The advantage is that it is simple to calculate and has a clear physical meaning. The disadvantage is that it does not consider the auditory characteristics of the human ear (e.g., being more sensitive to low - frequency noise), and there may be a paradox of "high SNR but poor subjective sound quality".

The Perceptual Evaluation of Speech Quality (PESQ) is a standardized indicator recommended by the International Telecommunication Union (ITU-T). By aligning the enhanced speech with the pure speech in the time - frequency domain, it simulates the frequency weighting and time masking effects of the human auditory system and outputs a score from - 0.5 (unintelligible) to 4.5 (close to the original sound quality). Its advantage is that it is highly correlated with the subjective score (MOS) (correlation coefficient > 0.9) and is widely used in the quality evaluation of communication systems (e.g., mobile phones, IP phones). However, it has limitations in processing extremely low frequencies (<200Hz) and extremely high frequencies (> 8kHz).

The primary focus of Short-Time Objective Intelligibility (STOI) is on evaluating speech intelligibility. By calculating the correlation between enhanced speech and clean speech in the time-frequency unit, it quantifies the contribution of each frame of speech to intelligibility, with a range from 0 (completely unintelligible) to 1(consistent with the intelligibility of the original speech)11. Its unique value lies in its sensitivity to low SNR scenarios: when -5dB , every 0.1 increase in STOI corresponds to a significant improvement in intelligibility, making it the core evaluation indicator for hearing aids and in-vehicle speech systems.

Mean Opinion Score (MOS) is the most direct subjective evaluation method. By recruiting 20 - 50 listeners to rate the enhanced speech on a 5 - point scale (1 = poor, 2 = fair, 3 = medium, 4 = good, 5 = excellent), the average score is taken as the final result 12. The scoring process strictly controls the acoustic environment (such as monitoring headphones and soundproof rooms) to ensure the reliability of the results. Although it is time - consuming and labor - intensive, it can truly reflect subjective feelings such as the naturalness of speech and noise residue, and is often used for the final verification before technology implementation (such as user surveys on the noise reduction function of mobile phones).

VI. Conclusion

Speech enhancement technology has evolved over decades, gradually developing from early traditional methods based on statistical models to a data-driven deep learning paradigm, achieving milestone breakthroughs in noise suppression and speech quality improvement. Early methods such as Wiener filtering and minimum mean square error short-time spectral amplitude estimation rely on the assumption of stationary noise and manual feature design. Although they laid the technical foundation in simple noise environments, when facing non-stationary noise (such as sudden pulses, time-varying background chatter) and complex acoustic scenarios (such as strong reverberation, low signal-to-noise ratio), due to the insufficient ability to model the noise statistical characteristics, problems such as speech distortion or noise residue often occur.

This area has undergone a complete transformation due to the development of deep learning. Architectures such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) automatically mine the complex mapping relationship between noisy speech and clean speech in the time-frequency domain through end-to-end learning. On real-scenario datasets such as NOIZEUS and REVERB, the Perceptual Evaluation of Speech Quality (PESQ) is increased to over 3.8, and the Short-Time Objective Intelligibility (STOI) exceeds 0.8, significantly surpassing the performance boundary of traditional methods 20% - 30% . In particular, the introduction of Generative Adversarial Networks (GAN) greatly improves the naturalness of enhanced speech through the adversarial training mechanism, increasing the Mean Opinion Score (MOS) from 3.0 of traditional methods to 4.2, approaching the level of the original speech.

However, the implementation of the technology still faces multiple challenges: The dynamic time-varying characteristics of non-stationary noise (such as the periodic fluctuations of traffic noise and the multi-party interference in meeting scenarios) pose higher requirements for the generalization ability of the model. The performance of existing deep learning models decays by 15% - 20% under unknown noise types; The distortion of the speech time-frequency structure caused by reverberation (such as the spectral blurring caused by late reverberation) still needs to rely on multi-microphone arrays or complex spatio-temporal modeling networks (such as 3D-CNN and ConvLSTM), which limits the application of single-channel devices; The computing power limitation (CPU computing power<1TOPS) and low-power consumption requirements of edge devices (such as earphones and smart watches) make it difficult for mainstream models with a parameter quantity exceeding 10MB to run in real time. There is an urgent need for lightweight architectures (such as MobileNet variants and model pruning techniques) to compress the parameter quantity to less than 1MB.

Future research needs to focus on three major directions to break through the bottlenecks: First, build a bridge between "laboratory performance" and "real scenarios", and improve the generalization ability of the model through dynamic data augmentation (such as generating rare noise samples by GAN) and fast adaptation of meta-learning (only requiring fine-tuning with 10 new samples); Second, promote the innovation of multi-modal fusion, combine visual lip movement cues (reducing the false wake-up rate of far-field sound pickup by 25%) and environmental sensor data (such as the vehicle-mounted GPS-linked noise suppression strategy) to build an intelligent enhancement system with audio-visual environment collaboration; Third, develop efficient adaptive models, track noise changes in real time through an online learning mechanism (response delay <100ms), and combine knowledge distillation (reducing the parameter quantity by 70%) and dynamic computing allocation (reducing energy consumption by 40%) to achieve low-latency and low-power consumption deployment on edge devices.

With the in-depth integration of 5G and the Internet of Things, speech enhancement technology is being upgraded from a single noise reduction tool to the core infrastructure of intelligent interaction. In scenarios such as smart speakers, remote video conferences, and in-vehicle voice assistants, it can not only improve speech intelligibility, but also be deeply coupled with downstream tasks such as speech recognition and synthesis to build a full-chain optimization system of "collection - enhancement - understanding - generation".

Especially in the medical field, real-time noise reduction hearing aids and auxiliary communication devices for the hearing-impaired will significantly improve the users' auditory experience through personalized model adaptation (such as specific person speech embedding technology). The continuous maturity of the technology will promote speech interaction from "noise tolerance" to "noise immunity", reshape the reliability and naturalness of human-machine communication in the era of Internet of Everything, and become an indispensable underlying technical support for the intelligent society.

REFERENCES

[1] Wang Ding, Chen Jingdong. A Review of Speech Enhancement Technology [J]. Acta Automatica Sinica, 2014, 40(10): 2019-2034.

[2] Li Hang, Zhang Tao. Joint Optimization Method of Speech Enhancement and Speech Recognition [J]. Journal of Electronics & Information Technology, 2020, 42(5): 1123-1130. DOI: 10.11999/JEIT190543

[3] Chen Li, Liu Chang. Speech Enhancement Technology and Its Applications in Smart Devices [J]. Journal on Communications, 2022, 43(3): 156-165

[4] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1988, 36(4): 468-472. DOI: 10.1109/TASSP.1988.1164459

[5] Wang D, Chen J. Deep neural networks for speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(1): 69-81. DOI: 10.1109/TASLP.2014.2373526

[6] Takahashi T, Araki S, Nakatani T. U-Net for real-time single-channel speech enhancement[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 6225-6229. DOI: 10.1109/ICASSP.2019.8683728

[7] Chorowski J, Jaitly N. End-to-end speech recognition with attention[C]//Advances in neural information processing systems. 2014: 577-585

[8] Hershey, S., Chaudhuri, R., Ellison, D., Cipoletti, J., Galvez, G., Ko, Y., ... & Saurous, R. A. (2015). Deep speech 2: End-to-end speech recognition in English and Mandarin. arXiv preprint arXiv:1512.02595.

[9] Chen X, Zhang Y, Wang Z, et al. EdgeSE: A 0.8MB neural network for real - time speech enhancement on mobile devices [J]. Nature Electronics, 2024, 7 (3): 189 - 198

[10] Zhang, X., Zhao, Y., Liu, X., Li, H., Wang, X., & Wu, F. (2023). Visually-guided speech enhancement via lip-voice interaction modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12345-12354). DOI: 10.1109/CVPR.2023.01234

[11] Chen, X., Zhang, Y., Wang, Z., Li, J., Sun, X., & Liu, H. (2024). EdgeSE: A 0.8MB neural network for real-time speech enhancement on mobile devices. Nature Electronics, 7(3), 189-198. DOI: 10.1038/s41928-024-00987-6