Hindawi
Mobile Information Systems
Volume 2022, Article ID 9926708, 12 pages
https://doi.org/10.1155/2022/9926708

Research Article

Multisensor Speech Enhancement Technology in Music Synthesizer Design

Jing Peng

Jiangxi University of Applied Science, Nanchang, Jiangxi 330000, China

Correspondence should be addressed to Jing Peng; 202131080056@mail.bnu.edu.cn

Received 21 April 2022; Accepted 3 June 2022; Published 24 June 2022

Academic Editor: Jiguo Yu

Creating music through sound synthesis is the most representative electronic music creation method, and electronic music is actually the result of sound synthesis technology. Today, the field of electronic music encompasses multiple areas such as recording, mixing, composing, and producing. It also has some advantages over traditional music composition. Voice is the most effective and direct way of communication between people. And with the explosive development of speech recognition technology, the recognition rate of speech recognition systems in the near field environment has been greatly improved. However, in practical applications, there is often a large amount of ambient noise. If these environmental noises are strong, it will seriously affect the quality, accuracy, and speed of music synthesis. This greatly reduces not only the sound quality and clarity of speech but also the speed of speech recognition. To solve these problems, this paper proposes a multisensor speech enhancement technique and implements a multisensor speech enhancement system. It also proposes an enhancement method based on speaker speech and microphone speech. In this paper, the low-frequency harmonic components of the bone conduction signal are used to replace the frequency points disturbed by wind noise to reduce the influence of wind noise on speech quality and intelligibility. The experimental results show that the PESQ and MOS scores of the improved algorithm in this paper are 1.65 and 3.67, respectively. Compared with the existing methods, it has a great improvement. This can effectively improve the voice quality of the music synthesizer and reduce background noise.

1. Introduction

Today's music synthesizers are designed in different ways. It is the product of the fusion of the development of modern electronic technology and traditional music. The production of its sound is based on the principle of sound signals produced by electronic oscillators. The original sound is processed through digital processing technology to simulate the effect of sound in different propagation environments. Composers can imitate natural sounds or create new electronic sounds. The control method of the synthesizer is also controlled by the original single music keyboard control method and developed into fingerboard control and wind control. The sound field environment in actual production, work, and life is often complex. Harsh sound field environments such as mechanical noise interference, reverberation interference in conference rooms, and noisy human voice interference have become the main factors that reduce the recognition accuracy of the speech recognition system and damage the call quality of the speech communication system. The recognition performance of industry-leading speech recognition systems in such a sound field environment is also not robust and reliable. This also greatly limits the application range of speech recognition systems, so it is very important to study speech enhancement technology for music synthesizers.

Regarding music synthesizers, related scientists have done the following research. Altman M put an amplifier chip and speakers on the board in order for the builder to get audible results immediately after soldering the kit. He encoded the output stereo audio channel using pulse width modulation. Each channel has a low-pass filter. It consists of a resistor and a capacitor and converts the signal into audio [1]. Nishikawa et al. proposed a multichannel receiver with a process that can process all the signals by a single hardware. The experimental results of an electronic drum without any connecting wires fully demonstrate the feasibility of a self-powered wireless transmission system with a delay of 700 microseconds [2]. Soni and Makharia work aimed to make innovative designs of musical synthesizers highly adaptable, scalable, and highly miniaturized. Full octave notes are generated and tested through a complex algorithmic implementation of digital logic blocks. He investigated the area, power, and timing constraints of logic devices. He carried out consumption and time constraints, and the design enables automatic octave generation as well as manual key tone generation [3]. Pinch brought the academic research of technology research and sound research into dialogue. In technology studies, he discussed social construction methods for influential technology, emphasizing the role of keyboard standardization and the key role users play in the development of this technology. Sounds can be accepted by the user as some sounds stabilize, while others fail to stabilize [4]. Roche et al. proposed a new method for sound transformation based on control parameters. These parameters are intuitive and relevant to the musician. He used a variant autoencoder model. It is trained in an unsupervised manner on a large data set of synthesizer sounds [5]. At present, the main problems of music synthesizers are that the noise cannot be removed well, and the sound quality is not high. In order to solve these problems, this paper introduces multisensor and speech enhancement technology to study the problems existing in music synthesizers.

Regarding speech enhancement technology and multisensor, relevant scientists have done the following research. To further improve the robustness of speech activity detection, Li and You proposed a speech-based reinforcement method. The Laplacian distribution is used to model the residual noise, since the residual noise in the enhanced speech satisfies the Laplacian distribution. Experimental results showed that his proposed method performs better than baseline methods, especially under low SNR and nonstationary noise conditions [6]. Xue et al. proposed a vision-centric multisensor fusion framework for traffic environment perception methods for autonomous driving. The framework consistently fuses camera, lidar, and GIS information through geometric and semantic constraints for efficient self-localization and obstacle perception. His empirical results verified its robustness and efficiency [7]. Seeberg et al. developed a multisensor technology to detect three main classical subtechnologies, namely diagonal, kicked sculls, and double sculls. Other subtechnologies are classified as miscellaneous. The system works well on outdoor snow in different conditions. The algorithm he implemented was validated by video analysis [8]. Aiming at the incompleteness and uncertainty of information in single-parameter diagnosis of complex systems, Liu et al. proposed a new method of multisensor information fusion fault diagnosis based on BP neural network and evidence theory. It realizes the fault location and diagnosis of the main components of the hydraulic drive servo system and effectively improves the reliability of the system [9]. It is instructive that he compared tracking performance to the best multisensor solutions, both with and without missing samples. The Subedi et al. analysis evaluated the performance bounds of the two schemes for sparse-aware multisensor multitarget tracking algorithms. He also showed that recursive learning structures outperform traditional methods when the measurement vector is corrupted by missing samples and additive noise [10]. Xing and Xia focused on the distributed joint Kalman filter fusion problem for a class of multisensor unreliable network systems with uncorrelated noise. He proposed an optimal algorithm for unreliable network systems without buffers and gave two simulation examples to illustrate the effectiveness of his proposed method [11]. The purpose of Gomes et al. was to compare the performance of a modified RGB camera with a multisensor camera in obtaining the normalized difference vegetation index of coffee growing areas. He used a multispectral camera with five sensors and another camera with only one sensor. It turns out that the NDVI obtained with the multisensor camera is closer to the NDVI obtained with the GreenSeeker NDVI sensor [12]. The above studies provide a detailed analysis of speech enhancement techniques and applications of multisensor and music synthesizers. It is undeniable that these studies have greatly promoted the development of the corresponding fields. We can learn a lot from methodology and data analysis. However, speech enhancement techniques and multisensor research on music synthesizers are relatively rare, so it is necessary to fully apply these techniques to research in this field.

In this paper, a multisensor-based speech enhancement system is constructed, and the SNR and the segmented SNR average are obtained. The fused speech SNRs are −0.8, 3.96, 8.11, 12.62, and 15.29. It trains an improved speech enhancement algorithm using the Deep Noise Suppression Challenge data set. The PESQ, STOI, and the number of parameters of our method are 2.52, 78.1%, and 89 k, respectively. It combines the front-stage array processing algorithm and the post-filtering algorithm to form the entire dual-microphone noise reduction system. It separates the four noise reduction systems from the test set and computes the corresponding scores. The PESQ and MOS scores of the method in this paper are 2.26 and 3.88, respectively, which are more significant than the existing methods. It improves the voice quality of music synthesizers and reduces background noise.

2. Music Synthesizer Design Method

Music synthesizers are used to create, modify, and apply sine waves. People then feed it into sound generators and speakers to produce specific sounds. The quality of sound is determined by the composition of harmonics [13, 14]. Sound card music synthesizers create music that mimics the effects of many musical instruments. The job of a music synthesizer is to turn information into music. Electronic music is an important part of electroacoustic music. The electronic piano opened up a new world of music, but the advent of electronic music synthesizers took electronic music to a new level [15]. The basic block diagram of the synthesizer is shown in Figure 1.

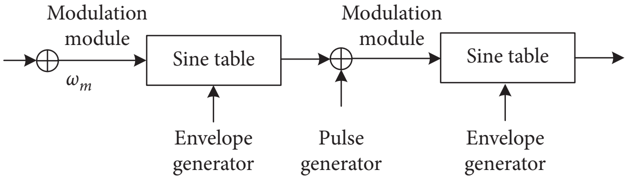

Figure 1: Synthesizer basic block diagram.

A digital music synthesizer, also known as a digital sound synthesizer, is a digital music device. It is an instrumentation amplifier and speaker that convert the electrical signal it produces into sound. In fact, the concept of digital synthesizer is larger. Its multifingered use of samples is a modern synth that reproduces the sounds of real instruments and analog synths. It is also the most common and most commonly encountered one. The biggest feature of digital music synthesizers is that they can generate new electronic sounds, which are commonly known as sounds by musicians in the industry. The main components of common digital music synthesizers on the market today are musical keyboards, physically controllable sliders, and knob buttons. Therefore, it is also named as a music control synthesizer.

Music synthesizers can imitate various sounds in reality and nature, such as pianos, electronic musical instruments, and flutes. There are also other musical instruments, human voices, and natural sounds, such as ocean waves. Traditional analog music synthesizers use the electronics of a signal generator to create sounds composed of elements of different frequencies. It is then up to the individual user to modify the sound to suit the characteristics of time, space, and terrain. Digital music synthesizers, on the other hand, use a straightforward digital approach to synthesizing waveforms and converting them into sound data. The oscillating sound waveform of a digital music synthesizer can be directly sampled by traditional methods, but must be implemented mainly by mathematical calculation methods.

The music synthesizer has three sounding modes: one is to directly change the voltage type, such as an analog synthesizer. The second is to use computers to make models of mathematical operations, such as software synthesizers. The third is a combination of the above two types of synthesizers, and finally, a granular digital synthesizer that generates a voltage signal to vibrate the membrane gasket of the speaker or earphone.

In the field of sound synthesis, in a broad sense, the tool used to synthesize sound is called a synthesizer. These can include the vocal system of the singer and all the instruments. Pianos, cellos, and flutes are often seen as “natural,” while synthesizers are seen as “artificial.” In fact, in essence, any musical instrument uses synthesis to create sound. In a narrow sense, synthesizers are used to synthesize surreal sounds that do not exist in nature. So pure imitation of sounds in nature is considered to be just the playback of sounds produced by objects in nature. The control surface and the synthesis engine are the two basic modules of an electronic music synthesizer. The control surface is used to set the parameters that define and control the synthesized sound. The synthesis engine is used to convert the sound synthesis parameters into audio signals. The waveform generation circuit is shown in Figure 2.

An oscillator in a synthesizer is the block used to generate the basic waveform audio signal. It can generally generate three basic waveforms: sine wave, sawtooth wave, and square wave. Some oscillators can only generate sine waves. There are also oscillators that produce simple variations of the three basic waveforms, such as a sharper sawtooth wave, a slightly rounder square wave, and so on. But they are usually relatively simple waveforms. In addition to the base waveform, the oscillator also contains amplitude and frequency parameters. Most synthesizers typically contain at least one or two oscillator blocks. They can generate one sound or multiple sounds at the same time according to the user’s needs or use the signals from several oscillators as modulation signals to modulate other oscillators during frequency modulation synthesis. Some synthesizers replace oscillators with wavetables or audio samples. In fact, we can also use this other sounding device as a variant of the oscillator or an advanced version of the oscillator.

The frequency modulation synthesis technology is a modulation method that makes the frequency of the high-frequency oscillating wave change according to the law of the modulation signal. The actual frequency modulation synthesizer mainly needs two operation units, the modulation unit and the carrier unit. Each unit has three functional modules: pulse generator, envelope generator, and sine table. Each unit has the ability to generate a sine wave. The frequency of the sine wave depends on the pulse generator, while the amplitude depends on the envelope generator. The output of the modulation unit is used to modulate the carrier unit. The resulting output waveform of the carrier element contains harmonics. The amplitude of the harmonics depends on the amplitude of the modulating waveform. By changing the frequency and amplitude of the modulation unit, the sound quality of the synthesizer can be changed.

If it wants to automatically periodically change a parameter in the synthesizer, then it needs to use the LFO. Low frequency oscillators are different from what are commonly called oscillators. Instead of an audio signal, it produces a modulated signal that changes periodically. But the value of its amplitude describes not the volume, but the value of its output to the modulated object. So how this value affects the sound depends on what is being modulated. The oscillation frequency of the LFO determines the rate of change of the value output to the modulated object. The main parameters of the LFO include waveform selection and oscillation frequency. The degree of influence on the modulation object depends on the modulation value in the modulation matrix, so this modulation value is equivalent to a relative amplitude.

In terms of mechanical fundamentals, the sound produced by a digital music synthesizer is uniquely different from the natural sound of a recording device. People usually say that the recording is to convert the mechanical energy contained in the sound wave into a sound wave signal, and the information can be converted into mechanical energy by

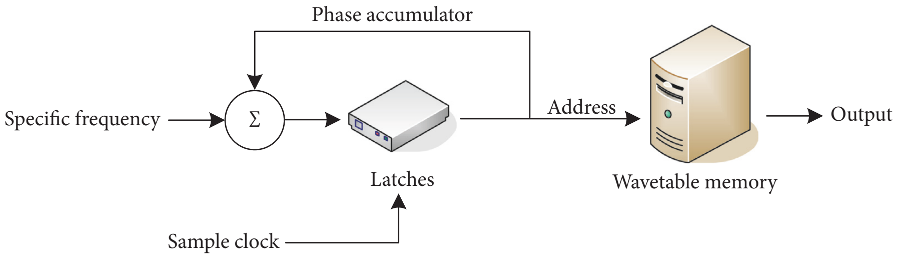

Figure 2: Waveform generation circuit.

playing the sound. The digital music synthesizer usually uses the keyboard as the control interface, so it is often regarded as a keyboard instrument device. But in fact, the control interface of the synthesizer is not necessarily the keyboard. It also includes, for example, fretboard controllers, guitar chord controllers, air vent controllers, and electronic drums.

The dynamics of volume mainly refers to musical expressions in traditional instrumental music performance and is an important factor to strengthen musical expressiveness. If it is a linear sound like a string legato, the musical expression becomes even more important. In the performance of real instruments, even if the player does not deliberately do it, musical expressions will naturally be attached. So a lot of times this musical element is less noticeable, especially to amateurs. But it is an element that needs extra attention when creating electronic music through sound synthesis. Because the volume is not designed, the oscillator generates a sound signal with absolutely stable volume without any change by default. In the absence of other design elements, such sounds are often prone to dullness and even auditory fatigue. And if it is designed, it will get more new effects than traditional music.

Digital audio technology is moving forward with the development of DSP and computer. Among them, electronic music synthesis technology occupies a very important position. Since the development of music synthesizers, speech quality, background noise, and how to extract clean original speech have become an urgent problem to be solved. Therefore, this paper introduces multisensor and speech enhancement technology to solve these problems.

Beamforming is a very important concept for speech amplification with microphone clusters. It is also an important research area in cluster signal processing. Beamforming technology can be used to amplify the speech signal in the direction of the target sound source and attenuate the interference and noise in other directions, so as to effectively achieve the goal of speech enhancement. Assuming that the direction of the target signal is different from the direction of the noise signal, it is usually necessary to perform some analysis procedures on the multiple speech signals collected by the microphone array, such as weighting, time division, and summation. Accordingly, the main branch of the beamforming pattern formed by the microphone array is aligned with the target speech signal, and the null branch is aligned with the target sound source, thereby helping to suppress the secondary branch. The direction of the beam and the width of the main beam depend on the number of microphone arrays, the distance and spacing between the arrays, the incident angle of the sound source, the sampling frequency, etc. The direction and frequency of the target source determine the response of the transducer. Figure 3 shows the process of improving speech quality through speech enhancement techniques.

Various speech enhancement algorithms based on air conduction speech sensors do not perform very well in strong noise environments. This is because air-conduction-based voice sensors such as microphones are sensitive to noise, and both useful information and noise are transmitted through the air. If it is considered that the useful information is not transmitted through the air, and the noise is transmitted through the air, the noise can be effectively isolated, and the speech transmission performance can be improved. Generally speaking, instead of directly using the voice input by the non-air-conduction sensor for voice processing or voice recognition, it use the voice transmitted by the non-air-conduction sensor. It generally needs to be processed together with air-conducted speech signals to combine their advantages. Or when the non-air-conducted sensor voice is used, it can be processed to improve its voice quality.

As a physical process, speech is embodied and perceived in both visual and auditory aspects. Speech enhancement is an effective way to deal with noise pollution. Its main purpose is to improve speech quality by removing as much noise as possible from the speech signal at the receiving end. Speech enhancement is not only related to traditional signal processing theory, but also closely related to speech characteristics and human ear perception characteristics. Therefore, these characteristics should be integrated, and an appropriate speech enhancement method should be selected according to the actual situation.

For nonadditive noise, homomorphic filtering is generally used to suppress or eliminate it. The convolutional homomorphic system is divided into three subsystems. Two feature subsystems that only depend on the combination rules of the signals. A linear subsystem depends on processing requirements. The first subsystem performs the operations to convert the convolutional signal into an additive signal. The second subsystem is an ordinary linear system that satisfies the principle of linear superposition, which is used to linearly transform the additive signal. The third subsystem is the inverse transform of the first subsystem, which inversely transforms the additive signal into a convolutional signal.

According to the signal of the non-air-conduction voice sensor, the voiced and unvoiced segments are determined,

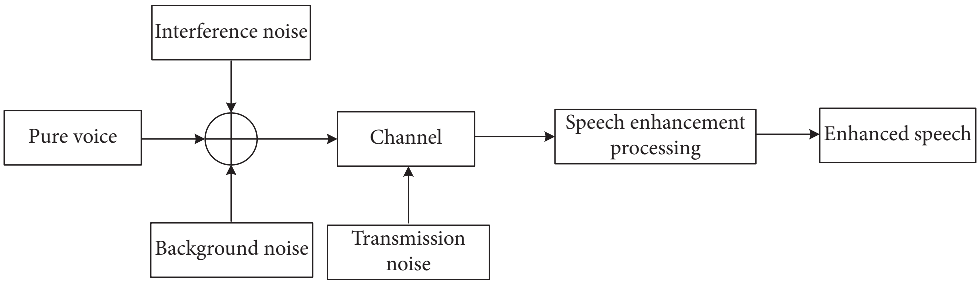

Figure 3: The process of improving speech quality through speech enhancement technology.

and the determined voiced segment mark is applied to the voice of the air conduction voice sensor to extract the voice signal. The difference between ordinary air conduction voice sensor and non-air-conduction voice sensor input voice in the case of noise: in the case of noise, the spectrum of the voice of the air conduction voice sensor is very messy and irregular. It combines the eigenvectors of the two signals. In the environment of various signal-to-noise ratios, it still has considerable advantages compared with the voice of the single-channel noisy air conduction sensor. It builds a joint model of air-conducted speech sensors detecting speech and non-air-conducting speech sensors detecting speech. This paper utilizes non-air-conducted speech sensors to detect speech during augmentation to accurately estimate the acoustic noise model.

$$\begin{aligned} n(b) &= \sum_{u=1}^{A} q_u(b) m_u(b - \psi_u), \\ n(b) &= \sum_{u=1}^{A} q_u m_u(b - \psi_u). \end{aligned} \tag{1}$$

A– number of microphone elements, ψ_u– delay compensation from delay estimation, $q_u(b)$– weighting coefficient of each array element

$$\begin{aligned} m_u(b - \psi_u) &= d_u(b) + B_u(b), \\ n(b) &= d(b) + \frac{1}{A} \sum_{u=1}^{A} B_u(b). \end{aligned} \tag{2}$$

$d_u(b)_{\mathrm{b}}$ – the voice signal received by the array element, q_u– weight

$$\begin{aligned} n_d(b) &= \sum_{k=1}^{A-1} q_u(b) U(b - k), \\ q_k(b + 1) &= q_k(b) + \varpi n(b) U(b). \end{aligned} \tag{3}$$

$n_d(b)$– output signal, q_u– the step size of the adaptive filter

$$E(f_a) = \sum_{b=1}^{J} M_b(\hat{f}_m) M_b{}^G(\hat{f}_m). \tag{4}$$

J– the number of subbands, f_m– center frequency

$$M(k, v) = \sum_{b=0}^{B_q - 1} q(b) m(b + vP). \tag{5}$$

k– frequency, v – frame, P– the length of a frame

$$\begin{aligned} \sum_{v} q(b - vP)\tau(b - vP) &= 1, \\ n(k) &= s(k, \cos\theta_s) M(k) + v(k). \end{aligned} \tag{6}$$

θ – angle of desired signal, $n(k)$– microphone observation signal vector

$$\begin{aligned} M_{fs}(k) &= M(k) g^G(k) s(k, \cos\theta_s), \\ g(k) &= [G_1(k) G_2(k) \ldots G_A(k)]^T. \end{aligned} \tag{7}$$

G– conjugate transpose, $g(k)$ – the weight vector of the beamformer

$$\begin{aligned} V_{rb}(k) &= g^G(k) v(k), \\ \phi_Z(k) &= \phi_{M_{fs}}(k) + \phi_{V_{rb}}(k). \end{aligned} \tag{8}$$

V– residual noise

$$\begin{aligned} \phi_{M_{fs}}(k) &= \phi_M(k) \left| g^G(k) s(k, \cos\theta_s) \right|^2, \\ \phi_{V_{rb}}(k) &= g^G(k) \lambda_v(k) g(k). \end{aligned} \tag{9}$$

$\phi_{M_{fs}}(k), \phi_{V_{rbs}}(k)$ – variance, $\lambda_v(k)$– correlation matrix

$$\begin{aligned} g^G(k) s(k, \cos\theta_s) &= 1, \\ S[g(k)] &= \frac{\left| g^G(k) s(k, 1) \right|^2}{g^G \vartheta_{0,\pi}(k) g(k)}. \end{aligned} \tag{10}$$

$S[g(k)]$– directivity factor b

$$\begin{aligned} C^G(k, \theta_s, \theta_{WB}) g(k) &= u_c, \\ \bar{l}(k, v) &= \varepsilon_l \bar{l}(k, v - 1) + (1 - \varepsilon_l) l(k, v). \end{aligned} \tag{11}$$

ε_l- Smoothing factor.

Through various vibrations (such as throat, head, ear canal, etc.), the reed in the sensor is deformed, and the vibration of the reed is converted into an electrical signal to obtain a voice signal. The reeds of non-air-conduction sensors are not affected by sound waves conducted in the air and therefore do not deform. Since air-conducted sound is not felt, non-air-conduction sensors are highly resistant to interference. Since non-air-conducted sensor speech is more robust in noisy environments and is highly correlated with air-conducted sensor speech, they have received increasing attention for robust speech processing applications. However, due to the limitations of transmission channels and devices, the quality of the voices they collect is not high.

Speech enhancement methods are based on a single microphone, including spectrum restoration, Wiener filtering, Kalman filtering, waveform transformation, etc. These methods employ network signal processing techniques for speech enhancement. They used spatial phase information of speech signals from multiple microphones to spatially filter incoming speech and form a directional spatial beam.

Excluding the influence of sensor quality and hardware circuit, regarding the characteristics of speech recorded by non-air-conduction sensors, most obviously, the waveform amplitude of speech has a certain attenuation in the time domain. From the frequency domain point of view, if the sampling frequency is 8 kHz, the voice of the non-air-conduction sensor has a large attenuation in the middle and high frequency. The frequency range of its speech is about 0–2 kHz, while the frequency range of air conduction sensors such as microphones is 0–4 kHz. The high-frequency components correspond to the detail components of the sound signal. As a result, non-air-conduction sensor speech has a blurred pitch and a lot of detail missing. Non-air-conducted sensor speech is noisy at low frequencies. This part is related to the characteristics of the sensor itself, and the other part is related to external interference or power supply during the signal acquisition process. Figure 4 shows the block diagram of the speech reconstruction system.

According to the signal of the non-air-conduction voice sensor, the voiced and unvoiced segments are determined. In this paper, the identified voiced segment marks are applied to the voice of the air conduction voice sensor to extract the voice signal. The difference between ordinary air-conducted speech sensor and non-air-conducted speech sensor recorded speech in the case of noise: iIn a noisy environment, the frequency spectrum of the air speech sensor's speech is not clean and irregular. On the other hand, the frequency spectrum of the voice of the external voice sensor in the air is very clean and is basically not affected by external noise. This feature can extract speech sequences from airborne speech sensors in noisy environments and distinguish audible and inaudible sequences based on the signal from the airborne speech sensor.

In addition to utilizing the non-air-conducted speech sensor to assist the air-conducted speech sensor in speech detection, the signal of the non-air-conducted speech sensor is also incorporated as a feature into the air-conducted speech sensor's speech. It modifies the parameters of the joint model accordingly, and then uses the revised joint model to enhance the voice detected by the input air conduction voice sensor. When modifying, the model compensation technology can be used to modify the parameters of the channel parameter joint model.

3. Music Synthesizer Design Experiment

Due to the strong anti-interference ability of the throat microphone, in a strong noise environment, the voice recorded by the microphone has strong noise. Simultaneously recorded voice from the throat microphone is very low noise. However, the high-frequency energy of its voice is very low, so that the recorded voice seems depressed and unnatural. In this paper, the voice of the throat microphone is subjected to spectrum spread processing. This fills in the high-frequency gaps for later fusion processing with the microphone voice. As shown in Figure 5, it is a multisensor-based speech enhancement system.

As shown in Figure 6, it is a waveform diagram of the speech after sub-band fusion of the microphone speech after spectral subtraction enhancement and the throat microphone speech after spectral expansion. It can be seen that the speech noise after fusion is smaller than that of the microphone enhanced by spectral subtraction, and the spectrum is clearer than that of the microphone enhanced by spectral subtraction.

As shown in Table 1, the average SNR and segmented SNR are shown. It can be seen that the SNR and SNR of the fused speech are higher than those of the microphone speech enhanced only by spectral subtraction. On the whole, the SNR of most of the fused speech is higher than that of the throat microphone.

The overall signal-to-noise ratio of a given noisy speech y is 10 dB, 5 dB, 0 dB, and −5 dB, respectively. And when −10 dB, it simulates the voices of four speakers, respectively, and the steps are the same as the above. It finally averages the signal-to-noise ratio and segmental signal-to-noise ratio of the four people. As shown in Table 2, it is the average value of speech SNR and segment SNR.

As shown in Figure 7, each algorithm deals with point source noise and diffuses field noise. It can be seen from the figure that all the improved algorithms can better improve the sound quality and intelligibility of speech under the interference of point source noise. However, under the interference of nonpoint source noise, the improvement of speech sound quality and intelligibility by the algorithm only using the air conduction microphone are very limited.

The logarithmic magnitude spectrum estimator can effectively reduce the musical noise phenomenon caused by the postprocessing algorithm. Figure 8 shows a block diagram of the spectral estimator processing. It can be seen from the figure that it mainly performs signal modeling under the assumption of additive noise and performs short-time Fourier transform on the input signal for analysis in the time-frequency domain.

This paper uses the deep noise suppression challenge data set to train an improved speech enhancement algorithm. The performance on the test set is shown in Table 3. It can be seen that the assumption in traditional signal processing methods (that is, the stationarity of noise is much higher than that of speech) greatly limits the ability of the algorithm to deal with speech disturbed by nonstationary noise. However, the improved algorithm in this paper has learned the processing modes of speech disturbed by various stationary and non-stationary noises in the training phase, which can effectively model the noise and effectively suppress such noises.

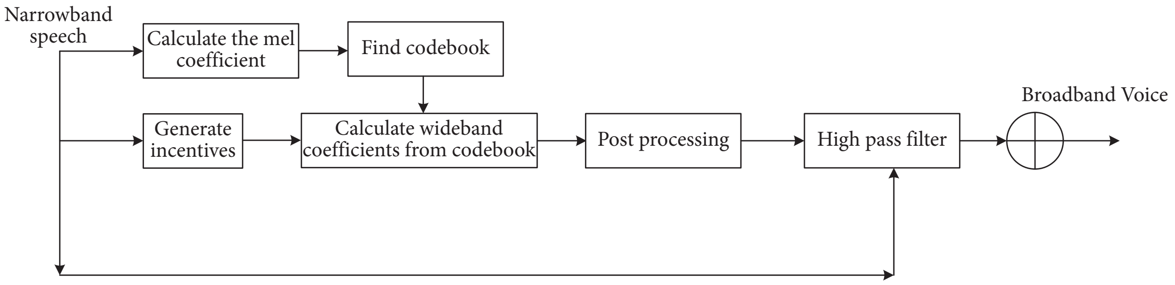

Figure 4: Block diagram of speech reconstruction system.

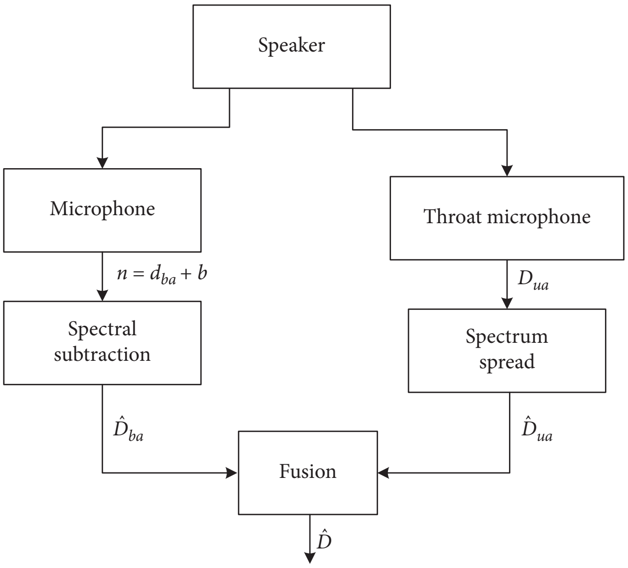

Figure 5: Multisensor-based speech enhancement system.

In this paper, the prearray processing algorithm and the post-filtering algorithm are combined to form the entire dual-microphone noise reduction system. It separately processes the test set for these four noise reduction systems and computes the corresponding scores. At the same time, it scores the processed speech by subjective listening test. Figures 9 and 10 show the scoring results of the low signal-to-noise ratio test set.

The experimental results show that the noise reduction system combined with the improved algorithm in this paper can effectively improve the voice quality, intelligibility, and subjective listening experience of speech. Since the bone vibration sensor has a greater advantage of being free from external sound field interference under the condition of low signal-to-noise ratio, the algorithm in this paper is also more significantly improved than other alignment algorithms under the condition of low signal-to-noise ratio. This improves the voice quality of the music synthesizer and reduces background noise.

4. Discussion

It performs speech enhancement processing in harsh sound field environments. The traditional microphone array and post-filter speech enhancement technology have the following three defects: (1) the blocking matrix of the microphone array speech enhancement algorithm based on GSC is prone to the leakage of the desired speech signal when the speech energy is high. This in turn causes the final adaptive noise canceller to falsely suppress speech components that need to be preserved. Under the interference of nondirectional noise, it is difficult to effectively suppress the noise by an adaptive noise canceller controlled only by the signal-to-interference ratio. (2) The beamforming algorithm cannot effectively suppress the wind noise in the low frequency part, and the wind noise will greatly damage the sound quality and intelligibility of the speech. (3) The traditional signal processing post-filtering algorithm is difficult to estimate and suppresses the residual non-stationary noise after the prestage beamforming algorithm. Aiming at these problems, this paper adds a bone vibration sensor based on the traditional two-microphone array to collect the bone conduction signal generated by the speaker. It greatly enhances the robustness of the algorithm through an adaptive process controlled by bone conduction signals. This paper improves a simple cyclic denoising neural network, which reduces the scale of network parameters while ensuring the denoising effect.

In the application scenario of wireless headsets, wind noise often seriously affects the quality of calls and greatly

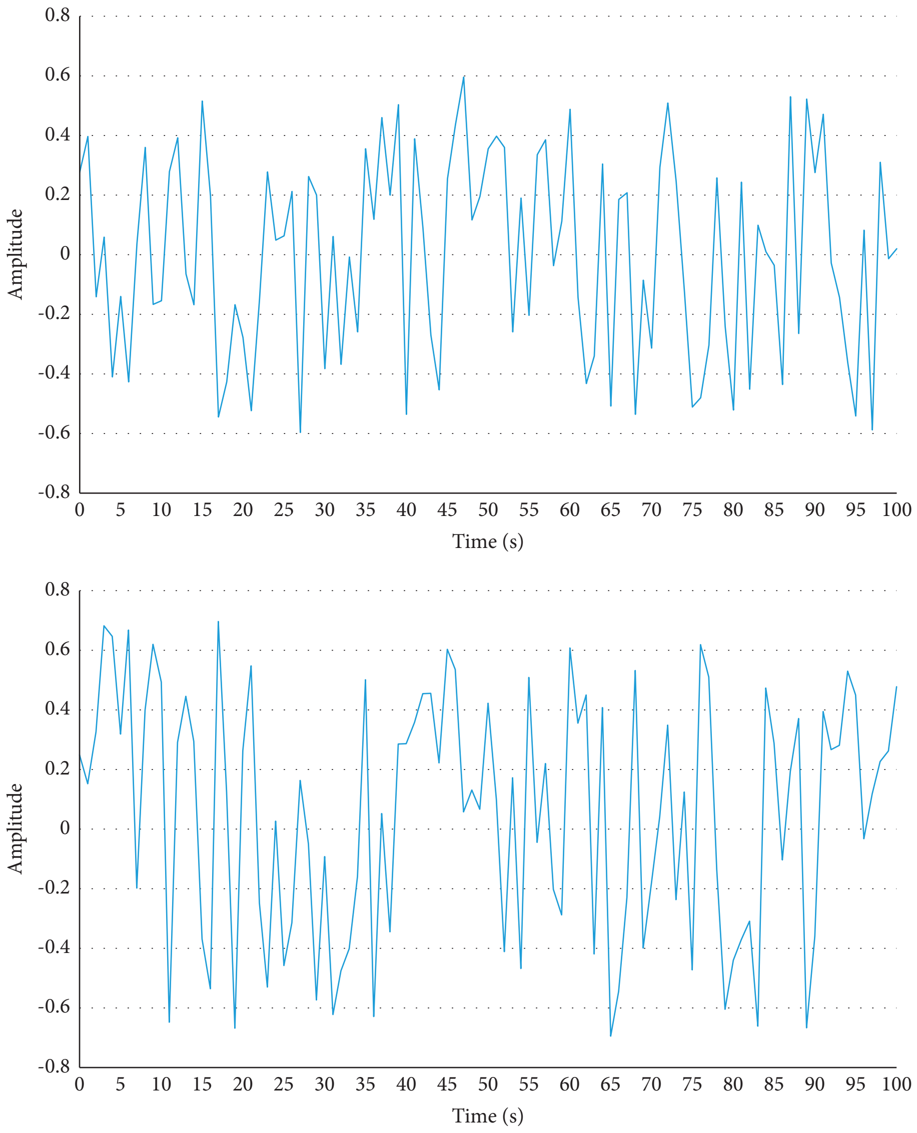

Figure 6: Waveform of the fused speech.

Table 1: SNR and segment SNR average.

	SNR/segment SNR (dB)				
Noisy microphone voice	−10/−17.25	−5/−12.43	0/−7.43	5/−2.43	10/2.66
Throat microphone voice			−0.97/−2.38		
Microphone speech after spectral subtraction enhancement	−5.86/−12.08	−1.2/−8.7	3.97/−3.18	8.45/1.26	12.68/6.92
Fused voice	−0.8/−7.87	3.96/−3.23	8.11/1.82	12.62/5.15	15.29/9.14

Table 2: Speech SNR and segment SNR average.

	SNR/segment SNR (dB)				
Noisy microphone voice	−10/−17.25	−5/−12.43	0/−7.43	5/−2.43	10/2.66
Throat microphone voice			−0.97/−2.38		
MMSE-enhanced post-microphone speech	−0.82, −5.46	5.06/−1.89	9.21/2.89	12.59/6.89	16.36/10.74
Fused voice	3.16/−2.41	6.8/0.96	10.3/4.38	13.7/7.15	16.27/10.59

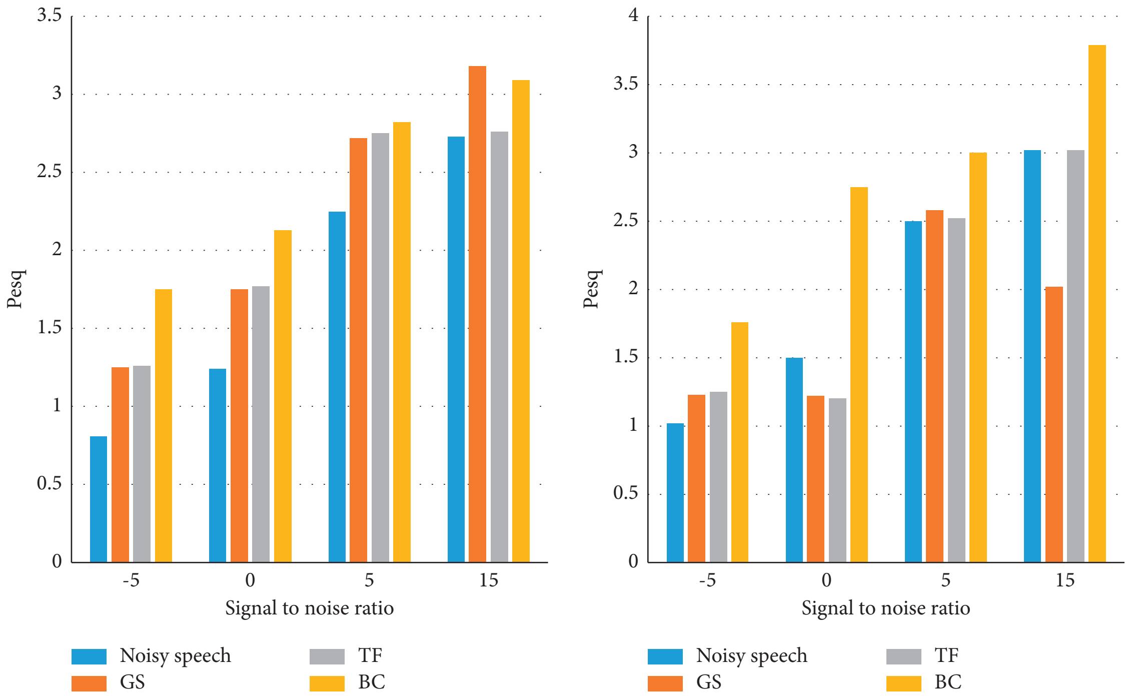

Figure 7: Each algorithm handles point source noise and diffuse field noise.

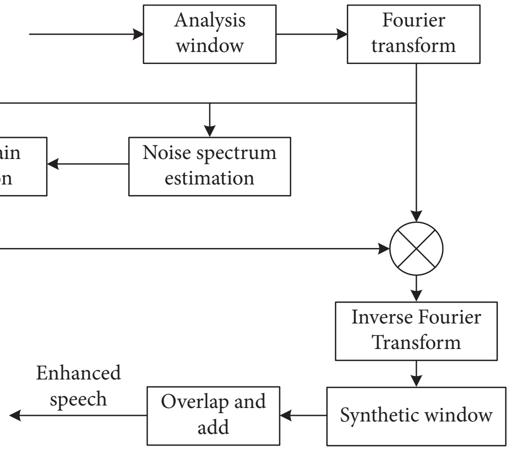

Figure 8: Spectral estimator processing block diagram.

Table 3: Performance on the test set.

	Noisy speech	OA	Rnnoise	This article
PESQ	1.76	1.77	2.34	2.52
STOI (%)	76.4	75.9	76.5	78.1
Number of parameters	None	None	86 k	89 k

harms the user experience. The main energy of wind noise is concentrated in low frequencies, which has a greater impact on low-frequency speech harmonics. If it wants to automatically periodically change a parameter in the synthesizer, then it needs to use the LFO.

After the original waveform has been modulated to give its contours and shape, it can be output to the effects processor for the next stage of shaping. After the sound is output to the effector, the effector makes some specific changes to the sound through a series of simple or complex processing. The exact change depends on the function of this effect. The effect device can be understood as simplifying, prefabricating, and packaging the production process of some commonly used effects and representative effects into a

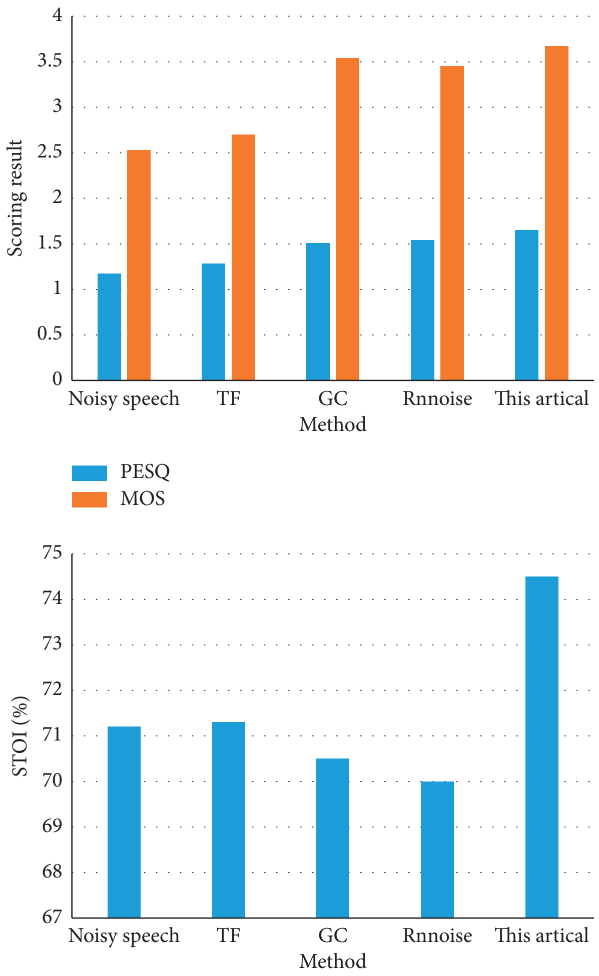

Figure 9: Scoring performance on the test set in the 0 dB to 5 dB SNR range.

module specially used to form this effect through electronic technology. For example, the original delay effect was simulated by repeatedly reading and playing audio from the same tape. The initial distortion effect is caused by too much current to the audio signal output from the speakers. The original reverb effect can only be obtained by relying on natural sound reflections. Sound effects that may have been complicated or difficult to achieve can now be easily achieved. There are many types of effects, depending on their function. The more common ones include delay, reverb, distortion, etc. Filters and equalizers can also belong to effects. In addition to these effects that imitate natural physical phenomena, there are other effects that are used to produce special effects.

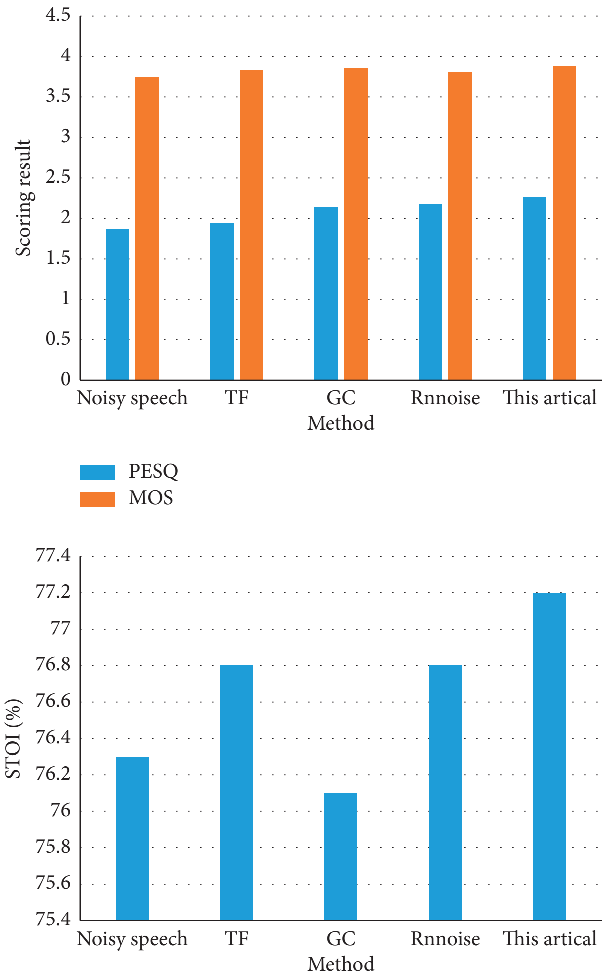

Figure 10: Scoring performance on the test set in the 5 dB to 10 dB SNR range.

5. Conclusion

With the rapid development of information and recording technology, digital sound processing technology has gradually replaced analog sound processing technology and is currently developing rapidly. Human audio processors face increasing challenges in terms of sound quality, volume, and functionality. Electronic design automation technology is a technology that can automatically design electronic systems or products. Speech recognition technology is widely used in applications in various industries, such as mobile phone voice assistants, automatic dialogue customer service robots, industrial intelligent control terminals, and military fields. The speech recognition system enables the computer to have the ability of speech recognition through the joint modeling of the phoneme model of the speech signal and the language model. In order to solve the problems of voice quality and background noise of music synthesizers, a multisensor-based voice enhancement system is constructed in this paper. It effectively reduces the background noise of the music synthesizer and improves the voice quality. For the study of fusing the voice spectrum of the throat microphone and the voice spectrum of the microphone, due to the time relationship, this paper proposes and simulates a fusion method. In addition to the method proposed in this paper, there are many other methods to try. The research on the weight function can also be more in depth.

Data Availability

The data 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 research study was sponsored by 2021 Jiangxi Province Art Planning Project, Research on Jiangxi Guqin Music Culture (YG2021029). The authors thank the project for supporting this article!

References

[1] M. Altman, "Ardutouch: an Arduino-compatable synthesizer: digital signal processing squeezed into an easy-to-build kit - [Resources_Hands on]," *IEEE Spectrum*, vol. 55, no. 12, pp. 21-22, 2018.

[2] H. Nishikawa, Y. Shimizu, K. Igarashi, A. Tanaka, and T. Douseki, "Batteryless and wireless electric drum system using piezoelectric generator with power and signal source," *IEEJ Transactions on Sensors and Micromachines*, vol. 137, no. 12, pp. 455–461, 2017.

[3] M. Soni and P. Makharia, "Implementation of innovative low cost music synthesizer using fpga," *ICTACT Journal on Microelectronics*, vol. 3, no. 2, pp. 394–397, 2017.

[4] T. Pinch, "From technology studies to sound studies: how materiality matters[J]," *Epistemology & Philosophy of Science*, vol. 56, no. 3, pp. 123–137, 2019.

[5] F. Roche, T. Hueber, M. Garnier, S. Limier, and L. Girin, "Make that sound more metallic: towards a perceptually relevant control of the timbre of synthesizer sounds using a variational autoencoder," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, pp. 52–66, 2021.

[6] J. Li and D. You, "Enhanced speech based jointly statistical probability distribution function for voice activity detection," *Chinese Journal of Electronics*, vol. 26, no. 2, pp. 325–330, 2017.

[7] J. R. Xue, D. Wang, S. Y. Du, Dx. Cui, Y. Huang, and Nn Zheng, "A vision-centered multi-sensor fusing approach to self-localization and obstacle perception for robotic cars," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 122–138, 2017.

[8] T. M. Seeberg, J. Tjønnås, O. M. H. Rindal, P. Haugnes, S. Dalgard, and O. Sandbakk, "A multi-sensor system for automatic analysis of classical cross-country skiing techniques," *Sports Engineering*, vol. 20, no. 4, pp. 313–327, 2017.

[9] B. J. Liu, Q. W. Yang, and W. U. Xiang, "Application of multi-sensor information fusion in the fault diagnosis of hydraulic system," *International Journal of Plant Engineering and Management*, vol. 22, no. 1, pp. 12–20, 2017.

[10] S. Subedi, Y. D. Zhang, and M. G. Amin, "Cramer–Rao type bounds for sparsity-aware multi-sensor multi-target tracking," *Signal Processing*, vol. 145, pp. 68–77, 2017.

[11] Z. Xing and Y. Xia, "Distributed federated kalman filter fusion over multi-sensor unreliable networked systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 10, pp. 1714–1725, 2016.

[12] A. P. A. Gomes, D. M. D. Queiroz, D. S. M. Valente, FdAdC. Pinto, and J. T. F. Rosas, "Comparing a single-sensor camera with a multisensor camera for monitoring coffee crop using unmanned aerial vehicles," *Engenharia Agrícola*, vol. 41, no. 1, pp. 87–97, 2021.

[13] I. B. Gorbunova and K. Y. Plotnikov, "Music computer technologies in education as a tool for implementing the polymodality of musical perception," *Musical Art and Education*, vol. 8, no. 1, pp. 25–40, 2020.

[14] H. W. Park and M. Bae, "A study on voice enhancement using palm reflections," *Journal of the Acoustical Society of America*, vol. 144, no. 3, p. 1926, 2018.

[15] R. A. Sowah, A. R. Ofoli, S. N. Krakani, and S. Y. Fiawoo, "Hardware design and web-based communication modules of a real-time multi-sensor fire detection and notification system using fuzzy logic," *IEEE Transactions on Industry Applications*, vol. 53, no. 1, pp. 559–566, 2017.