Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 9890456, 1 page
https://doi.org/10.1155/2023/9890456

Retraction

Retracted: Research and DSP Implementation of Speech Enhancement Technology Based on Dynamic Mixed Features and Adaptive Mask

Journal of Electrical and Computer Engineering

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Yang and Y. Tang, "Research and DSP Implementation of Speech Enhancement Technology Based on Dynamic Mixed Features and Adaptive Mask," *Journal of Electrical and Computer Engineering*, vol. 2022, Article ID 7287072, 10 pages, 2022.

Hindawi
Journal of Electrical and Computer Engineering
Volume 2022, Article ID 7287072, 10 pages
https://doi.org/10.1155/2022/7287072

Research Article

Research and DSP Implementation of Speech Enhancement Technology Based on Dynamic Mixed Features and Adaptive Mask

Jie Yang and Yachun Tang

School of Electronics and Information Engineering, Hunan University of Science and Engineering, Changsha 425199, China

Correspondence should be addressed to Yachun Tang; 1799@huse.edu.cn

Received 19 March 2022; Revised 12 April 2022; Accepted 27 April 2022; Published 27 May 2022

Academic Editor: Wei Liu

A deep learning speech enhancement algorithm based on dynamic hybrid feature and adaptive mask and DSP implementation is proposed in this paper, which solves the problem of feature loss and improves the performance of speech enhancement. The dynamic features incorporate the log Mel power spectrum, Mel cepstral coefficients, and Multiresolution Auditory Cepstral Coefficients (MRACC) and capture the speech transient information by deriving the derivatives to comprehensively represent the nonlinear structure of speech and reduce distortion. To make the system improve the speech quality while reducing the speech distortion as much as possible, a soft mask that can be adaptively adjusted considering the signal-to-noise ratio information is proposed, which can be automatically adjusted according to the different speech signal-to-noise ratio information to obtain the mask value under the corresponding signal-to-noise ratio conditions, and phase difference information that can improve the speech intelligibility is incorporated in it. Then, an improved deep neural network model is designed to effectively improve the speech enhancement performance. Finally, the hardware and algorithm software design of the DSP-based speech enhancement system is given. Experimental simulations are carried out for multiple voices in different noise backgrounds. The experimental results indicate that the performance indexes of the proposed method are significantly improved compared with the existing speech enhancement methods, which verifies the feasibility and superiority of the proposed method.

1. Introduction

As artificial intelligence technology develops, the application of speech signal processing technology is becoming more and more widespread, and speech enhancement, as one of the key aspects, is also a hot spot for researchers' attention. Speech enhancement is a technique to extract useful signals from noisy backgrounds, reduce interference, and minimize distortion and can be applied in artificial intelligence, hearing aids, speech recognition, and other fields [1]. Currently, speech enhancement methods can be divided into two categories: unsupervised and supervised. Unsupervised speech enhancement is mostly based on unreasonable assumptions such as smooth noise and uncorrelated speech noise, which leads to weak ability to suppress nonsmooth noise and produces speech distortion. Representative algorithms include spectral subtraction and Wiener filtering [2]. Supervised speech enhancement suppresses noise by learning the statistical properties of the signal, which has obvious advantages in low signal-to-noise environments and nonsmooth noise, and can be divided into two types based on shallow and deep models. Shallow layer models include Hidden Markov and shallow neural networks. This model limits the learning ability and performance cannot be effectively improved because the number of layers and the number of nodes per layer are small and the data used for training is also small. Deep-layer models are able to learn interspeech nonlinear relationships in depth, which greatly improves their generalization performance in unknown noisy environments [3]. It can be roughly divided into 3 classes: feature mapping-based speech enhancement, acoustic features of input, and output signals. Speech

enhancement is based on time-frequency masking, with input acoustic features and output time-frequency masking. Signal approximation-based speech enhancement, a fusion of the first two methods, trains the model to predict the masked value, and the final optimization goal is to estimate the mean square error of speech versus pure speech so that the network converges to an optimal point. Therefore, signal approximation-based speech enhancement has better performance in dealing with nonsmooth noise and has become a hot research topic [4, 5].

In literature [6], the nonlinear relationship between noisy speech features and the time-frequency mask is learned by deep neural networks, and a series of time-frequency-based masks such as Ideal Binary Mask (IBM), Ideal Ratio Mask (IRM), and Ideal Amplitude Mask (IAM) are compared. The quality and intelligibility of the enhanced speech are optimal when IRM is selected as the learning target for speech enhancement. Literature [7] learns nonlinear mappings by deep neural networks using three types of time-frequency masks: IBM, IRM, and Phase Sensitive Mask (PSM). Literature [8] uses recurrent neural networks and two masks, IBM and IRM, to achieve speech enhancement. However, IRM does not consider the phase information which is closely related to speech intelligibility, and it is determined according to the proportion of speech energy in the sum of speech and noise energy under different signal-to-noise ratio conditions, which cannot be automatically adjusted according to the different signal-to-noise ratios, and it is easy to cause the loss of target speech components. Speech features can characterize the characteristics of speech signals, and different speech features represent different speech attributes.

In recent years, deep learning has achieved great results in the fields of image classification and speech recognition, and researchers have started to introduce deep neural networks into speech enhancement problems to obtain better speech enhancement performance. Deep learning generally uses neural networks to process data, and the commonly used neural networks are deep neural networks (DNN), convolutional neural network (CNN), recurrent neural networks (RNN), and so forth. Literature [9] combines deep learning networks with minimum mean square error to improve speech enhancement performance. Literature [10] proposed RNN and deep neural networks for speech enhancement and designed two mean square errors as learning objectives to control speech distortion and noise reduction. There are also speech enhancement methods based on deep models such as temporal convolutional neural networks and adversarial generative networks. The Mel-Frequency Cepstral Coefficient (MFCC) is commonly used for time-frequency-domain features, but the Mel filter is prone to leakage at high frequencies, thus losing effective speech features, and cannot better simulate the crossover characteristics of the basilar membrane of the human ear. Literature [11] combines MFCC and normalized spectral subband centroids (NSSC) prime and implements speech enhancement using adversarial generative networks. The current speech enhancement algorithms tend to ignore the variation of speech phase information, resulting in less-than-optimal enhancement.

Digital audio is ubiquitous in modern life, and, with the development of Internet and media technologies and the rise of IoT infrastructure, digital media is widely used in a variety of scenarios such as cloud monitoring, webcasting, and remote intercom [12]. There are various hardware implementations of digital audio processing algorithms, such as the commonly used FPGA implementation. FPGAs allow static repetitive programming and online debugging, which greatly improve the fault tolerance of the product, but this comes at the expense of hardware cost and energy efficiency. As a result, FPGAs typically have lower resource utilization than dedicated chips [13]. DSPs dominate the market with their software and hardware resources suitable for digital signal processing. The use of DSP chips to develop digital audio processing devices takes into account the disadvantages of low resource utilization of FPGAs and long development cycles of dedicated audio processing chips and they are widely used in modern digital signal processing [14].

Based on the above research works, a deep learning speech enhancement algorithm and DSP implementation based on dynamic hybrid features with adaptive mask is proposed to improve the performance of speech enhancement. Firstly, three features of noisy speech are extracted and spliced to obtain static features, and then the first-order and second-order difference derivatives are found to capture the transient signals of speech and fuse them into dynamic features to reduce speech distortion. Secondly, in order to minimize the distortion while filtering the background noise, an adaptive soft mask with automatic adjustment according to the signal-to-noise ratio is proposed as the learning target. The soft mask also incorporates the phase difference information of speech, which can improve the masking effect and enhance the intelligibility of speech. Then, a causal gated recurrent unit- (CGRU-) based learning is designed to enable the network model to be trained more stably under more relaxed conditions. Finally, DSP software and hardware implementations are given, and the advantages of the proposed algorithm are verified by designing experiments.

Section 2 is an introduction to related work. Section 3 is about the recommendation algorithm based on music gene. Section 4 is a recommendation algorithm based on improved knowledge graph. Section 5 is the hybrid recommendation algorithm. Section 6 is the conclusion.

2. Improved Deep Neural Network Model

2.1. Gated Recurrent Unit (GRU). A schematic diagram of the gated recurrent neural network unit is shown in Figure 1, where i_n, b_n, and b_{n-1} are the input, output, and output of the current moment and the previous moment, respectively. r_n, k_n, and $\widetilde{b}_n$ are the reset gate, update gate, and candidate hidden state, respectively. The gated recurrent unit (GRU) neural network alleviates the network overfitting problem to some extent due to the gating

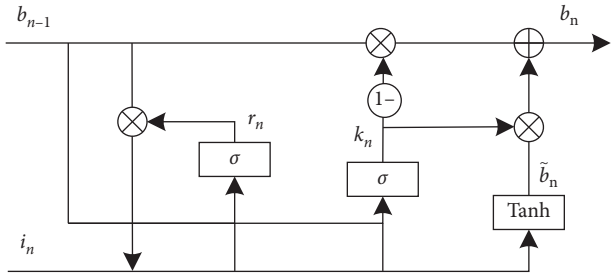

Figure 1: Gated recurrent unit (GRU).

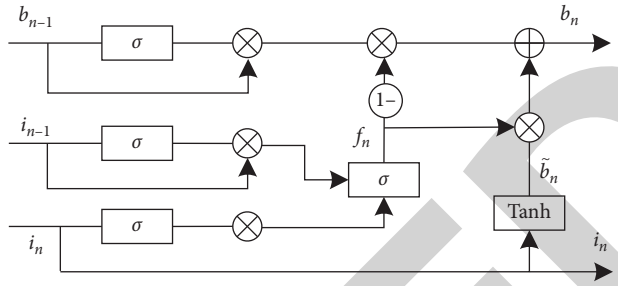

Figure 2: Causal gated recurrent unit (CGRU).

mechanism. The network is able to learn longer temporal relationships. Gated recurrent unit (GRU) neural network, as an improvement of long short-term memory (LSTM) network, has improved in the complexity of network structure, the number of network parameters, and other metrics. As shown in Figure 1, compared to the three-gate structure of the long short-term memory (LSTM) network, the gated recurrent unit (GRU) neural network has only reset gate r_n and update gate k_n.

The unit update relation of the gated recurrent neural network can be expressed by the following formula:

$$\begin{aligned} k_n &= \sigma\left(M_k i_n + P_k b_{n-1} + h_k\right), \\ r_n &= \sigma\left(M_r i_n + P_r b_{n-1} + h_r\right), \\ g_n &= \tanh\left(M_g P_g\left(b_n - 1 \odot r_n\right) + h_g\right), \\ b_n &= \left(1 - k_n\right) \odot b_{b-1} + k_n \odot g_n. \end{aligned} \tag{1}$$

In formula (1), M and P are the weight matrices. h is the bias term. They are trainable parameters. $\odot$ is the Hadamard product and σ is the Sigmoid activation function.

2.2. Cause-and-Effect Gated Recurrent Neural Unit (CGRU). A schematic diagram of the causally gated recurrent neural unit (CGRU) designed in this paper is shown in Figure 2. In order to solve the problem of fixed-time delay in traditional neural network speech enhancement due to the symmetric window with noncausal (input of $2N+1$ frames), in this study, a causal (input of $N+1$ frames) network input is used.

Since a causal network input is used, the feature information of the speech signal obtained by the neural network decays to a 0.5 multiple of the noncausal input. To reduce its impact on the neural network learning, this paper makes full use of the feature information of the speech signal from the previous N frames and incorporates the input features i_{n-1} of the previous moment in the CGRU network.

As shown in Figure 2, the output b_n of the CGRU neural network at the current moment is determined by the input i_{n-1} of the previous moment and the output b_{n-1} of the previous moment together with the input i_n at the current moment. This makes full use of the speech signal features of the previous frame. i_n, b_n, i_{n-1}, and b_{n-1} are the input and output of the current moment and the input and output of the previous moment, respectively. Inspired by the spatial attention mechanism and the gated linear unit (GLU), this paper first computes the weighted feature vectors of i_n, i_{n-1}, and b_{n-1} in the unit input of the CGRU neural network as follows:

$$\begin{aligned} \hat{i}_n &= \sigma\left(M_i i_n\right) \odot i_n, \\ \hat{i}_{n-1} &= \sigma\left(M_{i-1} i_{n-1}\right) \odot i_{n-1}, \\ \hat{b}_{n-1} &= \sigma\left(M_{b-1} b_{n-1}\right) \odot b_{n-1}. \end{aligned} \tag{2}$$

After calculating formula (2), $\hat{i}_{n-1}$ and $\hat{i}_n$ are used to calculate the forgetting gate f_n, which can be expressed by the following formula:

$$f_n = \sigma\left(M_n \hat{i}_n + M_{n-1} \hat{i}_{n-1} + h_f\right). \tag{3}$$

Unlike GRU, the candidate hiding status of CGRU is determined only by the current input i_n.

$$\tilde{b}_n = \tanh\left(M_b i_n + h_b\right). \tag{4}$$

The output b_n of the network unit at the current moment is determined by the candidate hidden state $\tilde{b}_n$ and forgetting gate f_n, with the band-weighted feature $\hat{b}_{n-1}$ of the previous moment's output, as in the following formula:

$$b_n = f_n \odot \tilde{b}_n + \left(1 - f_n\right) \odot \hat{b}_{n-1}, \tag{5}$$

where $\odot$ is the Hadamard product. σ is the Sigmoid activation function. In order to reduce the complex structure of the network, only one forgetting gate f_n is used in the CGRU network. Meanwhile, to address the problem of speech enhancement performance degradation due to the reduction of the input speech signal feature information in causal speech enhancement, this paper makes full use of the speech signal features of previous frames, the input i_{n-1} of the previous moment fused in the input i_n of the current moment, and, at the same time, the gated linear unit (GLU) mechanism to control the delivery of feature information, which greatly improves the performance of the network.

3. Speech Enhancement Algorithm in This Paper

3.1. Dynamic Characteristics. As shown in Figure 3, different speech features reflect different properties of the speech signal. LMPS smoothes the spectrum after a Mel filter bank and eliminates the effect of harmonics,

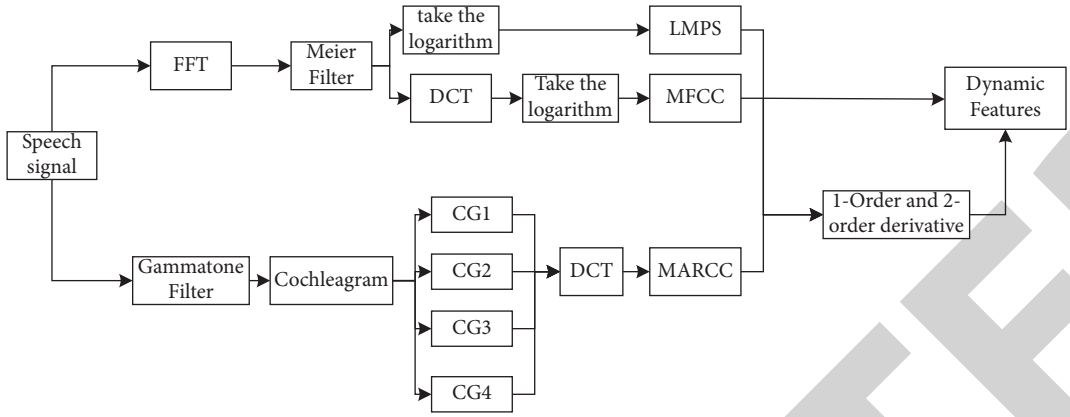

Figure 3: Block diagram of dynamic features extraction.

highlighting the resonant peaks of speech. MFCC reflects the relationship between the various dimensions of the noisy speech power spectrum. MRACC is an improved MRCG feature with four cochleagram sparse representations with different resolutions, which can represent global and local information. In order to fully represent the speech nonlinear structure, the three features are combined and complementary to obtain a relatively complete static feature. The first-order and second-order differential derivatives are then derived for the spliced features to capture the transient information of speech. The differential features describe the connection of adjacent frames of speech and avoid relying only on the network to obtain time-varying information of speech. The combination of dynamic and static features improves the inadequate incompleteness of existing features in representing speech structure, resulting in less distortion in reconstructed speech and high speech intelligibility.

The specific extraction process is as follows:

(1) The speech signal is Mel filtered logarithmically and discrete cosine transformed to obtain the logarithmic Mel power spectrum and Mel cepstral coefficient features. Meanwhile, after gammatone filtering, four CochleaGram (CG)-64 channels were obtained and windowered. Four different resolutions of CG1, CG2, CG3, and CG4 were obtained by power law transform. After splicing, the whole was performed with discrete cosine transform to obtain MRACC features.

(2) The three speech features are spliced to obtain the splicing static feature W; namely,

$$W(x,w) = \left[W_{\mathrm{LMPS}}(x,w); W_{\mathrm{MFCC}}(x,w); W_{\mathrm{MRACC}}(x,w)\right], \tag{6}$$

where x denotes the number of frames and w is the feature dimension index. $W_{\mathrm{LMPS}}(x,w)$, $W_{\mathrm{MFCC}}(x,w)$, and $W_{\mathrm{MRACC}}(x,w)$ denote the LMPS, MFCC, and MRACC features, respectively.

(3) The first-order and second-order difference derivatives are found for the spliced static features to obtain the difference features ΔW and $\Delta(\Delta W)$:

$$\Delta W(x,w) = \frac{\sum_{z=1}^{2} z\left(W(x+z,w) - W(x-z,w)\right)}{\left(2\sum_{z=1}^{2} z^2\right)^{1/2}},$$

$$\Delta\left(\Delta W(x,w)\right) = \frac{\sum_{z=1}^{2} z\left(\Delta W(x+z,w) - \Delta W(x-z,w)\right)}{\left(2\sum_{z=1}^{2} z^2\right)^{1/2}}, \tag{7}$$

where z is the index, representing the first and last two frames of the current frame.

(4) The obtained features are fused to form dynamic features Ω.

$$\Omega(x,w) = \left[W(x,w); \Delta W(x,w); \Delta\left(\Delta W(x,w)\right)\right]. \tag{8}$$

3.2. Adaptive Soft Mask. In a deep neural network-based speech enhancement system, the performance of the learning target has a direct relationship with the effect of speech enhancement, which determines the degree of distortion and the amount of residual background noise in the enhanced speech. Among many learning targets, using IRM as the learning target for speech enhancement is the most effective, and its value is taken according to the pure speech energy and noise energy in each time-frequency unit, which can effectively improve the quality of enhanced speech and filter out the background noise. However, since IRM is used to filter out noise under different signal-to-noise conditions with the same technical means and cannot be automatically adjusted according to different signal-to-noise information, the problem of eliminating useful speech components while retaining noise components often occurs. In the traditional IRM, only the amplitude information of speech is considered, and the phase information, which affects the intelligibility of speech, is ignored. Therefore, this paper proposes a new adaptive soft mask, which can be automatically adjusted according to the different speech signal-to-noise information to obtain the mask value under the corresponding signal-to-noise conditions. At the same time, it can incorporate the phase information of speech to improve the speech intelligibility while enhancing the speech quality.

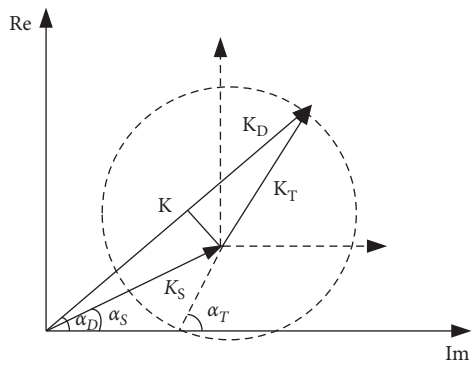

Figure 4: Phase geometry relationship diagram.

Figure 4 shows the geometric relationship of the phases. K_D denotes noisy speech. K_S denotes pure speech. K_T denotes the amplitude of noisy speech. α_D, α_S, and α_T are the phases of noisy speech, pure speech, and noisy speech, respectively, and, from Figure 4, it can be seen that

$$K_S^2 = K_D^2 + K_T^2 - 2K_D K_T \cos(\alpha_T - \alpha_D). \tag{9}$$

The phase difference information between noisy speech and noisy speech can be derived from the defined formula of the a priori S/N ratio ξ and the a posteriori S/N ratio γ.

$$\begin{aligned}\cos(\alpha_{TD}) &= \cos(\alpha_T - \alpha_D)\\ &= \frac{K_D^2 + K_T^2 - K_S^2}{2K_D K_T}\\ &= \frac{\gamma + 1 - \xi}{2\sqrt{\gamma}}.\end{aligned} \tag{10}$$

According to the geometric relationship in the figure, it can be concluded that

$$\begin{aligned}\cos(\alpha_T - \alpha_D) &= \frac{(K_D - K)}{K_T},\\ \cos(\alpha_{DS}) = \cos(\alpha_D - \alpha_S) &= \frac{K}{K_S}.\end{aligned} \tag{11}$$

Therefore, the phase difference information between pure speech and noisy speech can be expressed as

$$\begin{aligned}\cos(\alpha_{DS}) &= \frac{K}{K_S}\\ &= \frac{K_D - K_T\cos(\alpha_T - \alpha_D)}{K_S}\\ &= \frac{K_D/K_T - \cos(\alpha_T - \alpha_D)}{K_S/K_T}\\ &= \frac{\sqrt{\gamma} - \gamma + 1 - \xi/2\sqrt{\gamma}}{\sqrt{\xi}}\\ &= \frac{\gamma + \xi - 1}{2\sqrt{\gamma\xi}}.\end{aligned} \tag{12}$$

If $\cos(\alpha_{DS}) < 0$ or $\cos(\alpha_{TD}) < 0$ is not favorable for speech recovery, it is set to 0 as the minimum threshold. After incorporating the phase difference information in the time-frequency mask, the new mask $\overline{R}$ is obtained as follows:

$$\overline{R} = \frac{S^2(n,f)\max(\cos(\alpha_{DS}(n,f)),0)}{S^2(n,f)\max(\cos(\alpha_{DS}(n,f)),0) + T^2(n,f)\max(\cos(\alpha_{T\,D}(n,f)),0)}. \tag{13}$$

In the above formula, $S^2(n,f)$ denotes the speech energy in the fth band of the tth frame. $T^2(n,f)$ denotes the noise energy in the fth band of the tth frame. In the process of testing the performance of the mask $(\overline{R}(n,f))^{\beta}$ with different values of β composition, it is found that each of the new mask $\overline{R}(n,f)$ $(\beta = 1)$ and its square root mask $\sqrt{\overline{R}(n,f)}$ $(\beta = 0.5)$ has advantages. So the two are combined in a certain ratio to make the best effect of speech enhancement, and the ratio mask R is obtained as

$$R(n,f) = \alpha\overline{R}(n,f) + (1-\alpha)\sqrt{\overline{R}(n,f)}. \tag{14}$$

Experimentally, it is proved that the best effect is obtained when α is 0.7, so $\alpha = 0.7$. The obtained ratio mask R incorporates the phase information of speech and combines the advantages of different power masks.

3.3. Neural Network Speech Enhancement Based on Dynamic Features and Adaptive Mask. In the training phase, with the objective of minimizing the minimum mean square error cost function, as well as the dynamic features of the training set from the speech data samples, adaptive ratio mask is extracted as the input of the neural network model. In order to keep the training process stable, both inputs and outputs are mean-variance-normalized, and the optimal network model is trained and saved. In the testing phase, the normalized dynamic features of the test sample set are extracted and input to the trained neural network model to predict the adaptive mask. Finally, the phase-reconstructed speech with noisy speech is combined with the output to get the best enhancement result. The block diagram of the neural network speech enhancement system based on dynamic features and adaptive ratio mask is shown in Figure 5.

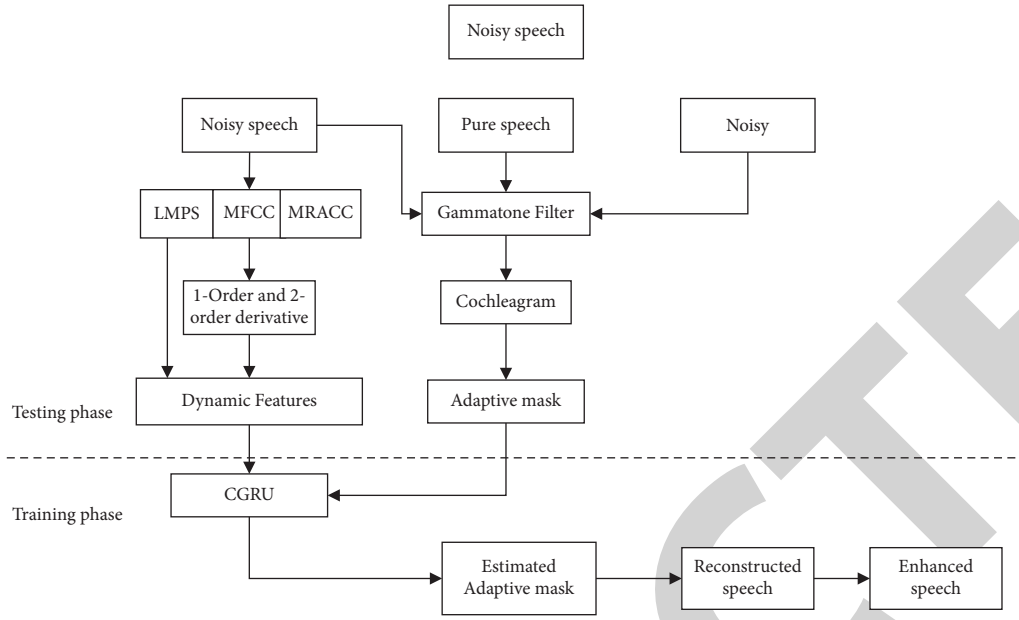

Figure 5: Block diagram of a neural network speech enhancement system based on dynamic features and adaptive mask.

4. Speech Enhancement Hardware and Software Design

4.1. Hardware Design of DSP-Based Speech Enhancement System. TI Production TMS320F281x series DSP provides a variety of peripheral communication interfaces, such as serial communication interface (SCI), serial peripheral interface (SPI), and multichannel buffered serial port (McBSP). Among them, McBSP supports full-duplex communication mechanism and provides double-buffered transmit and triple-buffered receive registers, allowing continuous data stream transmission. The data length is programmed to allow direct serial connection with industry standard decoders (CODEC), analog interface chips (AIC), and so forth. AIC23 is a sigma-delta high-performance audio codec chip with integrated 16-bit A/D and D/A converters inside for seamless connection to the DSP's McBSP. Its sampling rate can be programmed by DSP to achieve the reception and transmission of voice signals at high speed. At the same time, DSP with its high-speed frame processing capability, flexible use, low power consumption, and other advantages has gradually become the first choice of digital voice processing. Therefore, this paper selects TMS320F2812 chip as the main control chip with AIC23 and the corresponding peripheral circuit to complete the system hardware design. However, the DSP comes with a limited program and data memory capacity, usually difficult to meet the needs of speech processing. Therefore, 256 K × 16-bit SRAM is extended off-chip as external data memory and 512 K × 16-bit FLASH as external program memory. The system architecture block diagram is shown in Figure 6.

AIC23 has a separate control interface and data interface. The control interface is used to configure 11 registers inside the device, set the operating status of the audio chip, and initialize AIC23. The control interface works in SPI and I2C mode, which can be selected through the chip pins. The data interface transmits the data for AD conversion and DA conversion through the DIN and DOUT pins to achieve a seamless connection with the McBSP. The operating mode of the data interface can be set to DSP mode by the digital audio format register, while making the AIC23 work in the main mode; that is, the AIC23 provides the clock source and generates the shift clock and frame synchronization signal for serial communication through the divider. CLKX, CLKR, and BCLK are clock synchronization signals. CLKR and CLKX are connected by a 0 Ω resistor. FSX, FSR, LRCIN, and LRCOUT are frame synchronization signals. Before achieving normal communication between the data interface and McBSP, serial data needs to be continuously transferred to the control interface through the SPI port of the DSP for the purpose of configuring AIC23. Set the AIC23 clock to normal mode with a sample rate of 8 K and set the appropriate input/output signal gain. At the same time, AIC23 also has a function that other audio processing chips do not have, namely, analog bypass setting, which sends the input analog signal directly out for playback without going through AD and DA conversion, which is very important for system debugging.

The microphone captures the noisy speech signal and inputs it to AIC23 for anti-alias filtering, as well as A/D conversion, and passes it through McBSP to DSP chip for noise reduction processing. At the same time, the processed data is passed back to AIC23 via McBSP for D/A conversion and reconfiguration filtering. Usually AIC23 has a built-in headphone driver circuit, so there is no need for external driver processing, but the voice signal after noise reduction processing is output directly from the headphone.

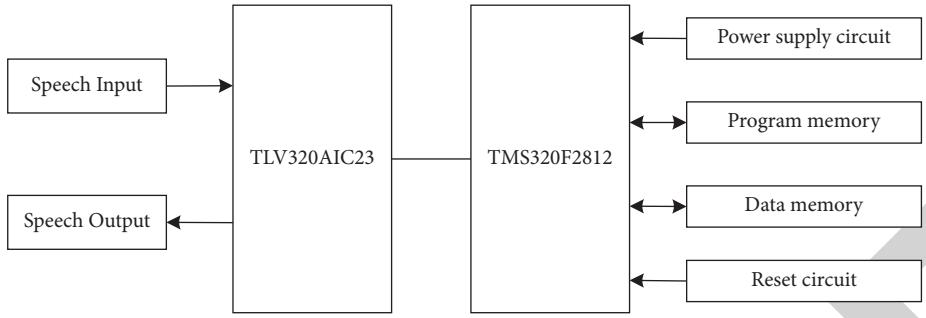

Figure 6: DSP speech enhancement system hardware structure design.

4.2. DSP-Based Speech Enhancement Algorithm Software Design. The performance of the algorithm is verified by first writing a speech noise reduction algorithm based on skip line and decoupling through MATLAB. Then the algorithm is rewritten in C language and assembly language in CCS integrated development environment and downloaded to DSP for online simulation debugging. The system software implementation flow is shown in Figure 7.

(1) Allocate program and data memory space rationally, with program segments and lookup table data defined in FLASH for read operations only. The data segment is allocated in DARAM, which can perform read-and-write operations at the same time to avoid running delays caused by calls and jumps.

(2) Initialize the CPU frequency of the DSP by configuring the on-chip clock mode register CLKMD.

(3) Using the SPI port of the DSP, initialize each register inside the AIC23 to set its operating mode, the number of bits of data transmission, sampling rate, and so forth. Initialize the McBSP and complete the configuration of each serial port register to ensure its normal communication with AIC23.

(4) Open a data buffer, because the speed of voice enhancement processing sometimes cannot keep up with the speed of data reception. To avoid frame loss, open a multiframe data buffer to save the unprocessed data. The speech enhancement algorithm is called during CPU idle time to complete the noise reduction process of the cached data.

(5) Open the serial port receive interrupt and start receiving data. During each interrupt processing, one piece of voice data is received and one piece of processed voice data is sent back to AIC23 for subsequent processing and output via McBSP.

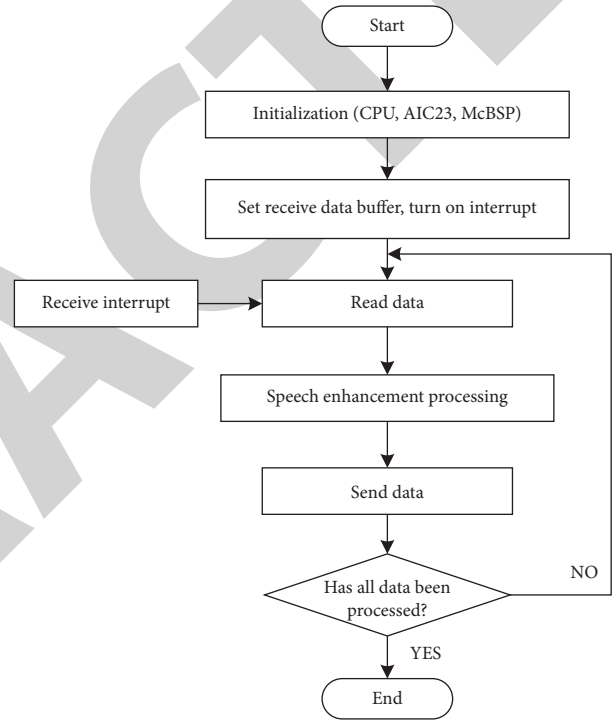

Figure 7: System main program flow chart.

5. Experimental Results and Analysis

5.1. Experimental Data and Experimental Setup. In order to verify the effectiveness of the proposed method, Pure Speech selects randomly 2000 strips of audio data from the training set using the TIMIT speech dataset as the training set. The 500 randomly selected strips of audio data in the test set are used as the test set. The noise of the training set is selected using the 100-species ambient noise from the literature, and the noise of the test set is selected using the 15-species noise from the NOISE-92 noise library as the selection of the test set noise. The 2000 randomly selected audio strips from the TIMIT training set and 100 ambient noises were randomly mixed at three signal-to-noise ratios of −5, 0, and 5 to generate a training dataset of 8000 tracks. The 500 strips of pure speech data randomly selected from the TIMIT test set and 15 noises from the NOISE-92 noise library are randomly mixed with three signal-to-noise ratios of −5, 0, and 5 to generate a test dataset of 2000 strips of noisy speech data. When extracting the features, the sampling frequency of pure speech and noise is set to 8000 HZ, and the frame length is 256 (about 31 ms) with frame shift of 128.

The network construction and training are done in the environment of Keras/TensorFlow 2.0. The initial learning rate of the network is set to $1e-4$, and the decay coefficient of the learning rate is set to $1e-6$ in order to make the network converge better, and the maximum number of

learning iterations is 50 times. The network is trained with batch gradient descent algorithm, and the batch size is set to 256. Adam is used for iterative optimization of the gradient descent algorithm. The loss function of the training network is selected using the mean absolute error (MAE).

In order to verify the effectiveness of the proposed method, 4-layer GRU, SRNN, SRU, and 4-layer CGRU network structures are designed in this paper with 512 neural network units in each layer.

5.2. Experimental Results and Analysis. To verify the effectiveness of the proposed algorithm, we do test experiments on the 4-layer GRU, SRNN, SRU, and 4-layer CGRU network structure models, respectively. The four noises in the test set, factory2, destroyerengine, buccaneer1, and hfchannal, are mixed with 500 pure speeches in the test set at three signal-to-noise ratios of −5, 0, and 5, respectively, and then the four models are compared to verify the effectiveness of the proposed algorithm.

The performance metrics for speech enhancement were selected as Short-Term Objective Intelligibility of Speech (STOI) and Perceptual Evaluation of Speech Quality (PESQ), with STOI ranging from 0 to 1 and PESQ ranging from −0.5 to 4.5. The larger the value is, the higher the enhanced speech quality and the speech intelligibility are. As shown in Tables 1 and 2, the average speech perception quality and average speech short time intelligibility are obtained from different network model structures.

The analysis of the mean Perceptual Evaluation of Speech Quality (PESQ) and the mean Short-Term Objective Intelligibility of Speech (STOI) in Tables 1 and 2 shows that the simple recurrent neural network (SRNN) has the worst speech enhancement effect relatively, and the simplified recurrent unit (SRU) neural network and the gated recurrent unit (GRU) neural network get relatively good speech enhancement effect. This is precisely because the simple recurrent neural network does not learn this long-term dependency. The gating mechanism used by GRU and SRU largely enhances the learning ability of the network. Compared with the three other networks, the causal speech enhancement network CGRU proposed in this paper outperforms the traditional network structure in terms of speech quality and short-time intelligibility of speech. In addition, the cell structure of the CGRU network adopts the gating mechanism of gated recurrent neural networks in order to make full use of the input features previous feature information. In the output feature calculation of the current network, CGRU not only integrates the input i_n at the current moment with the output b_{n-1} at the previous moment but also incorporates the input i_{n-1} at the previous moment. This makes full use of the previous N-frame feature information of the speech signal, as shown in Figure 8.

In order to verify the effectiveness of the overall algorithm in this paper, it is compared with other existing algorithms.

In Algorithm 1, the MRACC feature and IRM, which have the best effect among the 3-species features, are used to train the neural network. In Algorithm 2, the joint LMPS, MFCC, MRACC, and IRM were used to train the neural network. In Algorithm 3, the dynamic features and adaptive mask r are used to jointly train the neural network. In Algorithm 4, speech enhancement method based on the CGRU depth model was trained jointly with dynamic features and adaptive mask in this paper. In Algorithm 5, there is end-to-end speech enhancement network with ultra-lightweight channel attention.

Table 1: Average Perceptual Evaluation of Speech Quality (PESQ).

Noise	Signal-to-noise ratio (dB)	SRNN	SRU	GRU	CGRU
factory2	−5	2.346	2.515	2.426	2.495
	0	2.651	2.806	2.729	2.843
	5	2.877	3.102	3.065	3.127
buccaneer1	−5	1.861	1.952	1.763	1.952
	0	2.240	2.333	2.201	2.347
	5	2.533	2.654	2.587	2.668
destroyerengine	−5	1.903	2.030	2.106	2.228
	0	2.212	2.296	2.240	2.329
	5	2.521	2.578	2.526	2.600
hfchannel	−5	1.801	1.952	1.706	1.835
	0	2.157	2.177	2.111	2.285
	5	2.469	2.513	2.502	2.599

Table 2: Average Short-Term Objective Intelligibility of Speech (STOI).

Noise	Signal-to-noise ratio (dB)	SRNN	SRU	GRU	CGRU
factory2	−5	0.773	0.789	0.786	0.805
	0	0.852	0.868	0.870	0.888
	5	0.901	0.916	0.912	0.929
buccaneer1	−5	0.623	0.625	0.584	0.636
	0	0.745	0.748	0.720	0.760
	5	0.833	0.836	0.826	0.852
destroyerengine	−5	0.624	0.616	0.595	0.633
	0	0.749	0.752	0.734	0.767
	5	0.841	0.850	0.841	0.865
hfchannel	−5	0.655	0.655	0.645	0.671
	0	0.768	0.774	0.773	0.792
	5	0.847	0.862	0.861	0.879

(1) Comparing the results of Algorithm 1 and Algorithm 2, the enhancement performance of the input for the spliced features is better than that of the single feature, and the signs of the enhanced speech has improved by 1.3 dB on average, PESQ has improved by 0.12 dB on average, and STOI has improved by 0.03 dB, which verifies that the spliced features can better suppress the background noise.

(2) The comparison between the results of Algorithm 2 and Algorithm 3 shows that, after inputting new dynamic features and adaptive mask into the neural network, SegSNR improved by 0.99 dB on average, PESQ improved by 0.45 dB on average, and STOI improved by 0.02 dB. Experimental results proved the effectiveness of the combination of dynamic features and adaptive ratio mask, and the joint

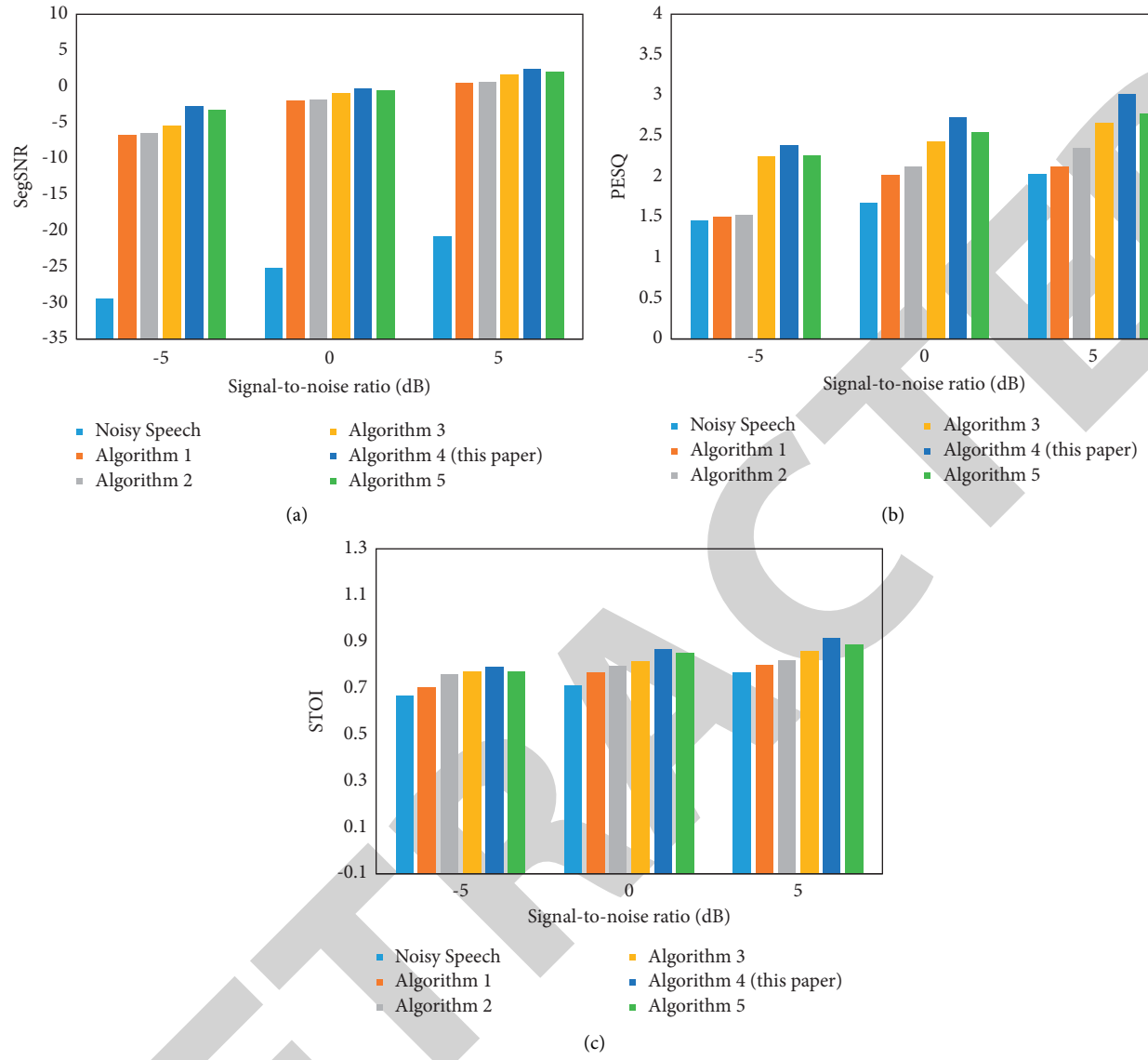

Figure 8: Comparison of three results of SegSNR, PESQ, and STOI under different algorithms in white-noise environment. (a) SegSNR; (b) PESQ; (c) STOI.

optimization can obtain enhanced speech with less distortion and better listening. The joint optimization can yield enhanced speech with less distortion and better listening experience.

(3) The comparison between the results of Algorithm 3 and Algorithm 4 shows that the proposed CGRU depth model better models the correlation of noisy speech signals with the same dynamic features and adaptive mask, and the output of the current moment of this network unit fuses the input and output of the previous moment. The problem of fixed time delay in real-time speech enhancement systems is solved, and better speech enhancement performance is obtained at the same time.

(4) The comparison between the results of Algorithm 4 and Algorithm 5 demonstrates that the dynamic features in this paper's method can comprehensively represent the speech nonlinear structure, and the integrity of the speech spectrum is maintained in combination with the adaptive mask that is automatically adjusted according to the difference of speech signal-to-noise ratio information. The speech enhancement performance of the deep CGRU model is further improved.

6. Conclusion

The dynamic feature joint adaptive mask optimization neural network speech enhancement algorithm is proposed. In this paper, a causal gated recurrent unit (CGRU) neural network is designed, solving the problem of real-time speech enhancement system. The dynamic features and adaptive mask are used as the input of the deep neural network to learn the complex mapping relationship between noisy

speech and pure speech in a supervised manner. The new features improve the neural network's ability to estimate the pure speech spectrum, and the new mask accurately represents the time-frequency masking value of each time-frequency unit. Finally, this paper gives the hardware implementation of the speech enhancement algorithm. The experimental results show that the algorithm can reduce the distortion of the enhanced speech under different noise and different signal-to-noise ratio conditions, and the speech quality and intelligibility are significantly enhanced with better enhancement effect. The next step will be to further improve the real-time performance of the speech enhancement system with the research goal of reducing the network complexity, and methods to speed up speech enhancement algorithms on DSP will be investigated.

Data Availability

The labeled datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work is supported by the General Project of Hunan Natural Science Foundation (no. 2018JJ2147), part of Youth Project of Hunan Natural Science Foundation (no. 2018JJ3203) or Project of Hunan Science and Technology Department (no. 2019ZK4018), and Hunan University of Science and Engineering Computer Application Special Subject Funding.

References

[1] N. Das, S. Chakraborty, J. Chaki, N. Padhy, and N. Dey, "Fundamentals, present and future perspectives of speech enhancement," *International Journal of Speech Technology*, vol. 24, no. 4, pp. 883–901, 2021.

[2] R. Jaiswal and D. Romero, "Implicit wiener filtering for speech enhancement in non-stationary noise," in *Proceedings Of The 2021 11th International Conference On Information Science And Technology (icist)*, pp. 39–47, IEEE, Chengdu, China, May 2021.

[3] Y. Hu, Y. Liu, S. Lv et al., "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," 2020, https://arxiv.org/abs/2008.00264.

[4] Y. Wang, H. Jia, and H. Ji, "Feature joint optimization of deep belief network for speech enhancement," *Computer Engineering and Applications*, vol. 55, no. 9, pp. 38–42, 2019.

[5] Z. Xu, S. Elshamy, and T. Fingscheidt, "Using Separate Losses for Speech and Noise in Mask-Based Speech enhancement," in *Proceedings Of The Icassp 2020-2020 Ieee International Conference On Acoustics, Speech And Signal Processing (icassp)*, pp. 7519–7523, IEEE, Barcelona, Spain, May 2020.

[6] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using f and recurrent neural networks," *IEEE Access*, vol. 8, pp. 160581–160595, 2020.

[7] N. Saleem and M. I. Khattak, "Multi-scale decomposition based supervised single channel deep speech enhancement," *Applied Soft Computing*, vol. 95, Article ID 106666, 2020.

[8] S. Chakrabarty and E. A. P. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[9] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, 2019.

[10] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *Proceedings Of The Icassp 2020-2020 Ieee International Conference On Acoustics, Speech And Signal Processing (icassp)*, pp. 871–875, IEEE, Barcelona, Spain, May 2020.

[11] F. Faraji, Y. Attabi, B. Champagne, and W. P. Zhu, "On the use of audio fingerprinting features for speech enhancement with generative adversarial network," in *Proceedings Of The 2020 Ieee Workshop On Signal Processing Systems (Sips)*, pp. 1–6, IEEE, Coimbra, Portugal, September 2020.

[12] D. Shah, T. Shah, and S. S. Jamal, "Digital audio signals encryption by Mobius transformation and Hénon map," *Multimedia Systems*, vol. 26, no. 2, pp. 235–245, 2020.

[13] Z. Yi-Fan and L. Qi-Bin, "Design Of aes/ebu audio transceiver system based on fpga," in *Proceedings Of The 2020 5th International Conference On Communication, Image And Signal Processing (ccisp)*, pp. 130–134, IEEE, Chengdu, China, December 2020.

[14] N. Pekez, A. Popović, and J. Kovačević, "Ethernet TCP/IP-based audio interface for DSP system verification," *IEEE Consumer Electronics Magazine*, vol. 10, no. 1, pp. 45–50, 2021.

[15] A. Pandey and D. L. Wang, "TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proceedings Of The icassp 2019-2019 Ieee International Conference On Acoustics, Speech And Signal Processing (icassp)*, pp. 6875–6879, IEEE, Brighton, UK, Apirl 2019.