IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 19, 2019, accepted May 15, 2019, date of publication May 22, 2019, date of current version July 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918147

Fusion Feature Extraction Based on Auditory and Energy for Noise-Robust Speech Recognition

YANYAN SHI[1], JING BAI[1], PEIYUN XUE[1], AND DIANXI SHI[2,3]

[1]College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China
[2]Artificial Intelligence Research Center (AIRC), National Innovation Institute of Defense Technology (NIIDT), Beijing 100071, China
[3]Tianjin Artificial Intelligence Innovation Center (TAIIC), Tianjin 300457, China

Corresponding authors: Jing Bai (bj613@126.com) and Dianxi Shi (dxshi@nudt.edu.cn)

ABSTRACT Environmental noise can pose a threat to the stable operation of current speech recognition systems. It is therefore essential to develop a front feature set that is able to identify speech under low signal-to-noise ratio. In this paper, a robust fusion feature is proposed that can fully characterize speech information. To obtain the cochlear filter cepstral coefficients (CFCC), a novel feature is first extracted by the power-law nonlinear function, which can simulate the auditory characteristics of the human ear. Speech enhancement technology is then introduced into the front end of feature extraction, and the extracted feature and their first-order difference are combined in new mixed features. An energy feature Teager energy operator cepstral coefficient (TEOCC) is also extracted, and combined with the above-mentioned mixed features to form the fusion feature sets. Principal component analysis (PCA) is then applied to feature selection and optimization of the feature set, and the final feature set is used in a non-specific persons, isolated words, and small-vocabulary speech recognition system. Finally, a comparative experiment of speech recognition is designed to verify the advantages of the proposed feature set using a support vector machine (SVM). The experimental results show that the proposed feature set not only display a high recognition rate and excellent anti-noise performance in speech recognition, but can also fully characterize the auditory and energy information in the speech signals.

INDEX TERMS Cochlear filter cepstral coefficients, Teager energy operators cepstral coefficients, principal component analysis, speech recognition.

I. INTRODUCTION

Speech is the material shell and acoustic representation of language, and is one of the most easily accessible carriers of information for humans. It is a vital component of research in the field of human-computer interaction and intelligent communication due to its ability to convey various information sources. Speech recognition is a technology for realizing intelligent human-computer interaction with broad application prospects and value. Its main purpose is to communicate with a computer, so that the computer can convert the speech signal into corresponding text or commands through the process of understanding and recognition. Interpretation of human spoken language through technology has a diverse range of applications including in air transport, intelligent homes, disaster rescue, medical diagnostics, and other human-computer interaction fields [1]. At present, research into speech recognition is mainly focused on feature extraction and pattern recognition. As an important element of speech recognition, feature extraction has a large influence on the performance of the system [2]. Therefore, methods to extract the most information-capable, noise-less, easily classified, and stable new features from speech must be developed. In order to achieve integrity of speech information, strategies to integrate and optimize the different types of features that have been proposed also require further research to establish a set of anti-noise speech features with the best classification performance.

Currently, the most widely effective speech features are based on the auditory characteristics of the human ear. The human ear has good anti-noise attributes, and an increasing number of researchers are studying these auditory features to establish a speech feature model which is more consistent with human auditory characteristics [3]. Mel frequency cepstral coefficient (MFCC) is the most popular feature currently, and the majority of research is optimized for this feature. However, studies have shown that the recognition performance of MFCC features in low signal-to-noise (SNR)

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

ratio environments are not ideal, resulting in poor stability of the speech recognition system [4]. In addition, MFCC is extracted based on Fourier transform, though Fourier transform is only suitable for the processing of stationary signals. In view of the non-stationary time-varying characteristics of speech signals, Peter Li proposed an auditory-based transform to process audio signals, and used this method to extract cochlear filter cepstral coefficients (CFCC) features for robust speaker identification [5], [6]. Studies have shown that this auditory-based transform compensates for the shortcomings of Fourier transform, with the advantages of less harmonic distortion and better spectral smoothness [7].

Recent development in CFCC features have been successfully applied within various applications [8]–[13]. Li and Huang [8] applied the study of CFCC features for robust speaker identification under mismatched conditions, and compared this with the traditional feature parameters such as MFCC, perceptual linear predictive (PLP), and relative spectral perceptual linear prediction (RASTA-PLP). Xin and Changchun of Beijing University of Technology described a blind bandwidth extension method based on CFCC, in which bandwidth extension of wideband audio signals achieved superior wideband audio auditory quality [9]. Li *et al.* added Teager energy operator (TEO) to CFCC features and proposed TEO-CFCC feature parameters for speaker recognition in noisy environments [10]. Patel T B combined CFCC and derivative features with MFCC for detection of natural vs. spoofed speech [11], [12], while Zuoqiang and Yong realized the fusion of phase information and CFCC in robust speaker identification systems [13]. Although some scholars have studied the CFCC feature parameter, there are very few studies focused on the auditory characteristics of the human ear and robustness of auditory features in speech signals.

Auditory mechanism studies have shown that nonlinear signal processing mechanisms in the cochlea contribute significantly to auditory production and robustness [14]. Traditional CFCC features only consider the use of cubic roots and logarithmic functions to complete the process of nonlinear loudness transformation, which does not fully conform to the auditory mechanism in the auditory system. Thus, from a physiological point of view, considering the saturation relationship between the spike rate of auditory neurons and sound intensity, new cochlear filter cepstral coefficients (NCFCC) are extracted by the power-law nonlinear function which can simulate the auditory characteristics of human ear. Speech enhancement technology is then introduced in the front-end of NCFCC feature extraction to reduce the influence of noise. That is, spectral subtraction, recursive least square, and least mean square are used as preprocessors to remove noise, respectively, to further improve the SNR of the speech signal. Three new robust feature parameters are then extracted: fusion feature based on power-law nonlinearity function and spectral subtraction (FFPSS), fusion feature based on power-law nonlinearity function and recursive least square (FFPRLS), and fusion feature based on power-law nonlinearity function and least mean square (FFPLMS). In addition, from the perspective of speech enhancement, the energy tracking transform characteristics of noisy speech are analyzed and Teager energy operator cepstral coefficient (TEOCC) is extracted. Combining the above robust features and their first order difference, TEOCC is then used to form fusion feature sets. Finally, principal component analysis (PCA) is applied to feature selection and optimization of the feature set to obtain the optimal feature set in order to achieve the best performance by the speech recognition system.

II. PROPOSED NCFCC FEATURE EXTRACTION

A. COCHLEAR FILTER CEPSTRAL COEFFICIENTS (CFCC)

The feature extraction procedure for CFCC consists of four parts: a series of cochlear filter banks model based on auditory transform, hair cell function, nonlinearity, and discrete cosine transform (DCT) [6]. The following subsection briefly describes auditory transform and procedure for estimating the CFCC and proposed NCFCC features.

1) AUDITORY TRANSFORM

As a new method of processing non-linear signals, auditory transform is equivalent to converting time-domain signals into frequency-domain signals through cochlear filter banks. The cochlear filter function is used as the basis function of the wavelet, completing the whole process of sound transmission from the outer ear to the basement membrane, with an existing inverse transform [8].

Let $\psi(t)$ be the impulse response of the basilar membrane of cochlear $\psi(t) \in L^2(R)$, in which the function $\psi(t)$ satisfies the following conditions:

①. It integrates to zero:

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \tag{1}$$

②. It is square integrable or has finite energy:

$$\int_{-\infty}^{+\infty} |\psi(t)\,dt|^2 < \infty \tag{2}$$

③. It satisfies:

$$\int_{-\infty}^{+\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega = C \tag{3}$$

where $0 < C < \infty$, and

$$\Psi(\omega) = \int_{-\infty}^{+\infty} \psi(t) e^{-j\omega t} d\omega \tag{4}$$

Let $f(t)$ be any square integrable function. The auditory transform of $f(t)$, with respect to $\psi(t)$ as the impulse response of the basilar membrane in the cochlea, is defined as:

$$T(a,b) = \int_{-\infty}^{+\infty} f(t)\psi_{a,b}(t)\,dt \tag{5}$$

where $\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)dt$, a, b are real, and a is a scale or dilation variable. By changing a, the central frequency of an impulse response function can be shifted. Subscript b is a time shift or translation variable. If a is known, $\psi_{a,0}(t)$ moves a unit along the time axis to get $\psi_{a,b}(t)$. Note that $1/\sqrt{a}$ is an energy normalizing factor. It ensures that the energy stays the same for all a and b, providing:

$$\int_{-\infty}^{+\infty} \left|\psi_{a,b}(t)\right|^2 dt = \int_{-\infty}^{+\infty} |\psi(t)|^2 dt \tag{6}$$

A typical cochlear impulse response function or cochlear filter can be defined as:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\frac{(t-b)^{\alpha}}{a}\exp\left[-2\pi f_L \beta\left(\frac{t-b}{a}\right)\right] \times\left[\cos 2\pi f_L\left(\frac{t-b}{a}\right)+\theta\right]u(t-b) \tag{7}$$

where $\alpha > 0, \beta > 0$, parameters α and β determine the frequency domain shape and width of the cochlear filter. Subscript α and β are taken as the generally empirical value, $\alpha = 3, \beta = 0.2$, $u(t)$ is the unit step function, and the value θ is the initial phase. The value of a can be determined by the current filter, the central frequency f_c, and the lowest frequency f_L of the cochlear filterbank, which is denoted as:

$$a = f_L / f_c \tag{8}$$

2) OTHER OPERATIONS IN CFCC EXTRACTION

As an important part of the auditory system, human cochlear inner ear hair cells transform the vibration signals transmitted from the basement membrane of the cochlea into analyzable nerve impulse signals of the brain, and then transmit them to the auditory nerve fibers [15]. The following nonlinear function of hair cell describes this motion:

$$h(a,b) = [T(a,b)]^2 \tag{9}$$

where $T(a,b)$ is the filterbank output of speech signal $f(t)$.

The hair cell output of each filterbank is converted into a representation of the nerve spike count density in a duration associated with the current band central frequency, which is computed as:

$$S(i,j) = \frac{1}{d}\sum_{b=1}^{l+d-1} h(i,b), \quad l = 1, L, 2L\cdots; \forall i,j \tag{10}$$

where $d = max\{3.5\tau_i, 20ms\}$ is the window length, τ_i is the period of the i band. $\tau_i = 1/f_c$, and L is the window shift duration.

The output of the above formula is further applied to scales of loudness functions as cubic root nonlinearity, providing:

$$y(i,j) = [S(i,j)]^{1/3} \tag{11}$$

Finally, discrete cosine transform (DCT) is applied to decorrelate the feature dimensions. It generates the cochlear filter cepstral coefficients as a new auditory-based speech feature, which is computed as:

$$cfcc(i,n) = \sqrt{2/M}\sum_{m=1}^{M-1} y(i,m)\cos\left(\frac{\pi n(m-1/2)}{M}\right) \quad 0 \le m \le M \tag{12}$$

where M is the number of filters.

B. NEW COCHLEAR FILTER CEPSTRAL COEFFICIENTS (NCFCC)

The CFCC is a feature parameter that simulates the auditory characteristics of the human ear. It imitates the basilar membrane function of the human ear auditory system by auditory transform, and uses cubic root function or logarithmic function to complete the non-linear transformation process [15]. Although the importance of auditory nonlinearity has been confirmed in many studies, the impact of peripheral nonlinearity remains less understood. Existing research shows that the non-linearity of the response of the auditory system to the sound signal obeys exponential compression, and gradually increases from low frequency to high frequency [16]. However, cubic root nonlinearity or logarithmic nonlinearity function compresses the entire frequency domain. This nonlinearity does not fully conform to the auditory mechanism in the auditory system, which causes some issues. Hence, in an actual speech recognition system, the extraction method of CFCC feature parameters is not ideal in a low SNR environment. The power-law nonlinear function can be roughly approximated to the relationship curve between spike rate and sound intensity, and the characteristics of the power-law nonlinear function are consistent with human auditory, that is, the dynamic characteristics of output are not entirely dependent on the magnitude of input [17]. Therefore, NCFCC features obtained by using the power-law nonlinear function to simulate human ear auditory characteristics are effective. The equation of this function used in the experiment is as follows:

$$y(i,j) = [S(i,j)]^{0.101} \tag{13}$$

III. SPEECH ENHANCEMENT AND ROBUST FEATURE EXTRACTION

A. PROPOSED FFPSS FEATURES

1) SPECTRAL SUBTRACTION

Spectral subtraction is a simple method used in the frequency domain estimation speech enhancement algorithm. Its principle is that the power spectrum of pure speech signal can be obtained by subtracting the power spectrum of noise from the power spectrum of speech signal with noise [18].

Let $y(n)$ be speech signal with noise, $s(n)$ is pure speech signal, $d(n)$ is noise, and the relationship between them is:

$$y(n) = s(n) + d(n), \quad 0 \le n \le N-1 \tag{14}$$

where n the data points, and N is frame length.

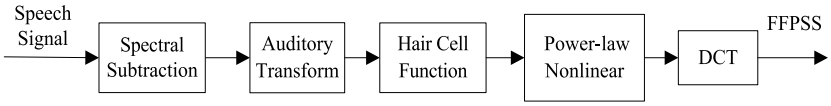

FIGURE 1. Flow diagram of the extraction process of FFPSS feature parameters.

Their representation in the fourier transform domain is given by:

$$Y(\omega) = S(\omega) + D(\omega) \tag{15}$$

As speech is assumed to be uncorrelated with background noise, the short-term power spectrum of $y(n)$ has no cross-terms, hence:

$$E\left|\left|S(\omega)\right|^2\right| = E\left|\left|Y(\omega)\right|^2\right| - E\left|\left|D(\omega)\right|^2\right| \tag{16}$$

where $S(\omega)$, $D(\omega)$, $Y(\omega)$ is the short-term power spectrum of $s(n)$, $d(n)$, and $y(n)$.

For a short-time stationary process in a frame, use:

$$|S(\omega)|^2 = |Y(\omega)|^2 - \lambda_d(\omega) \tag{17}$$

in which $\lambda_d(\omega)$ is the statistical average of silent segment $|D(\omega)|^2$. Therefore, the amplitude of the speech signal after spectral subtraction can be expressed as:

$$\hat{S}(\omega) = [|Y(\omega)|^2 - E(|D(\omega)|^2)]^{1/2} = [|Y(\omega)|^2 - \lambda_d(\omega)]^{1/2} \tag{18}$$

2) FFPSS FEATURES EXTRACTION

Spectral subtraction is introduced in the front-end of feature extraction to suppress background noise and further improve the clarity of the speech signal. The speech signal is pre-processed first in a process which includes pre-emphasis, endpoint detection, and frame windowing. Formula (16) is then used to subtract the spectrum amplitude of noise from the spectrum amplitude of the noise signal, providing the spectrum amplitude of pure signal. Based on the phase insensitivity of speech, the phase angle information before spectral subtraction is directly used to reconstruct the signal after spectral subtraction to obtain the denoised speech. Finally, the denoised speech signal is extracted using a power-law nonlinear function which simulates the auditory characteristics of human ears, and a new feature parameter FFPSS is obtained. The extraction process is illustrated in fig. 1.

B. PROPOSED FFPRLS FEATURES

1) RECURSIVE LEAST SQUARE

Recursive least square (RLS) is an adaptive filtering method [19]. The filter has two inputs. The first is input speech signal, which is represented by $x(n) = [x(1), x(2), \ldots, x(N)]$. The other input is the expected output signal, which is represented by $d(n) = \left[d(1), d(2), \ldots, d(N)\right]$. The impulse response of the filter is given by $[w(1), w(2), \ldots, w(N)]$, and the output of the filter is:

$$y(n) = \sum_{k=1}^{M} w_k(n) x(n-k+1), \quad n = 1, 2, \ldots N \tag{19}$$

where N is data length, M is the number of filters, and N must be greater than M. The RLS algorithm requires the sum of squares of all errors to be minimum [20]. Therefore, the error signal $e(n)$ is defined as:

$$e(n) = d(n) - y(n) = d(n) - w^T x(n) \tag{20}$$

where $w = [w_1, w_2, \ldots, w_M]^T$, the weighted sum of squares of $e(n)$ is expressed as:

$$\varepsilon(n, w) = \sum_{i=0}^{n} \lambda^{n-i} |e(i)|^2 \tag{21}$$

in which λ is defined as the forgetting factor or weighting factor, $0 \leq \lambda < 1$, and λ is to give new and old data different weights to ensure the fast response ability of the algorithm to changes in the input process.

To minimize the sum of weighted squares, the recursive formula of RLS can be expressed as:

$$w(n) = w(n-1) + R_x^{-1}(n) x^T(n) e(n) \tag{22}$$

where $R_x(n) = \sum_{i=1}^{n} \lambda^{n-i} x(i) x^t(i)$, it can be seen that the RLS algorithm is based on the previous parameter estimation. the algorithm uses the recursive method to revise the results according to the new observation data so as to recursively deduce the new parameter estimation.

2) FFPRLS FEATURES EXTRACTION

The RLS adaptive filter is used as a pre-processor of the speech signal to denoise the signal. The enhanced signal is then processed by auditory transformation and hair cell function, and the power-law nonlinear function is used to simulate the auditory characteristics of human ears to complete the loudness conversion process. Finally, the FFPRLS features are obtained by DCT transformation. The extraction process is shown in Fig. 2.

C. PROPOSED FFPLMS FEATURES

1) LEAST MEAN SQUARE

The least mean square (LMS) algorithm is a commonly used adaptive filtering algorithm [21]. It carries out automatic adjustment of the current filter parameters according to the estimation of the filter parameters at the previous moment to adapt to the statistical characteristics of signal and noise changes, thus realizing optimal filtering.

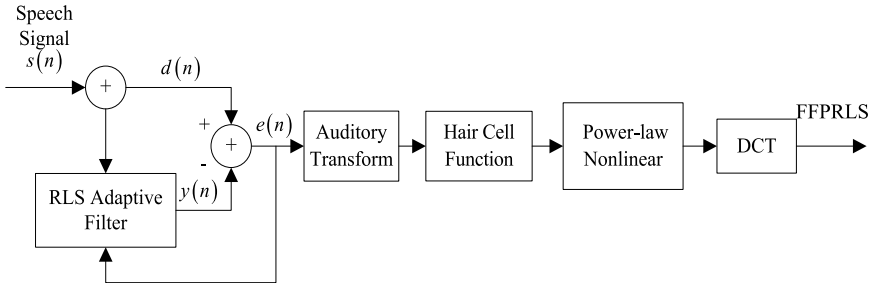

FIGURE 2. Flow diagram of the extraction process of FFPRLS feature parameters.

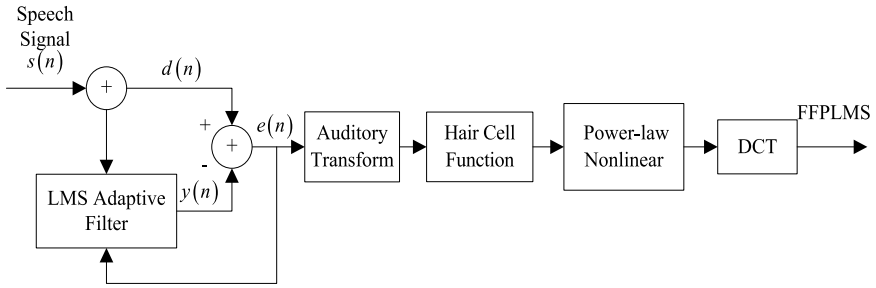

FIGURE 3. Flow diagram of the extraction process of FFPLMS feature parameters.

Let the formula $X_1(n), X_2(n), \ldots, X_m(n)$ be the input signal sequence, and $d(n)$ be the desired output signal. The error signal $e(n)$ is defined as:

$$e(n) = d(n) - \sum_{i=1}^{M} w_i x_i(n) \tag{23}$$

where w_i is the weight coefficient of the filter.

For convenience of study, the above formula (23) is represented as a vector, and the vector of input signal is defined as $X(n) = [x_1(n), x_2(n), \ldots, x_M(n)]^T$. The weight vector is given by $W(n) = [w_1(n), w_2(n), \ldots, w_M(n)]^T$, and formula (21) can be expressed as:

$$y(n) = W^T X(n) \tag{24}$$

$$e(n) = d(n) - y(n) = d(n) - W^T X(n) \tag{25}$$

The LMS algorithm uses the steepest descent method based on random gradient to find the optimal solution of weighted vector. Among them, the random gradient estimation is unbiased, and the random gradient estimation is unbiased.

The initial value of the weight vector is set and adjusted along the direction of negative gradient until the optimal value is found. The computation of the iteration formula used in this paper is as follows:

$$W(k+1) = W(k) - \mu \nabla(k) \tag{26}$$

where μ is a constant called the convergence factor, which can control the convergence speed and stability. It can be seen that the key two steps of LMS algorithm are calculating gradient $\nabla(k)$ and selecting convergence factor μ.

Here, the value of $\nabla(k)$ is roughly calculated by taking the error quadratic $e^2(k)$ as the estimated value of mean square error $E\left[e^2(k)\right]$. Hence, providing:

$$\hat{\nabla}(k) = \nabla\left[e^2(k)\right] = 2e(k)\nabla[e(k)] \tag{27}$$

where $\hat{\nabla}(k) = -2e(k)X(k)$, and the iteration formula is as follows:

$$W(k+1) = W(k) + 2\mu e(k) X(k) \tag{28}$$

The selection of convergence factor μ is derived from the updated formula of the weight coefficient vector.

When μ is satisfied, $0 < \mu < \frac{1}{\lambda_{\max}}$ tends to infinity, the weighted vector converges to the optimal wiener solution, and is given as:

$$\lim_{k \to +\infty} E\{W(k)\} = R_{xx}^{-1} R_{xd} \tag{29}$$

where λ is the maximum eigenvalue of the autocorrelation matrix R_{xx}.

2) FFPLMS FEATURES EXTRACTION

An LMS adaptive filter is used as the noise canceller of the speech signal, and the estimated value $y(n)$ of the output noise is as close as possible to the noise signal in $d(n)$. The FFPLMS features are then extracted by auditory transformation, hair cell function, the power-law nonlinear function, and dct from the signal after de-noising. The extraction process is illustrated in fig. 3.

IV. PROPOSED TEOCC FEATURE EXTRACTION

The TEO [22] is a type of non-linear difference operator proposed by kaiser. it has the characteristics of tracking the non-linear energy of a signal, can reasonably present the transformation of signal energy, and restrain the influence of zero-mean noise on speech signal to enhance the signal and extract features in speech recognition.

Let $x(n)$ be a discrete-time signal, and the definition of TEO is:

$$\psi[x(n)] = x(n)^2 - x(n+1)\,x(n-1) \tag{30}$$

where $\psi[x(n)]$ is output of TEO, $x(n)$ is the sampling value of the discrete signal at n point.

Let $x(n)$ be a speech signal with additive noise, $s(n)$ be pure speech signal, and $\omega(n)$ be zero-mean additive noise. This relationship can be Expressed as:

$$x(n) = s(n) + \omega(n) \tag{31}$$

The TEO of $x(n)$ is given by:

$$\psi[x(n)] = \psi[s(n)] + \psi[\omega(n)] + 2\tilde{\psi}[s(n), \omega(n)] \tag{32}$$

where $\tilde{\psi}[s(n), \omega(n)]$ is mutual teager energy of $s(n)$ and $\omega(n)$, and

$$\tilde{\psi}[s(n), \omega(n)] = s(n)\,\omega(n) - 0.5s(n-1)\,\omega(n+1) - 0.5s(n+1)\,\omega(n-1) \tag{33}$$

Both $s(n)$ and $\omega(n)$ are zero mean and independent of each other, providing:

$$E\left\{\tilde{\psi}[s(n), \omega(n)]\right\} = 0 \tag{34}$$

$$E\{\psi[x(n)]\} = E\{\psi[s(n)]\} + E\{\psi[\omega(n)]\} \tag{35}$$

Compared with TEO energy of pure speech signal, The TEO energy of noise can be neglected, according to:

$$E\{\psi[x(n)]\} \approx E\{\psi[s(n)]\} \tag{36}$$

Thus, TEO can eliminate the influence of zero-mean noise and achieve speech enhancement. The application of TEO in feature extraction can not only better reflect the energy change of speech signal, but also suppress noise and enhance speech signals, achieving good results in speech recognition.

The average TEO energy of each speech signal is calculated according to formula (30), and normalized and logarithmic data are obtained as:

$$\hat{\psi}[x(n)] = \log\left\{\psi[x(n)] \big/ \max(\psi[x(n)])\right\} \tag{37}$$

One-dimensional TEOCC is then obtained by dct transformation.

V. FUSION AND OPTIMIZATION OF FEATURE PARAMETERS

All speech features attempt to fully represent the complete characteristics of speech signals to achieve better recognition results. However, a certain kind of feature generally contains only part of the speech information, and the original feature parameters reflect the static characteristics of speech signals because human ears are more sensitive to dynamic parameters [23]. Therefore, the combination of dynamic and static feature parameters makes the dynamic and static information complementary, so as to more accurately describe the characteristics of speech. This paper is based on the above three robust features that are extracted, and their first-order difference. The three fusion feature sets are obtained by adding TEOCC which reflects the change of signal energy.

To reduce the storage of feature data and further obtain the optimal feature set, the fusion features are analyzed by principal component analysis to reduce dimension and recognition time, and further improve the performance of the recognition system. In the current research, PCA [24] is a statistical analysis method based on orthogonal transformation. This method uses covariance matrix to linearly combine the data with high correlation dimension into data with less correlation dimension. That is to say, fewer features are used to replace most of the information of the original features, so as to construct more representative feature vectors. Based on the fusion features sets above, this paper uses PCA algorithm to reduce their dimension, and obtains the principal component matrix by setting the threshold of cumulative contribution rate, in which the cumulative contribution rate reaches 97 %. The cumulative contribution rate is calculated by:

$$\alpha_p = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{38}$$

in which λ denotes the eigenvalue of each dimension, and p is the preceding p principal component.

VI. EXPERIMENT PREPARATION

A. KOREAN ISOLATED WORDS DATABASE

The isolated words database is used for performing isolated word recognition from speech signals. The vocabulary sizes used here are 10 words and 20 words. The corpus consists of 10 digits and 40 command words, with 16 speakers repeating each word three times. For the experiment in this study, recordings of the utterances of nine speakers are used as the training set, and the utterances of the remaining seven speakers are used as the test set.

B. CLASSIFICATION

Constructing a reasonable and efficient speech recognition model is the most important research challenge in the field of speech recognition technology. Currently, for speech recognition tasks, both linear and nonlinear classifiers are used. Researchers have experimented with different model classifiers for improving speech recognition. The most widely used classifiers for speech recognition are Hidden Markov Model (HMM) [25], Gaussian Mixture Model (GMM) [26], and Support Vector Machines (SVM) [27]. Among them, the theoretical reasoning and geometric description of SVM

TABLE 1. Comparison of recognition rates of CFCC features extracted from different functions in different SNR environments (%).

Vocabulary	Functions which simulating the auditory characteristics of human ears	SNR(dB)					Average
		0	5	10	15	20	
10	cubic root	64.76	63.81	64.76	67.62	68.10	65.81
	logarithmic	73.14	79.84	82.81	87.14	86.19	81.82
	power-law nonlinear	75.24	80.48	83.33	88.10	88.57	83.14
20	cubic root	59.43	59.61	61.69	61.64	62.24	60.92
	logarithmic	66.10	73.42	78.07	80.45	83.99	76.41
	power-law nonlinear	67.23	73.83	79.96	81.10	84.38	77.30

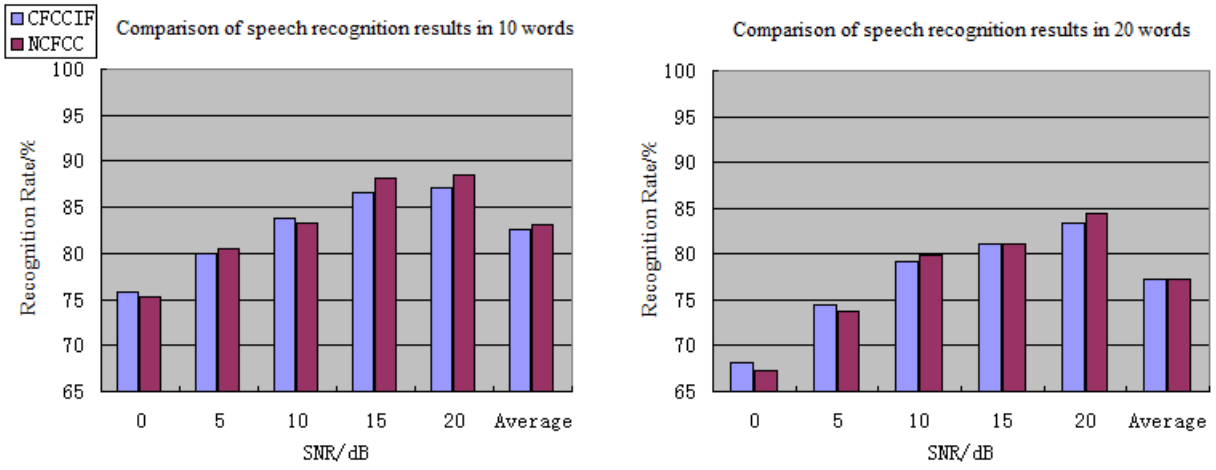

FIGURE 4. Comparison of speech recognition result of NCFCC and CFCCIF features.

are strict and intuitive, and its generalization ability is strong. This classifier can deal with the problems of high dimension inseparability and dimension disaster, which are difficult to solve in traditional machine learning, and realize the application of classification and regression in complex situations. To carry out the classification of complex non-linear speech data, this paper adopts SVM to map the original features into a high-dimensional space. The choice of kernel function is radial basis function (RBF).

VII. EXPERIMENTAL SETUP AND ANALYSIS OF RESULTS

To verify the validity and robustness of the proposed feature set, the following four experimental schemes are designed.

A. VALIDITY AND VERIFICATION OF ROBUSTNESS OF THE NCFCC FEATURE

Two experiments are designed to verify the validity of NCFCC feature parameters, and the results are then discussed.

Experiment 1: Traditional CFCC features are first extracted by the cubic root function or logarithmic function which can simulate the auditory characteristics of the human ear [8]–[11]. Next, the NCFCC feature parameters are extracted from the power-law nonlinear function proposed in this paper. Finally, the results of the two experiments are compared. The experimental results are provided in Table 1.

Experiment 2: To further test the recognition performance of NCFCC features in different SNR environments, the recognition rates of NCFCC and CFCCIF features proposed in reference [11] are compared. The experimental results are shown in Fig. 4.

It can be observed in Table 1 that for the task of speech recognition, a higher accuracy are extracted compared to traditional CFCC features. The robustness of the proposed system is obviously improved. The findings also confirm the feasibility and validity of the proposed feature NCFCC in isolated word speech recognition systems.

TABLE 2. Comparison of speech recognition based on five features (%).

Vocabulary	Features	SNR(dB)					AVERAGE
		0	5	10	15	20	
10	CFCCIF	75.71	80.00	83.81	86.67	87.14	82.67
	NCFCC	75.24	80.48	83.33	88.10	88.57	83.14
	FFPRLS	78.57	82.86	86.94	88.57	89.52	85.29
	FFPSS	79.48	85.71	88.10	89.18	89.93	86.48
	FFPLMS	81.90	85.71	89.52	90.48	90.48	87.62
20	CFCCIF	68.16	74.54	79.18	81.05	83.43	77.27
	NCFCC	67.23	73.83	79.96	81.10	84.38	77.30
	FFPRLS	68.83	77.18	82.18	82.87	84.73	79.16
	FFPSS	74.26	79.70	82.52	83.96	84.76	81.04
	FFPLMS	74.16	80.02	84.09	85.90	86.36	82.11

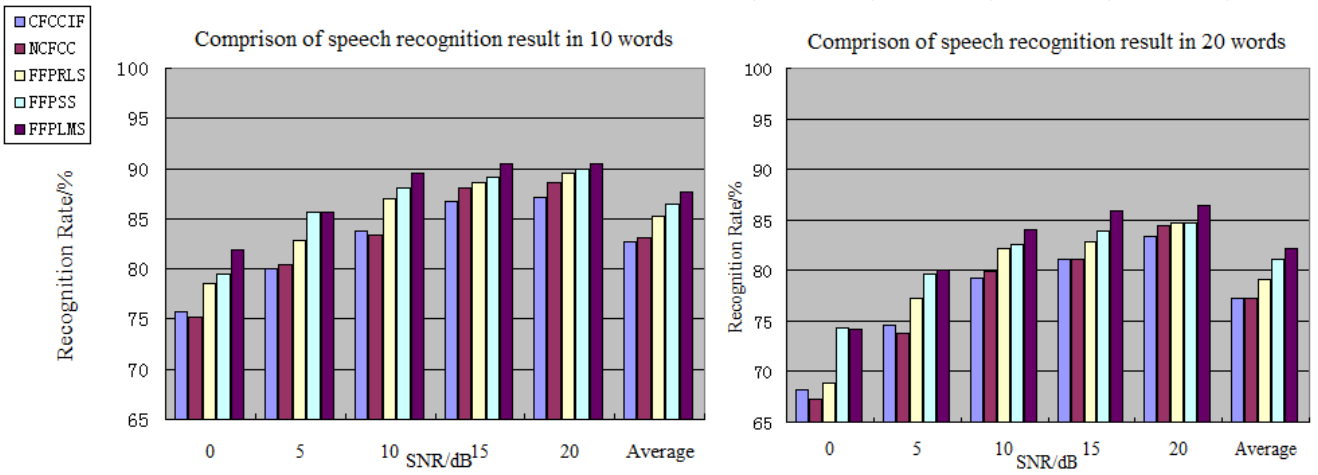

FIGURE 5. Comparison of experimental results.

According to the experimental results shown in Fig. 4, compared with CFCCIF, NCFCC features have a superior recognition effect. The overall speech recognition rate under 10 words is increased by 0.47 %. The increase is small under 20 words, and the overall speech recognition rate is increased by 0.03 %. This result proves that the power-law nonlinear function can better simulate the auditory characteristics of human ears to extract more representative speech features. However, it can also been seen in Fig. 1 that the recognition rate of the NCFCC features are lower than CFCCIF features when the SNR is 0 dB, illustrating that the recognition effect of the NCFCC features in a low SNR environment is not ideal.

B. VALIDITY AND VERIFICATION OF ROBUSTNESS OF THE FFPSS, FFPRLS AND FFPLMS FEATURES

In this paper, the CFCCIF features, NCFCC features, FFPSS features, FFPRLS features, and FFPLMS features extracted above are used as the input of SVM for speech recognition comparison experiments. The results of the recognition are provided in Table 2 and Fig. 5. These results confirm the validity and robustness of the three features based on speech enhancement.

Experimental results show that when using different vocabulary and different values of SNR, compared with the above five features, the recognition rate can be improved. From

TABLE 3. Comparison of speech recognition of FFPSS feature set (%).

Vocabulary	Experiment	Features	SNR(dB)					AVerage
			0	5	10	15	20	
10	Experiment 1	FFPSS	79.48	85.71	88.10	89.18	89.93	86.48
	Experiment 2	FFPSS_D	80.95	86.19	90.48	90.95	90.00	87.71
	Experiment 3	FFPSS_D+TEOCC	82.38	90.48	93.81	93.81	94.29	90.95
20	Experiment 1	FFPSS	74.26	79.70	82.52	83.96	84.76	81.04
	Experiment 2	FFPSS_D	74.86	80.30	82.71	86.03	87.34	82.25
	Experiment 3	FFPSS_D+TEOCC	79.52	83.83	86.80	89.01	90.50	85.93

TABLE 4. Comparison of speech recognition of FFPRLS feature set (%).

Vocabulary	Experiment	Features	SNR(dB)					AVerage
			0	5	10	15	20	
10	Experiment 1	FFPRLS	78.57	82.86	86.94	88.57	89.52	85.29
	Experiment 2	FFPRLS_D	79.52	86.19	88.57	89.18	90.00	86.69
	Experiment 3	FFPRLS_D+TEOCC	81.43	89.05	91.90	91.43	92.38	89.24
20	Experiment 1	FFPRLS	68.83	77.18	82.18	82.87	84.73	79.16
	Experiment 2	FFPRLS_D	73.56	80.30	83.27	86.41	87.90	82.29
	Experiment 3	FFPRLS_D+TEOCC	74.67	83.27	87.55	89.39	89.94	84.96

Table 2, it can be observed that NCFCC features have only slight speech recognition advantages over CFCCIF features. The three robust features FFPRLS, FFPSS, and FFPLMS extracted in this paper have higher recognition effect than NCFCC features. The average recognition rates are increased by 2.15 %, 3.34 % and 4.48 % for 10 words, and 1.86 %, 3.74 % and 4.81 % for 20 words, respectively. By comparing Table 2 with Fig. 5, it can be determined that the features extracted by combining speech enhancement with feature extraction have certain advantages in recognition rate and robustness. It further illustrates the potential of LMS adaptive filter as the preprocessor of speech signal, and the extracted FFPLMS features demonstrate better robustness.

C. VALIDITY AND VERIFICATION OF ROBUSTNESS OF FUSION FEATURE SET

The performance of discriminative features, as well as the dynamic and static information from speech signals on the basis of FFPRLS, FFPSS, and FFPLMS features are next analyzed. The first-order difference parameters are obtained and combined to form new mixed features, which are expressed by FFPSS_D, FFPRLS_D, and FFPLMS_D, and then used to identify isolated speech. Additionally, the above mixed features are combined with TEOCC to form fusion feature sets, which can use the recognition model SVM to classify speech signals. The experimental results are shown in Table 3, Table 4, and Table 5.

TABLE 5. Comparison of speech recognition of FFPLMS feature set (%).

Vocabulary	Experiment	Features	SNR(dB)					AVerage
			0	5	10	15	20	
10	Experiment 1	FFPLMS	81.90	85.71	89.52	90.48	90.48	87.62
	Experiment 2	FFPLMS_D	82.43	87.62	90.48	91.90	90.95	88.68
	Experiment 3	FFPLMS_D+TEOCC	82.86	90.95	92.38	94.29	94.29	90.95
20	Experiment 1	FFPLMS	74.16	80.02	84.09	85.90	86.36	82.11
	Experiment 2	FFPLMS_D	74.86	80.67	84.71	86.22	87.71	82.83
	Experiment 3	FFPLMS_D+TEOCC	79.70	83.83	86.80	89.76	89.94	86.01

TABLE 6. speech recognition result of optimized feature set (%).

Vocabulary	Experiment	Features	SNR(dB)					AVerage
			0	5	10	15	20	
10	Experiment 1	FFPLMS_D+TEOCC	82.86	90.95	92.38	94.29	94.29	90.95
	Experiment 2	PCA-Features	90.52	90.48	93.33	94.81	94.81	92.79
20	Experiment 1	FFPLMS_D+TEOCC	79.70	83.83	86.80	89.76	89.94	86.01
	Experiment 2	PCA-Features	83.33	86.67	88.81	90.95	92.38	88.43

Vertical analysis of Table 3, 4, and 5 illustrates the following:

(1) Comparing the recognition results of experiment 1 and experiment 2, it can be observed that the recognition rate of three dynamic and static combination features are higher than that of single static features in different SNR environments. The advantage of the FFPRLS_D features are more prominent in 0 dB 20 words and 15 dB 20 words, which are 4.73 % and 3.54 % higher than the FFPRLS feature, respectively. The average recognition rate of dynamic and static combination features are improved to varying degrees, which shows that the combination of dynamic and static features can more effectively represent the information of speech signals. The result provides further proof that the combination of features has an optimization effect on recognition performance compared with single feature.

(2) Comparing the recognition results of experiment 2 and experiment 3, it can be seen that after adding the TEOCC feature that embodies nonlinear energy characteristics, the recognition effect of the fusion feature set is further improved compared to that of dynamic and static combination features. From the average recognition rate, the recognition effect of three fusion feature sets are better than that of dynamic and static combination features. The fusion feature set FFPSS_D+TEOCC is 3.68 % higher than FFPSS_D. This result illustrates that TEOCC contains the effective information of speech signal and can be used as an auxiliary feature parameter to improve the performance of a speech recognition system.

(3) Comparison of experiment 1, 2, and 3 confirms that the three fusion feature sets proposed have the highest recognition rate compared with single static feature and dynamic-static combination features. Among them, the recognition rates of fusion feature set FFPSS_D+TEOCC and FFPLMS_D+TEOCC in the case of 20 dB 10 words are as high as 94.29 %. The result shows that the combination of

TEOCC feature reflects the non-linear energy characteristics of speech signals and eliminates the noise of the speech signal, thus improving the classification performance of speech recognition systems, and further verifying the effectiveness of feature fusion.

D. VALIDITY AND VERIFICATION OF ROBUSTNESS OF FUSION FEATURE SET OPTIMIZATION

To further confirm the validity of the optimized feature set, a recognition rate comparison experiment between the fusion feature set FFPLMS_D+TEOCC and the optimized feature set is carried out. The optimized feature set is defined as PCA-Features at the input of SVM, and the experimental results are provided in Table 6.

It can be observed from Table 6 that after the fusion feature is optimized by PCA, the recognition rate is improved. This is because PCA analysis can reduce the correlation between the feature parameters, retain the important components in the characteristics, and remove the redundant features. This analysis can also highlight the differences between the feature parameters, so that the performance of the speech recognition system is further improved.

VIII. CONCLUSION AND FURTHER STUDY

In view of the non-stationary time-varying characteristics of speech signals, this paper used auditory feature CFCC in noisy speech recognition systems, improved the non-linear transformation process of feature CFCC, and proposed a new feature to improve the operation of speech recognition systems. Due to the poor recognition performance of this feature in low SNR environments, three speech enhancement methods were introduced into the feature extraction process, then three robust features were extracted. In addition, due to the incomplete voice information represented by the existing features, the energy feature TEOCC was fused, and the compensation effect of the feature TEOCC on the auditory features was verified. Finally, based on the feature redundancy problem of fusion feature set, a feature optimization method of principal component analysis was proposed, and its effectiveness was verified. In future research, we would like to consider finding a better speech enhancement method combined with feature extraction to achieve better speech recognition performance. In addition, the study of more robust feature sets are also the future research direction that needs to be further explored.

REFERENCES

[1] A. A. Karpov and R. M. Yusupov, "Multimodal interfaces of human–computer interaction," *Herald Russian Acad. Sci.*, vol. 88, no. 1, pp. 67–74, Jan. 2018.

[2] G. Y. Wang, Y.-M. Zhang, M.-L. Sun, X. Wang, and Y. Zhang, "Speech signal feature parameters extraction algorithm based on PCNN for isolated word recognition," in *Proc. Int. Conf. Audio. Lang. Image Process. (ICALIP)*, Jul. 2017, pp. 679–682.

[3] L. Mu, Y. Peng, M. Qiu, X. Yang, C. Hu, and F. Zhang, "Study on modulation spectrum feature extraction of ship radiated noise based on auditory model," in *Proc. IEEE/OES China Ocean Acoust. (COA)*, Jan. 2016, pp. 1–5.

[4] S. Seyedin, S. Gazor, and S. M. Ahadi, "On the distribution of Mel-filtered log-spectrum of speech in additive noise," *Speech Commun.*, vol. 67, pp. 8–25, Mar. 2015.

[5] Q. Li, "An auditory-based transfrom for audio signal processing," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2009, pp. 181–184.

[6] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2010, pp. 4514–4517.

[7] G. Yang, "Cochlear filter cepstral feature in speech recognition," Ph.D. dissertation, Taiyuan Univ. Technol., Taiyuan, China, 2011.

[8] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 6, pp. 1791–1801, Aug. 2011.

[9] L. Xin and B. Changchun, "Bandwidth extension of audio signals based on cochlear filter cepstral coefficients," *J. Tsinghua Univ. (Sci. Technol.)*, vol. 53, no. 6, pp. 913–916, 2013.

[10] L. Li, D. An, D. Zhao, C. Rong, and S. Ma, "TEO-CFCC characteristic parameter extraction method for speaker recognition in noisy environments," *Comput. Sci.*, vol. 89, no. 2, pp. 118–121, 2013.

[11] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2015, pp. 2062–2066.

[12] T. B. Patel and H. A. Patil, "Cochlear Filter and Instantaneous Frequency Based Features for Spoofed Speech Detection," *IEEE J. Sel.Topics Signal Process.*, vol. 11, no. 4, pp. 618–631, Jun. 2017.

[13] L. Zuoqiang and G. Yong, "Robust speaker identification based on CFCC and phase information," *Comput. Eng. Appl.*, vol. 51, no. 17, pp. 228–232, 2015.

[14] S. Ramakrishnan, "Cochlear implant stimulation rates and speech perception," in *Modern Speech Recognition Approaches With Case Studies*. Hoboken, NJ, USA: Wiley, 2012, ch. 10. doi: 10.5772/2569.

[15] T. Haji and S. Kitazawa, "Acoustic analysis of certain consonants using a computed model of the peripheral auditory system.," *Nippon Jibiinkoka Gakkai Kaiho*, vol. 97, no. 11, pp. 2055–2064, Nov. 1994.

[16] Y. F. Ma, K. A. Chen, M. Ma, and C. Zhang, "A new cepstrum coefficients applied to acoustic target recognition," *ACTA Armamentarii*, vol. 30, no. 11, pp. 1477–1483, 2009.

[17] Y. Qianqian, Z. Ping, and J. Xinxing, "The auditory feature extraction algorithm based on power-law nonlinearity function," *Microelectron. Comput.*, vol. 6, pp. 163–166, 2015.

[18] N. Upadhyay and A. Karmakar, "An improved multi-band spectral subtraction algorithm for enhancing speech in various noise environments," *Procedia Eng.*, vol. 64, pp. 312–321, Sep. 2013.

[19] D. López-Oller, N. Benamirouche, A. M. Gomeza, and J. L. Pérez-Córdoba, "Speech excitation signal recovering based on a novel error mitigation scheme under erasure channel conditions," *Speech Commun.*, vol. 97, pp. 73–80, Mar. 2018.

[20] V. S. Nataraj, M. S. Athulya, and S. P. Savithri, "Single channel speech enhancement using adaptive filtering and best correlating noise identification," in *Proc. IEEE 30th Can. Conf. Elect. Comput. Eng. (CCECE)*, Apr./May 2017, pp. 1–4.

[21] M. M. Dewasthale, R. D. Kharadkar, and M. Bari, "Comparative performance analysis and hardware implementation of adaptive filter algorithms for acoustic noise cancellation," in *Proc. Int. Conf. Inf. Process.*, Dec. 2016, pp. 124–129.

[22] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 1990, pp. 381–384.

[23] Z. Z. Chomorlig and Y. L. Xiang, "Research of feature extraction in mongolian speech based on an improved algorithm of MFCC parameter," *Adv. Mater. Res.*, vols. 542–543, pp. 833–837, Jun. 2012.

[24] W. Lan, S. Jia, S. Song, and K. Li, "Multi-classification spacecraft electrical signal identification method based on random forest," *J. Beijing Univ. Aeronaut. Astronaut.*, vol. 43, no. 9, pp. 1773–1778, Sep. 2017.

[25] N. Kritika and S. Madhu, "Automatic isolated digit recognition system: An approach using HMM," *J. Sci. Ind. Res.*, vol. 70, no. 4, pp. 270–272, Apr. 2011.

[26] K. K. R, L. Birla, and S. R. K, "A robust unsupervised pattern discovery and clustering of speech signals," *Pattern Recognit. Lett.*, vol. 116, pp. 254–261, Dec. 2018.

[27] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers Comput. Neurosci.*, vol. 9, p. 99, Aug. 2015.

YANYAN SHI is currently pursuing the master's degree with the College of Information and Computer, Taiyuan University of Technology, China. Her current research interests include speech feature extraction and speech recognition.

JING BAI was born in Taiyuan, Shanxi, China. She received the bachelor's degree in electronic information engineering, the master's degree in information and signal processing, and the Ph.D. degree in circuits and systems from the College of Information Engineering, Taiyuan University of Technology, Shanxi, in 1985, 2004, and 2010, respectively.

She is currently a Teacher, an Associate Professor, a Master Supervisor, and the Director of the Experiment Technology Center, College of Information Engineering, Taiyuan University of Technology. She edited a textbook named *Digital Signal and Logical Design* (2009). Her research interests include multimedia information system and intelligent information processing. In 2011, she received the Teaching Achievements Second Prize of Shanxi Province, and in 2012, she received the Science and Technology Progress Second Prize of Shanxi Province.

PEIYUN XUE was born Taiyuan, Shanxi, China. She received the bachelor's degree in electronic information engineering from the College of Information and Computer, Taiyuan University of Technology, Shanxi, in 2013, where she is currently pursuing the Ph.D. degree in electronic science and technology.

In recent years, she took on and participated in many national and provincial projects. Her research interests include speech signal processing, pathological phonetics, and data processing.

DIANXI SHI received the B.S., M.S., and Ph.D. degrees in computer science from the National University of Defense Technology, Changsha, China, in 1989, 1996, and 2000, respectively. He is currently a Researcher and the Deputy Director of the Artificial Intelligence Research Center, National Innovation Institute of Defense Technology. His research interests include distributed object middleware technology, software component technology, adaptive software technology, and intelligent unmanned cluster system software architecture.