Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9452976, 15 pages
https://doi.org/10.1155/2020/9452976

Research Article

Nonlinear Dynamic Feature Extraction Based on Phase Space Reconstruction for the Classification of Speech and Emotion

Ying Sun, Xue-Ying Zhang, Jiang-He Ma, Chun-Xiao Song, and Hui-Fen Lv

College of Information Engineering, Taiyuan University of Technology, Shanxi Province Jinzhong CityYuci District College Town, Taiyuan030600, China

Correspondence should be addressed to Xue-Ying Zhang; tyzhangxy@163.com

Received 21 October 2019; Revised 16 January 2020; Accepted 27 February 2020; Published 9 April 2020

Academic Editor: Paolo Crippa

Due to the shortcomings of linear feature parameters in speech signals, and the limitations of existing time- and frequency-domain attribute features in characterizing the integrity of the speech information, in this paper, we propose a nonlinear method for feature extraction based on the phase space reconstruction (PSR) theory. First, the speech signal was analyzed using a nonlinear dynamic model. Then, the model was used to reconstruct a one-dimensional time speech signal. Finally, nonlinear dynamic (NLD) features based on the reconstruction of the phase space were extracted as the new characteristic parameters. Then, the performance of NLD features was verified by comparing their recognition rates with those of other features (NLD features, prosodic features, and MFCC features). Finally, the Korean isolated words database, the Berlin emotional speech database, and the CASIA emotional speech database were chosen for validation. The effectiveness of the NLD features was tested using the Support Vector Machine classifier. The results show that NLD features not only have high recognition rate and excellent antinoise performance for speech recognition tasks but also can fully characterize the different emotions contained in speech signals.

1. Introduction

Language is the most effective medium of human communication. Language not only contains interpretable text but also contains a large amount of paralinguistic information that can reflect the emotional changes in a speaker. Interpretation of human spoken language through technologies such as speech recognition and affective computing have found a wide range of applications in diverse domains such as vehicle navigation, video surveillance, network video, and other human-computer interaction fields. Speech recognition refers to the ability of machines to convert spoken language into written text. To do this, a speech recognition system often needs to take into consideration the specific and nonspecific environment to recognize the content of speech accurately. Therefore, feature extraction and speech signal characterization are two important steps for accurate speech recognition. Currently, the most important feature extraction techniques used in speech recognition can be divided into (a) prosodic features [1], (b) phonetic features [2], (c) features based on the correlation characteristics of the spectrum [3,4], and (d) feature fusion [5]. The above features are characterized by the piecewise linearity of speech signals. However, studies have shown that speech signal generation is neither a linear process, nor a stochastic process, but rather a nonlinear process [6]. Thus, only using the piecewise linearity of speech signals in the time and frequency domains to extract speech feature will lead to the loss of some of the nonlinear features of speech signals, making the information being extracted incomplete.

With recent development in nonlinear analysis methods, they have been successfully applied in various fields [7–12]. Zbancioc [7] applied the Lyapunov index for the extraction of spectral coefficients of MFCC and LPCC features and achieved an emotion recognition accuracy of 75%; Firoozet al. [8] evaluated nonlinear dynamic features by reconstruction of speech signals using phase space reconstruction to improve the accuracy of automatic speech recognition. Spanish researcher Karmele Lopez applied the study of the chaotic characteristic of natural speech for the detection of Alzheimer's disease and pointed out detection of the speaker's lesions by extracting the fractal dimension features in natural speech [9, 10]. Xiang and Tan of Beijing Jiaotong University combined the chaotic features from speech with other common features to detect fatigue among automobile drivers [11]. Although some researchers have studied the chaotic characteristics of speech signals, very few studies have focused on the nonlinear dynamics and geometric features of chaotic characteristics in speech signals.

Aerodynamic studies have shown that people generate vortices in the channel boundary layer when they make sounds, and this vortex can eventually form turbulence [12]. The nature of this turbulence is chaotic. To verify the chaotic characteristics of speech signals, this paper explores this chaotic mechanism of speech signal generation, from three different analytical aspects: (a) power spectrum, (b) principal component analysis, and (c) phase space reconstruction. This research aims to provide a theoretical basis for extracting nonlinear dynamic features based on the chaotic characteristics of speech signals. By studying and analyzing the two main parameters of phase space during phase reconstruction of speech signals, the minimum time and embedding dimension, we realize the optimal phase space reconstruction. Then, we extract the nonlinear dynamics features from the phase space. By designing experiments to contrast the dynamic features and MFCC nonlinear features for speech recognition, we verify that nonlinear dynamic features of speech signals not only provide high accuracies and excellent noise cancellation performance for speech recognition but also help in identifying emotional cues in speech.

2. Chaos Theory and Verification of Chaotic Characteristics in Speech

2.1. Chaos Theory. Chaos is a seemingly irregular, random phenomenon that occurs in deterministic systems [13]. Although a chaotic system has no obvious cycle, and the form of motion seems disorderly, the internal structure is ordered, and it is a new existence form of nonlinear systems.

Nonlinear dynamics are mainly studied for describing a system or time series. The internal state of motion and the law of transformation of a nonlinear system or time series are analyzed qualitatively and quantitatively [6]. At present, the method of nonlinear dynamic analysis of time series has been maturing and has a relatively complete theoretical research background, covering different nonlinear modeling techniques and nonlinear representations [14], such as fractal dimensions, Lyapunov index, and Kolmogorov entropy, among others. These features can not only effectively distinguish the signal sequence due to chaotic characteristics but also effectively describe the motion state and variation of the signal. These features absent in traditional analysis methods give an advantage to nonlinear modeling.

2.2. Verification of Chaotic Characteristics in Speech. There are two basic features which are used to describe chaotic characteristics. The chaotic attractor of high-dimensional phase space reconstruction has (a) fractal dimension characteristics and (b) initial conditions which have great influence on the system [13]. If a time series has the above two characteristics, we can say that the time series itself is chaotic. Based on the above theory, this paper verifies the chaotic characteristics of speech signals from three aspects: (a) power-spectrum analysis [13], (b) principal component analysis [13], and (c) phase space reconstruction [13].

2.2.1. Power-Spectrum Analysis Method. From the time-domain waveform, we cannot intuitively determine whether the time series is periodic or disordered. However, its power spectrum can be used to identify these regularities. Analysis of the power spectrum can help determine whether the time series demonstrates chaotic characteristics. This analysis is based on two aspects: the number of peaks in the power spectrum and the broad-spectrum characteristics. If there are a finite number of peaks in the power spectrum, the time series is said to have a periodic sequence. However, if there is no obvious peak in the spectrum and it demonstrates a "wide-spectrum" characteristic, we can say that the time series is turbulent or chaotic. Therefore, power-spectrum analysis has evolved as a theoretical basis for judging whether the signal has chaotic characteristics.

In this paper, we analyze the power spectrum of the speech signals of a single word in the Korean isolated words database [15]. The analysis is done for four cases: "15 dB, 20 dB,""25 dB," and "clean." From Figure 1, we can see that the speech signals of the four SNR have a wide spectrum and no special peak. Therefore, it can be verified that the isolated speech signals are chaotic.

2.2.2. Principal Component Analysis. Principal component analysis (PCA) is an effective method to identify a time series which has chaotic characteristics. The steps for the calculations are as follows.

Given a time series $[x(1), x(2), \ldots, x(N)]$, the appropriate embedding dimension m is chosen to construct the matrix $X_{k\times m}$ $(k = N - (m - 1))$, which is represented as

$$X_{k\times m} = \frac{1}{\sqrt{k}} \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_2 & x_3 & \cdots & x_{m+1} \\ \vdots & \vdots & \cdots & \vdots \\ x_k & x_{k+1} & \cdots & x_N \end{bmatrix} = \frac{1}{\sqrt{k}} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}. \quad (1)$$

Then, the covariance matrix $A\,(A \in R^{m\times m})$ of the trajectory is calculated as

$$A_{m\times m} = \frac{1}{k} X_{k\times m}^{T} X_{k\times m}. \quad (2)$$

Then, the eigenvalues of the covariance matrix $A\,(A \in R^{m\times m})$ are solved to obtain $\lambda_i\,(i = 1, 2, \ldots, m)$. Next, we calculate the sum of all the eigenvalues λ and then sort the eigenvalues $\lambda_i\,(i = 1, 2, \ldots, m)$ in descending order. We calculate and plot the main component spectrum using

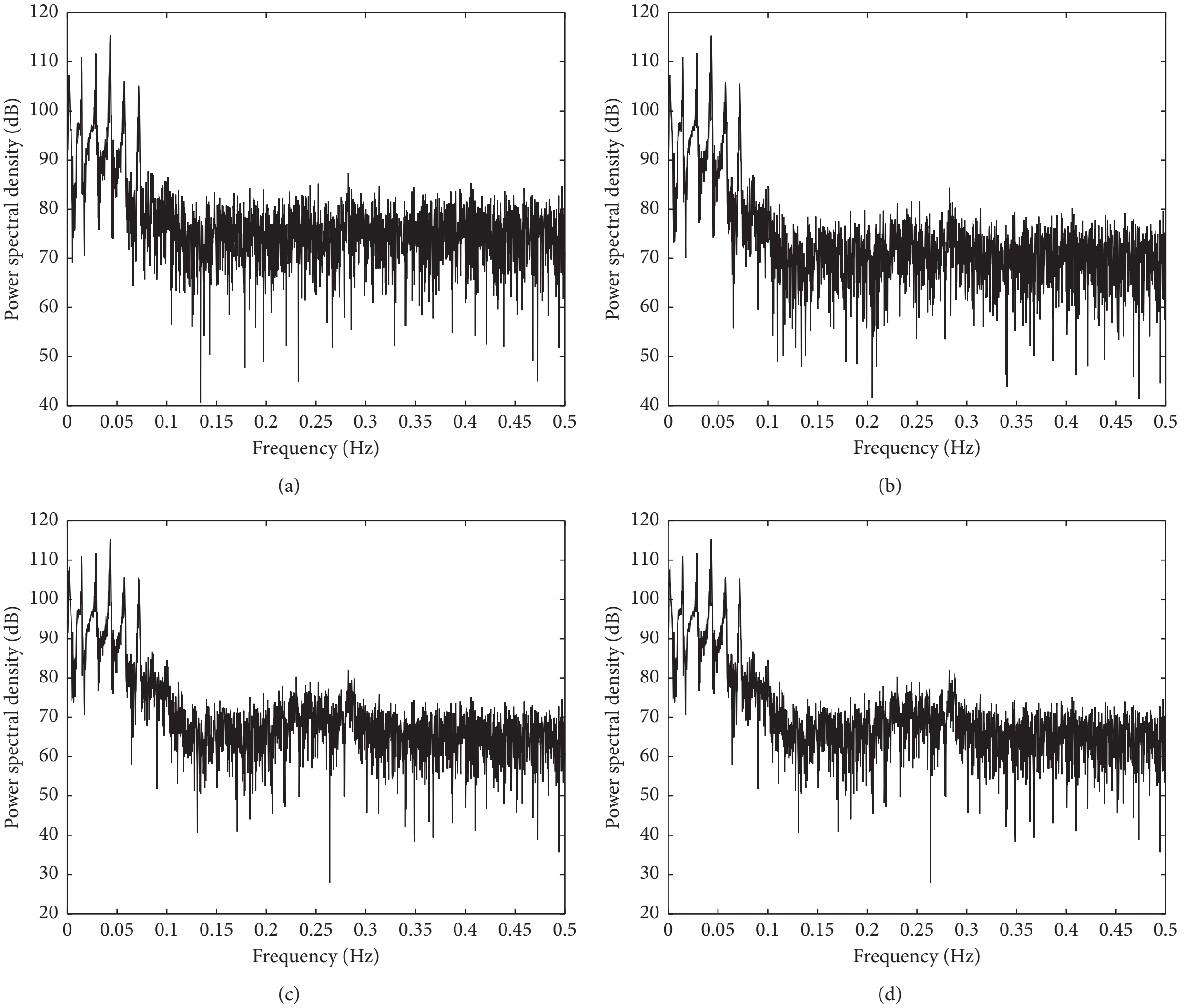

Figure 1: Power-spectrum analysis of isolated speech. Power spectrum of (a) 15 dB, (b) 20 dB, (c) 25 dB, and (d) 30 dB isolated speech signals.

$\ln(\lambda_i/\lambda) - i$ as the coordinates for the simulation graphics. If the principal component spectrum is a nearly straight line with a negative slope, this indicates that the signal has chaotic characteristics.

As shown in Figure 2, in this paper, we carry out the principal component spectral analysis of the four emotions—“happy,”“sad,”“neutral,” and “anger”—from semantic speech signals taken from the Chinese Affective Chinese database (CASIA) [16]. As can be seen from Figure 2, the covariance matrix is used to calculate the three eigenvectors ($i = 1, 2, 3$), and the resulting value $\ln(\lambda_i/\lambda)$ is calculated as a nearly straight line with a negative slope in the graph. Therefore, it can be shown that the emotional speech signals are chaotic.

2.2.3. Phase Space Reconstruction. Phase space reconstruction (PSR) is the first step to analyze nonlinear dynamic, commonly used in the embedding theorem proposed by Taken’s [17]. The essence of this method is to construct an m-dimensional space vector $\{x(t), x(t+\tau), \ldots, x(t+(m-1)\tau)\}$ by selecting the corresponding appropriate delay time τ and embedding dimension m from the one-dimensional time series $x(t)$. The reconstructed high-dimensional space is equivalent to the original space. Given the time series of the one-dimensional emotional speech signals x_i, $i = 1, 2, 3, \ldots, N$, we select the appropriate time delay τ and embedding dimension m. The sequence expression after phase space reconstruction can be written as

$$\overrightarrow{x_i} = \left(x_i, x_{i+\tau}, \ldots, x_{i+(m-1)\tau}\right). \quad (3)$$

The row vector $\overrightarrow{x_i}$ represents the location information of each single attractor required for phase space reconstruction. The definition of nonlinear dynamical systems indicates that these vectors are connected by a column to form a trajectory matrix. This information can be used to create the following PSR matrix:

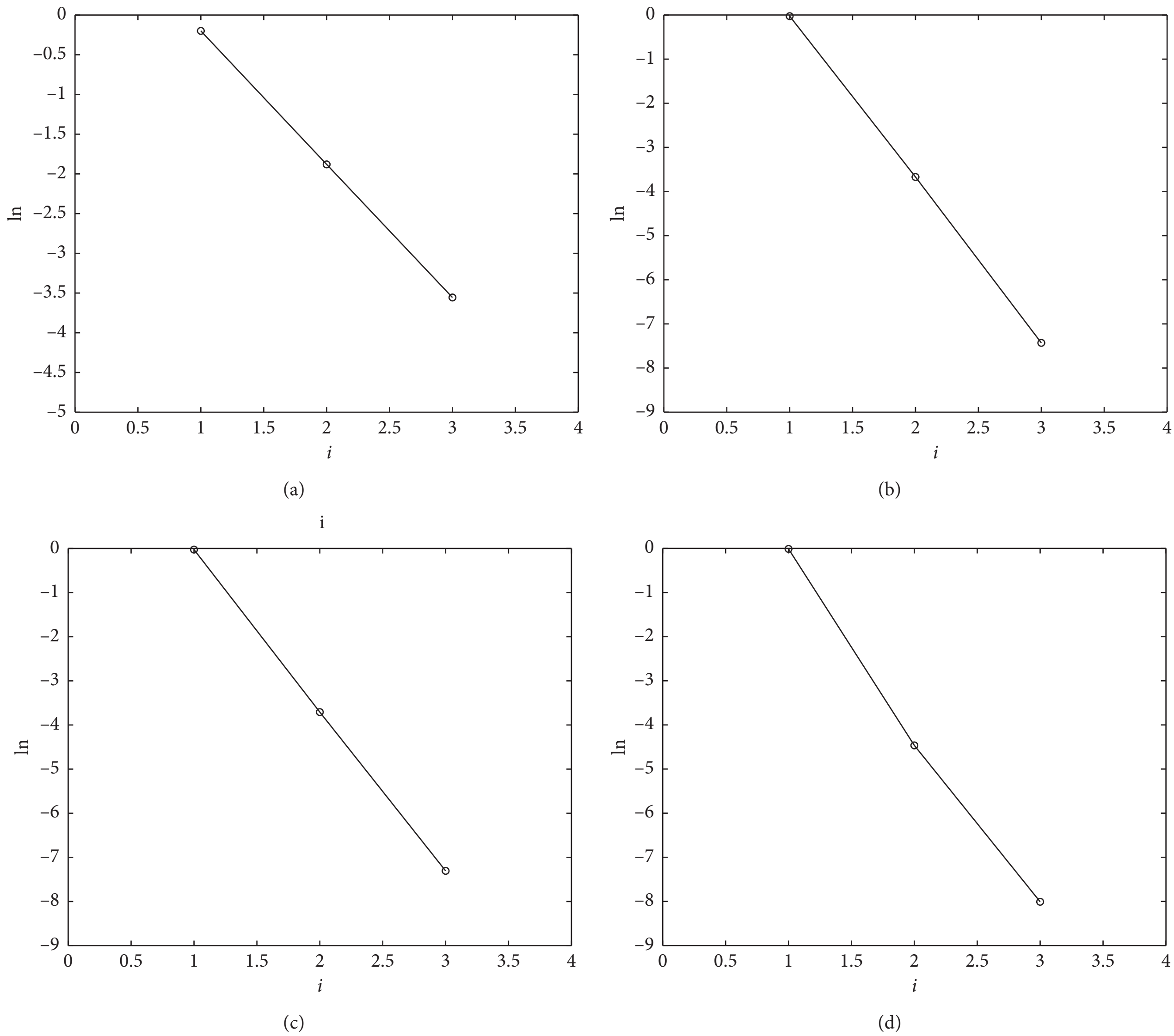

Figure 2: Principal component-spectrum analysis of different emotional speeches from CASIA emotional speech corpus. Principal component-spectrum analysis of (a) happiness, (b) sadness, (c) neutral, and (d) anger.

$$X = \begin{bmatrix} x_1 & x_{1+\tau} & x_{1+2\tau} & x_{1+(m-1)\tau} \\ x_2 & x_{2+\tau} & x_{2+2\tau} & x_{2+(m-1)\tau} \\ x_3 & x_{3+\tau} & x_{3+2\tau} & x_{3+(m-1)\tau} \\ \vdots & & & \\ x_N & x_{N+\tau} & x_{N+2\tau} & x_{N+(m-1)\tau} \end{bmatrix}. \quad (4)$$

The significance of a high-dimensional phase space is that the internal structure of the signal can be expanded. The signal can be projected onto a high-dimensional space, and the qualitative properties of the signal can be obtained by measuring and predicting the evolutionary trajectory in this space.

This paper reconstructs the phase space by measuring different emotions in the same semantics of the Berlin emotional speech database [18]. In this paper, we study the overall structure and motion trajectory of the speech signals under a one-dimensional time series and a three-dimensional phase space reconstruction for four emotional states: “happy,”“sad,”“neutral,” and “angry.” From Figure 3, we can see that the differences between the four kinds of emotional speech are mainly reflected in features such as the number of peaks, the peak size, and the number of zero crossings in the time-domain waveform. However, there are also significant differences in the overall structure and motion trajectory once the four kinds of emotional speech are reconstructed in a three-dimensional phase space. Therefore, a nonlinear dynamic model can be used to analyze the chaotic characteristics of speech signals.

3. Nonlinear Dynamic Feature Extraction from Speech

Phase space reconstruction is one of the key techniques used to study time series with chaotic characteristics. Taken’s

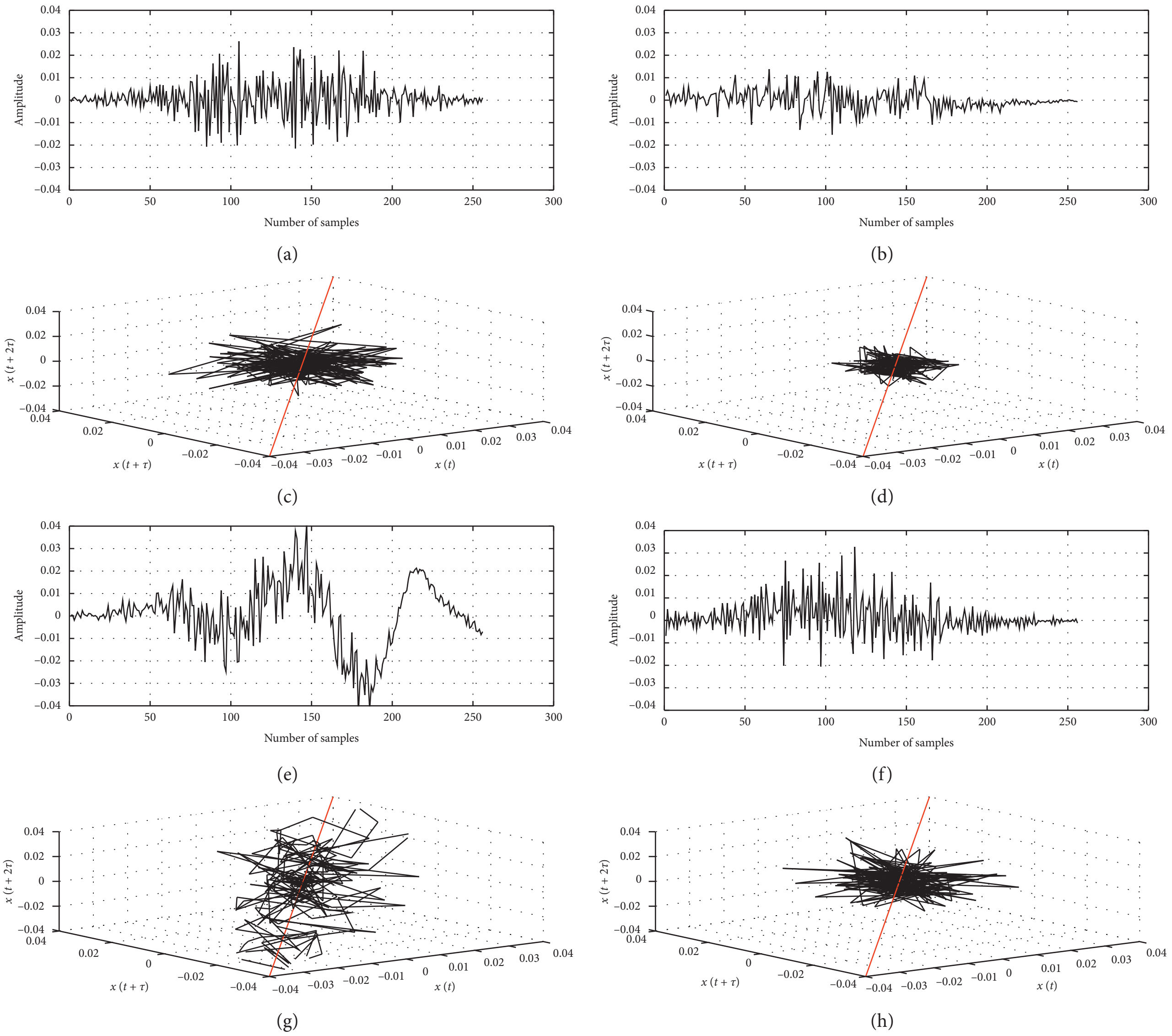

Figure 3: Time-domain waves and phase space reconstruction of the Berlin-DB emotional speech corpus. The time-domain wave of (a) happiness, (b) sadness, (e) neutral, and (f) anger. Phase space reconstruction of (c) happiness, (d) sadness, (g) neutral, and (h) anger.

embedding theorem [14] states that as long as the appropriate time delay τ and the embedded dimension m are appropriately selected, the one-dimensional emotional time series $X = [x_1, x_2, \ldots, x_N]$ can be mapped from a low-dimensional space to a high-dimensional space $X(t_i) = [x(t_i), x(t_i + \tau), \ldots, x(t_i + (m-1)\tau)]$ to realize phase space reconstruction. Here, $i = 1, 2, \ldots$, and ensure that the reconstructed phase space and the original one-dimensional voice signal retain information integrity. The emotional speech signals are analyzed under the reconstructed phase space, and then, the following nonlinear dynamic (NLD) features are extracted. The algorithm flow is shown in Figure 4.

3.1. Preprocessing. Since speech signals are nonstationary and time-varying and have short-time stationary characteristics, the following three steps are needed for the processing and analysis of speech signals: ①endpoint detection: the identification of the start and end points of the speech signals based on energy and zero rate; ②preemphasis: a first-order digital filter is used to preaccentuate the high-frequency part of the speech signals; ③window framing: a Hamming window is used for frame processing, with a frame length of 256 and a frame shift of 128.

3.2. C-C Algorithm. The purpose of phase space reconstruction is to extend the dynamic one-dimensional speech signals into a high-dimensional space to completely reveal the implicit information in the time series. However, we observed that the significant parameter delay time τ of the reconstruction phase space is strongly correlated with the embedded dimension m. Therefore, this paper chooses the C-C [18] method to calculate the delay time τ and the

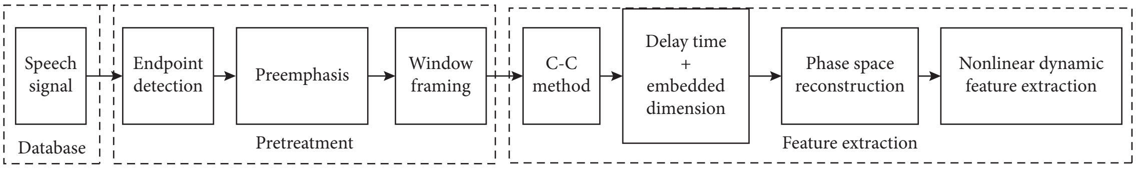

Figure 4: Flow chart of feature extraction algorithm.

window delay time τ_w. This paper also further obtains the embedded dimension m which is a part of the implicit information in the time series. In view of the current spatial coordinates, the geometric information is limited to a two- or three-dimensional space. This paper improves the C-C method and extends its speech time series to two- and three-dimensional phase spaces to extract five nonlinear geometric features (NLD-2) from the structural trajectory contours. The specific calculations are performed in the following steps:

(1) As shown in equation (5), the time series $X = [x_1, x_2, \ldots, x_N]$ is divided into t disjoint time subsequences:

$$\begin{aligned}\overline{x_1} &= \{x_1, x_{t+1}, x_{2t+1}, \ldots, x_{N-t+1}\},\\ \overline{x_2} &= \{x_2, x_{t+2}, x_{2t+2}, \ldots, x_{N-t+2}\},\\ &\vdots\\ \overline{x_t} &= \{x_t, x_{2t}, x_{3t}, \ldots, x_N\},\end{aligned} \quad (5)$$

where the length is $l = (N/t)$.

(2) The associated integral of the embedded time series is defined by the following function:

$$C(m, N, r, t) = \frac{2}{M(M-1)} \sum_{t \le i \le j \le M} \theta\left(r - d_{ij}\right), \quad r > 0, \quad (6)$$

where $M = N - (m-1)\tau d_{ij} = \|X_i - X_j\|$, and when $x < 0$, $\theta(x) = 0$, and $x \ge 0$, $\theta(x) = 1$.

(3) The $S_{(m,N,r,t)}$ of the subsequence $\overline{x_i}$ is defined using the associated integral $C(m, l, r, t)$ function:

$$S_{(m,N,r,t)} = \frac{1}{t} \sum_{s=1}^{t} \left[C_s(m, l, r, t) - C_s^m(1, l, r, t)\right]. \quad (7)$$

When $N \longrightarrow \infty$, $S_{(m,r,t)} = (1/t)\sum_{s=1}^{t}[C_s\ (m, r, t) - C_s^m\ (1, r, t)]\,(m = 2, 3, \ldots)$. If the time series is independently distributed, then for fixed m, t, when $N \longrightarrow \infty$, for all r, $S(m, r, t)$ is equal to zero. But the actual sequence is limited, and the sequence elements may be related, we actually get $S(m, r, t)$ which is generally not equal to zero so that the local maximum time interval can be located at the zero point of $S(m, r, t)$ or at the minimum time point for all the differences between the radii. Since this implies that these points are almost uniformly distributed, the maximum and minimum radii of the corresponding values are selected, and the difference $\Delta S_{(m,N,t)}$ can be written as

$$\Delta S_{(m,N,t)} = \max S\left(m, N, r_j, t\right) - \min S\left(m, N, r_j, t\right). \quad (8)$$

The above formula measures the maximum deviation of the radius r.

(4) To calculate the time delay τ and the window delay time τ_W, we must first calculate the following three components:

$$\begin{aligned}S_t &= \frac{1}{4} \sum_{m=2}^{3} \sum_{j=1}^{2} S\left(m, N, r_j, t\right),\\ \Delta S_t &= \frac{1}{2} \sum_{m=2}^{3} \Delta S_{(m,N,t)},\\ S_{\text{cor}}(t) &= \Delta S_t + \left|S_t\right|,\end{aligned} \quad (9)$$

where r_j is $r_j = (j\sigma/2)$ and σ is the mean square of the time series time delay. τ is the first value of S_t or the first minimum of ΔS_t corresponding to the value of the input t. The window delay time τ_W is the value of the input t corresponding to the minimum value of $S_{\text{cor}}(t)$.

(5) The embedded dimension m is calculated:

$$m = \left(\frac{\tau_W}{\tau} + 1\right). \quad (10)$$

3.3. Nonlinear Attribute Feature Extraction

(1) Minimum delay timethe known speech signal is represented as $[x(1), x(2), \ldots, x(N)]$. Here, we use the mutual information function to calculate the mutual information between the speech signals $x(i)$ and $x(j)$ at different time intervals. At the points where the mutual information of these two speech time series reaches the minimum, the correlation between the two variables is also minimal. This corresponding time interval is the minimum delay time τ. As shown in equation (11), this paper uses the average mutual information (MI) [19] to calculate the minimum delay timeτ:

$$I(\tau) = \sum_{i,j} p_{i,j}(\tau) \ln\left[\frac{p_{i,j}(\tau)}{p_i p_j(\tau)}\right], \quad (11)$$

where p_i and p_j, respectively, represent the probability of the sequence amplitude falling in the ith and jth segments, respectively. $p_{i,j}$ denotes the joint probability of the two-point amplitude of the sequence at time interval τ. The minimum delay which quantifies the disorder between two discrete variables corresponds to the moment of the first local minimum of the obtained mutual information function curve.

(2) Correlation dimension:the correlation dimension is a nonlinear representation of chaotic dynamics. It is used to describe the property of the dynamics and self-similarity of the structure of high-dimensional spatial speech and provides a quantitative analysis of the complexity of its structure. The more complex the corresponding system structure is, the greater will be the correlation dimension. The correlation dimension is calculated using the G-P algorithm [20]. As shown in equation (12), the G-P algorithm is a method proposed by Grassberger and Procaccia for calculating the correlation dimension:

$$D(m) = \frac{\ln C(r,m)}{\ln r}, \tag{12}$$

where $D(m)$ is the relational dimension and $C(r,m)$ is the correlation integral function. $C(r,m)$ is the ratio of the phase point between any $(\mathbf{X_i}, \mathbf{X_j})$ in the m-dimensional reconstruction space which is less than r, the ratio of all phases, and is defined as

$$C(r,m) = \frac{1}{M(M-1)} \sum_{i,j=1,i\neq j}^{N} \theta\Big(r - \left\|\mathbf{X_i} - \mathbf{X_j}\right\|\Big). \tag{13}$$

In Equation (13), the corresponding $\ln C(r,m) \longrightarrow \ln r$ curve is obtained by taking the minimum embedded dimension of m, and the correlation dimension can be obtained by fitting the local line of the curve.

(3) Kolmogorov entropy: it is a physical quantity used to accurately describe the degree of confusion in a time-series distribution. Grassberger and Procaccia proposed the correlation dimension analysis method. They demonstrated that the K entropy can be approximated using the K_2 entropy. The relationship between K_2 entropy and the correlation integral function $C(r,m)$ can be expressed as

$$K_2 = \frac{1}{m\tau}\log_2\frac{C(r,m)}{C(r,m+1)}. \tag{14}$$

This entropy calculated in equation (14) is the Kolmogorov entropy.

(4) Largest Lyapunov exponent: the Lyapunov exponent is used to quantify the average change in the rate of local convergence or divergence of adjacent orbits in the phase space. The maximum Lyapunov exponent x represents the degree of convergence or divergence of the orbit. When $\lambda_1 > 0$, as the value of λ_1 increases, the value of the orbital divergence and the chaos also increases. The paper uses the Wolf method [21] to obtain the maximum Lyapunov exponent. Here, we take the initial point $\mathbf{X_i}$ in the phase space and find its nearest neighbor point $\mathbf{X_{i_l}}$. The distance between them is represented as L_0. This distance is tracked over time as the adjacent orbits in the phase space converge or diverge. A point is retained when the distance L_i between the two points meets the set value ε after n iterations of tracking. Once this condition is met, the next moment is tracked.

When tracking the overlay M times, we can obtain the maximum Lyapunov exponent using the following equation:

$$\lambda_1 = \frac{1}{Mn}\sum_{i=0}^{M} \ln \frac{L_i}{L_0}. \tag{15}$$

Compared with other algorithms, this algorithm has advantages of fast computation, robustness to embedded dimension m, delay time τ , and noise.

(5) Hurst exponent: the Hurst exponent (H) measures the long-term memory of a time series. H lies within the range of 0-1. If $H > 0.5$, it indicates that the time series displays a long-term autocorrelation and the time series is highly correlated. This paper uses the rescaled-range analysis method [22] to calculate the H value. The rescaled-range is a nonparametric statistical method, which is not affected by the distribution of the time series. The method divides the one-dimensional speech signal with emotional content $[x(1), x(2), \ldots, x(N)]$ into M adjacent subsequences C of equal lengths. By calculating the cumulative deviation z_u and the standard deviation S_u for each subsequence and then calculating the weight difference of each sub-sequence R_u/S_u, we obtain the Hurst exponent using $R_u = \max z_u - \min z_u$. The calculation is as follows:

$$\frac{R_M}{S_M} = bH^M. \tag{16}$$

Here, b is a constant. By taking the logarithm of both sides of equation (16), we can obtain the value of H which is the Hurst exponent. For different emotional states contained in a speech signal, the changes in the value H are different. The Hurst exponent feature of the extracted emotional speech reflects the correlation between the emotion and the change.

3.4. Nonlinear Geometric Feature Extraction. After the one-dimensional speech signal is mapped to a high-dimensional space using phase space reconstruction, the speech signal is analyzed in the high-dimensional space. Next, the geometric features—which are the five trajectory-based descriptor contours—of the phase space reconstruction for different speech states are extracted. These five descriptors are detailed as follows:

(1) The first contour: the distance from the attractor to the center is expressed as $\overline{a} = [|\overrightarrow{a_1}|, |\overrightarrow{a_2}|, \ldots, |\overrightarrow{a_N}|]$:

$$|\overline{a_i}| = \begin{cases} \sqrt{a_i^2 + (a_i + \tau_i)^2}, \\ \sqrt{a_i^2 + (a_i + \tau_i)^2 + (a_i + 2\tau_i)^2}. \end{cases} \quad (17)$$

Among them, the two-dimensional space under the attractor is defined as $\overrightarrow{a_i} = (a_i, a_i + \tau_i)$, and the three-dimensional space under the attractor is defined as $\overrightarrow{a_i} = (a_i, a_i + \tau_i, a_i + 2\tau_i)$.

(2) The second contour: the length of the continuous trajectory between the attractors is expressed as$\overline{l} = [|\overrightarrow{l_1}|, |\overrightarrow{l_2}|, \ldots, |\overrightarrow{l_{N-1}}|]$:

$$|l_i| = |\overrightarrow{a_{i+1}} - \overrightarrow{a_i}|. \quad (18)$$

(3) The third contour: the trajectory of the continuous path between the attractors is expressed as $\overline{\theta} = [\theta_1, \theta_2, \ldots, \theta_{N-2}]$:

$$\theta_i = \frac{(\overrightarrow{a_i} - \overrightarrow{a_{i+1}}) \cdot (\overrightarrow{a_{i+1}} - \overrightarrow{a_{i+2}})}{|\overrightarrow{l_i}||\overrightarrow{l_{i+1}}|}. \quad (19)$$

(4) The fourth contour: the distance from the attractor to the marker line is expressed as $\overline{d} = [d_1, d_2, \ldots, d_N]$:

$$d_i = \begin{cases} \dfrac{(1,1) \otimes (a_i, a_i + \tau_i)}{\sqrt{2}}, \\ \\ \dfrac{(1,1,1) \otimes (a_i, a_i + \tau_i, a_i + 2\tau_i)}{\sqrt{3}}. \end{cases} \quad (20)$$

For the time delay $\tau = 1$, when the original waveform $x(t)$ is lagged, there will be a small difference between the two samples $x(t-1)$ and $x(t-2)$. This can be expressed as the identity [20]:

$$x(t) = x(t-1) = x(t-2). \quad (21)$$

From formula (20), we can observe that the upper form will not hold when the three attractors are different. Since the dynamic factors of the chaotic system are interactive, the data points produced in time will also be correlated [23]. Therefore, formula (21) represents the labeling line. The differences between the attractors can be obtained by analyzing the distances between the attractors and the labeling line.

(5) The fifth contour: the total length of the trajectory of the attractor is expressed as S:

$$S = \sum_{i=1}^{N} |ai|. \quad (22)$$

4. Experimental Preparation

4.1. Speech Corpora

4.1.1. Korean Isolated Words Database [15]. The isolated words database was used for performing speaker-independent, isolated word recognition from neutral (nonemotional) speech. The vocabulary sizes used in the experiments were 10 words, 20 words, 30 words, 40 words, and 50 words. The corpus consisted of ten digits and 40 command words with 16 speakers thrice repeating each word. For our experiment, we used the recording of the utterances of 9 speakers as the training set and the utterances of the remaining 7 speakers as the test set.

4.1.2. CASIA Database [16]. The CASIA database is a Chinese database developed in the Institute of Automation, Chinese Academy of Sciences. The recordings consisted of six acted (simulated) emotions (Neutral, Anger, Fear, Happiness, Sadness, and Surprised) by four professional speakers (2 females and 2 males). Each emotion category consists of 300 identical texts and 100 different texts. Recordings of readings of the same text with different emotions are useful for the comparison of acoustics and prosodic performance for different emotional states. Another 100 different texts with emotional content that matched the emotion being expressed made it easier for the articulating person to express their feelings better. The recordings were performed with a sampling rate of 16 kHz and a 16-bit resolution and were stored in PCM format.

4.1.3. Berlin Database [17]. The Berlin database is a German database recorded in an anechoic chamber at the Technical University Berlin. The database consists of 10 actors (5 females and 5 males) who simulated seven emotions (Neutral, Anger, Fear, Happiness, Sadness, Disgust, and Boredom). Each emotion category contains ten German sentences. The recordings were performed with a sampling frequency of 48 kHz and later downsampled to 16 kHz with high-quality recording equipment. In our experiments, we use happy, sad, neutral, and angry as the four basic emotions from the German Berlin speech library.

Taking into account the effect of the length of speech on the recognition results, this paper filters the database to obtain 363 German sentences and 1000 Chinese sentences with approximate speech length of five seconds. The results of the division of emotional speech into the training and test set are shown in Table 1.

4.2. Feature Extraction. Previous studies have demonstrated that prosodic features [24] and MFCC features [24] are highly efficient for distinguishing between different emotional states. In this paper, we first perform a series of preprocessing operations on the speech signals. Then, we extract the prosodic features and MFCC features for each speech frame. We also extract the NLD-1 and NLD-2 features based on the phase space reconstruction method described earlier in this paper. Then, we calculate the statistical functions for the above features. These statistical functions include the maximum and the minimum values, the mean, the variance, the median, the deviation, and the kurtosis. Finally, as shown in Table 2, we end up with a feature set of 150 dimensions. The normalized method of linear function

Table 1: Corpus setting for emotional speech experiment.

Database	Berilin-DB					CASIA				
Emotion	Happy	Sadness	Neutral	Anger	Fear	Happy	Sadness	Neutral	Anger	Fear
Training	47	42	53	55	46	132	132	132	132	132
Test	24	20	26	27	23	68	68	68	68	68
Sum	71	62	79	82	69	200	200	200	200	200

Table 2: Statistics of feature parameters extracted from speech.

Features	Dimensions	Statistics
Prosodic features	38	Speed Average zero-crossing rate Energy and its 1^{st}-order maximum, minimum, and mean values Fundamental frequency and its 1^{st}-order maximum, minimum, and mean values First formant and its 1^{st}-order maximum, minimum, and mean values Second formant and its 1^{st}-order maximum, minimum, and mean values Third formant and its 1^{st}-order maximum, minimum, and mean values
MFCC	60	The skewness, kurtosis, mean, variance, and median of the first 12 steps of MFCC The maximum, minimum, mean, median, and variance of the hurst exponent The maximum, minimum, mean, median, and variance of the minimum delay time
NLD-1 (nonlinear attribute features)	59	Correlation dimension's maximum, minimum, mean, median, and variance Kolmogorov entropy's maximum, minimum, mean, median, and variance The mean, median, and variance of the largest Lyapunov index The first, second, and third contours
NLD-2 (nonlinear geometric features)	23	The maximum, minimum, mean, variance, standard deviation, skewness, and kurtosis The maximum, minimum, skewness, and kurtosis of the fourth contour The fifth contour

transformation is used to eliminate the influence of different types of affective features, and then, the objective performance is evaluated synthetically.

4.2.1. Prosodic Feature Extraction. Prosodic features mainly describe the nonverbal information in the emotional speech signal, including the level, the length, the speed, the severity of speech, and the fluent speech information. Therefore, the prosodic feature, also known as the "supersegmental feature," is also recognized for its ability to recognize emotions.Therefore, we use speech speed, average zero-crossing rate, energy, fundamental, and formant as the prosodic feature.

4.2.2. MFCC Feature Extraction. The ability of the human ear to perceive the sound intensity is related to the frequency of the sound. At low frequencies, the perceived sound perception of a human ear is linear with the sound frequency. At high frequencies, due to the masking effect, the perception of the human ear to the sound is nonlinear with the frequency of the sound, so Mel frequencies are introduced to simulate auditory properties. This paper uses the expression: $f_{\text{mel}} = 1125 * \ln((1+f)/700)$. The ordinary frequency is converted to Mel frequency, and the first 12 steps of MFCC are extracted.

4.3. Classification. Constructing a reasonable and efficient speech recognition model is the most important research challenge in the field of speech recognition technology. It requires learning from a large training corpus, which can be used to explore a variety of acoustic features for mapping the corresponding path of the speech signals to achieve the correct identification. Currently, for speech recognition tasks, both linear and nonlinear classifiers are used. The linear ones include Naïve Bayes Classifier, Linear ANN (artificial neural network), and Linear SVM (support vector machine).The nonlinear ones include Decision Trees, k-NN (k-nearest neighbor algorithm), and Nonlinear ANN. Nonlinear classifiers also include SVMs, GMM (Gaussian mixture model), HMM (hidden Markov model), and sparse means classifiers, among others. Researchers have experimented with different model classifiers for improving speech recognition. The most widely used classifiers for speech recognition are HMM [25,26], GMM [27,28], ANN [29,30], and SVM [31,32]. In this paper, to improve the separability of data, the SVM classifier is used to generate a nonlinear mapping of the original features to a high-dimensional space; the choice of kernel function is Radial Basis Function (RBF).

5. Experimental Setup and Analysis of Results

To verify the validity and robustness of the proposed NLD feature set, we design the following two experiments. The first experiment consists of an analysis of the influence of PSR parameter selection on the NLD feature set. The second experiment verifies the validity of NLD features for speech recognition by comparing them with traditional acoustic features.

5.1. Influence of PSR Parameter Selection on NLD-2 Features. We design two experiments to verify the validity of two important parameters of phase space reconstruction and discuss the results under different parameters:

> Experiment 1: first, we generate the phase space reconstruction of speech signals using the delay time τ and the embedding dimension $m\,(\tau = 1, m = 3)$ as set in the document [20]. Next, the phase space reconstruction of speech signals is also carried out using the delay time τ and the embedding dimension m for each frame of the speech signal extracted using the improved C-C method. Finally, we compare the results of the two experiments.
>
> Experiment 2: in view of the current research on spatial coordinates, the geometric information is limited to the two- or three-dimensional space [13]. Therefore, we set the value of the embedded dimension as $m = 3$ and the delay time $\tau = 1, 2, \ldots, 5$. This is done to compare the experimental results for the delay times and embedding dimensions.

We reconstruct the phase space, based on the above two sets of experimental parameters. Next, we extract five kinds of NLD-2 features from the corresponding phase space of the Berlin-DB for the recognition of five basic emotions. The experimental results are shown in Table 3 and Figure 5.

From Table 3, we can observe the task of recognition of emotional speech, and we obtain a higher accuracy (75%) for the delay time and the embedded dimensions than those reported in the literature [20]. Our system demonstrates an increase of 33.3%, for the happiness category, while the recognition rates for sadness, anger, and fear are relatively low. However, from the perspective of average recognition rate, using NLD features extracted by our method based on the parameters of this paper, we obtain a recognition rate which is 2.5% higher.

According to the experimental results shown in Figure 5, the NLD-2 features based on the method of parameter setting cannot achieve the optimal recognition rate for the recognition of each emotion speech category. However, the overall recognition trend is relatively smoother than other approaches. At the same time, we are also able to achieve an optimal value for the average recognition rate. This indicates that the five NLD-2 features used to solve the delay time τ and the embedding dimension m are valid based on the method of improving the C-C. This also proves that compared with setting fixed values for the delay time τ and embedding dimension m, the method of using C-C to set the delay time τ and the embedding dimension m for each frame of the speech signal's phase space reconstruction yields better results for recognition of emotional speech signals.

5.2. Validity and Verification of Robustness of the NLD Features. In this paper, we used three methods to verify the validity of the extracted features.

5.2.1. Experimental Scheme 1: Speech Recognition of Isolated Words. The ten types of NLD features based on the PSR theory are combined with the MFCC features to identify isolated speech vocabulary. The experimental results are shown in Table 4 and Figure 6. These results verify the validity and robustness of the NLD features based on phase space theory.

The experimental results show that using different vocabulary and different values of signal-to-noise ratio (SNR), the recognition rate can be improved by the combination of NLD features and traditional linear speech acoustic features. Compared with the above four types of feature combination methods, from Table 4, we can see that the complementary effect of NLD-2 features is better than that of NLD-1 features. The effect of combination of the NLD features and the MFCC features yields optimal results. From the results, it can be seen that the recognition rate of the feature set comprising of the fusion of traditional linear acoustics with NLD features increases with the increase in the vocabulary size. This can be attributed to an increase in the training set. Therefore, the effective information from the speech signals can be better described by combining or complementing the NLD features with the traditional linear features of speech signals. But the overall recognition rate decreases with an increase in the number of words. This is because the fusion of the above features is not suitable for large words. Therefore, new features must be considered to improve the recognition effect for large vocabulary speech recognition.

5.2.2. Experimental Scheme 2: Single Language Emotion Recognition. The prosodic features, MFCC features, NLD-1 features, and NLD-2 features are combined to recognize single emotional speech from Berlin-DB and CASIA in two languages. The results of the recognition are shown in Tables 5 and 6.

The confusion matrix of the Chinese emotional speech recognition is provided in Table 5. We can see that compared with MFCC, NLD-1, and NLD-2, prosodic features achieve the best recognition rate for the happy emotional state. From the perspective of misjudgment, the misclassification of happiness and anger is the lowest for prosodic features. This indicates that prosodic features can effectively distinguish between happy and angry emotional states. From the overall recognition results, the overall recognition performance of MFCC is higher than that of the other three features and the recognition results for the anger class are optimal. NLD-1 features have better recognition effect for the neutral emotional voice, and NLD-2 has a better recognition for sadness and fear. The recognition performance of NLD alone is not optimal. This can be explained as that for emotional speech, NLD is used for effectively recognizing the effect of local emotion recognition only. It also indicates that the nonlinear feature can make up for the lack of speech chaos observed in previous studies.

In Table 6, we can see the confusion matrix of the Berlin German emotional speech corpus. The recognition effect of NLD-2 is better than that obtained using prosody, MFCC, and NLD-1. For happiness, NLD-2 correctly classifies 50 instances which is higher than the number of instances recognized using the other feature sets. From the

Table 3: Emotional speech recognition using different phase space parameters (%).

The choice of τ and m	Happy	Sadness	Neutral	Anger	Fear	Average
Literature [20] sets $\tau = 1$ and $m = 3$	41.7	90.0	73.1	88.9	73.9	73.3
The paper sets τ and $m = 3$	75.0	85.0	73.1	81.5	65.2	75.8

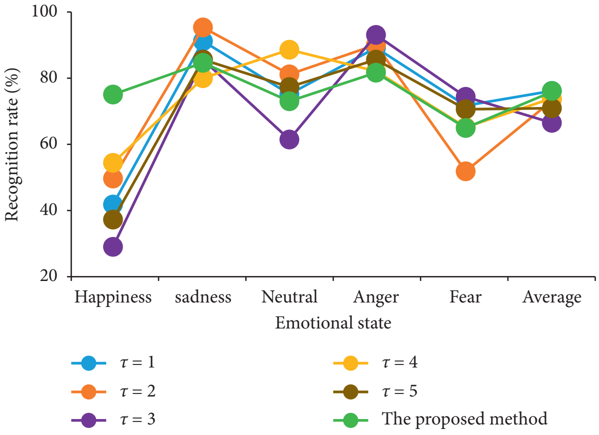

Figure 5: Comparison of emotional speech recognition results under different parameter settings.

Table 4: Recognition results of the two features for different SNR (dB).

Word no.	Feature type	SNE (dB)					
		15	20	25	30	Clean	Average
10	MFCC	91.90	91.90	91.90	93.81	94.76	92.85
	NLD-1 + MFCC	93.33	94.29	93.33	94.29	95.71	94.19
	NLD-2 + MFCC	93.33	94.29	93.33	94.29	95.71	94.19
	NLD + MFCC	92.85	95.24	92.38	95.24	95.71	94.28
20	MFCC	91.19	93.38	91.67	92.86	92.14	92.25
	NLD-1 + MFCC	92.38	92.81	93.57	93.10	92.86	92.94
	NLD-2 + MFCC	91.19	93.81	93.57	92.86	92.61	93.04
	NLD + MFCC	92.14	93.81	93.10	94.05	92.62	93.14
30	MFCC	87.62	89.68	91.27	92.22	89.05	90.32
	NLD-1 + MFCC	90.79	90.79	91.27	92.63	89.79	91.54
	NLD-2 + MFCC	87.62	91.11	92.06	92.22	90.79	90.76
	NLD + MFCC	88.73	91.11	92.06	92.63	91.27	91.16
40	MFCC	86.90	88.69	90.31	90.00	88.33	88.85
	NLD-1 + MFCC	87.14	90.35	91.31	90.36	88.69	89.57
	NLD-2 + MFCC	87.14	90.35	91.79	90.36	88.69	89.67
	NLD + MFCC	87.86	89.88	90.36	90.24	88.93	89.45
50	MFCC	84.29	87.71	88.95	78.76	85.82	85.11
	NLD-1 + MFCC	85.81	87.90	88.48	78.81	85.05	85.21
	NLD-2 + MFCC	85.52	87.52	89.05	78.76	85.52	85.27
	NLD + MFCC	85.52	87.52	89.05	88.10	87.71	87.58

recognition results of fear, the recognition performances of NLD-1 and MFCC reach the optimum values. From the overall recognition results, the recognition performance obtained using MFCC is superior to the other three types of features. This is because the MFCC features extracted for the sadness, neutral, anger, and fear yield the best recognition results. Comparing the results of emotion recognition in two languages, we can see that recognition result of emotional speech is not only related to the language type of the speech database but also has a close relationship with the features. The same feature yields different results for the representation of emotional information in different languages.

In Figure 7, we compare the results for single language emotional speech recognition for German and Chinese. We can see that for the recognition of the emotional speech, only prosodic features yield slightly better results in Chinese than in German. This is because in Chinese, we obtain the highest recognition rate for happy emotional speech. From the results of the recognition rate of the different features, the dominant features based on the recognition performance can be sorted as follows: MFCC > NLD-2 > NLD-1 > prosodic

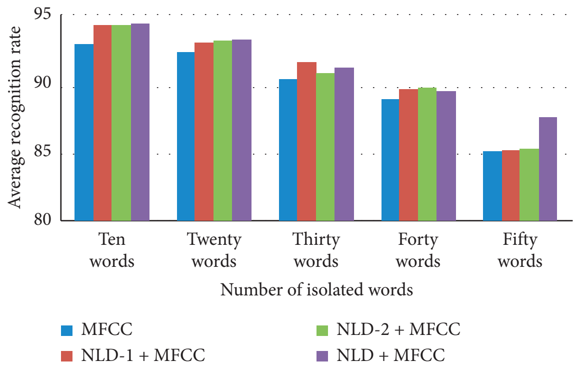

Figure 6: Comparison of speech recognition result of different isolated words.

Table 5: Confusion matrix for the CASIA emotional speech recognition for four feature types.

Prosody		Detected emotion				
Emotional state		Happy	Sadness	Neutral	Anger	Fear
True emotion	Happy	**152**	22	19	1	6
	Sadness	29	**93**	18	19	41
	Neutral	20	21	**113**	29	17
	Anger	7	24	34	**112**	23
	Fear	12	51	26	19	**92**
NLD-1						
Emotional state		Happy	Sadness	Neutral	Anger	Fear
True emotion	Happy	**137**	12	34	17	0
	Sadness	35	**90**	22	18	35
	Neutral	13	11	**128**	22	26
	Anger	11	20	19	**124**	30
	Fear	0	49	29	24	**98**
MFCC						
Emotional state		Happy	Sadness	Neutral	Anger	Fear
True emotion	Happy	**137**	4	48	8	3
	Sadness	17	**90**	7	10	76
	Neutral	49	10	**121**	16	4
	Anger	11	12	18	**148**	11
	Fear	3	75	7	19	**96**
NLD-2						
Emotional state		Happy	Sadness	Neutral	Anger	Fear
True emotion	Happy	**147**	12	30	11	0
	Sadness	39	**95**	21	16	55
	Neutral	11	14	**118**	20	37
	Anger	8	19	21	**116**	36
	Fear	0	51	14	27	**108**

Table 6: Confusion matrix for the Berlin-DB emotional speech recognition for four feature types.

Prosody		Detected emotion				
Emotional state		Happy	Sadness	Neutral	Anger	Fear
True emotion	Happy	**35**	1	6	23	6
	Sadness	35	**46**	12	1	1
	Neutral	2	8	**56**	3	10
	Anger	14	1	2	**64**	1
	Fear	11	5	13	15	**25**
NLD-1						
Emotional state		Happy	Sadness	Neutral	Anger	Fear
True emotion	Happy	**34**	0	4	18	5
	Sadness	0	**47**	14	0	1
	Neutral	0	3	**67**	0	9
	Anger	20	0	0	**59**	3
	Fear	7	5	8	4	**45**
MFCC						
Emotional state		Happy	Sadness	Neutral	Anger	Fear
True emotion	Happy	**48**	0	0	17	6
	Sadness	0	**59**	3	0	0
	Neutral	2	2	**71**	0	4
	Anger	12	0	0	**69**	1
	Fear	11	5	7	1	**45**
NLD-2						
Emotional state		Happy	Sadness	Neutral	Anger	Fear
True emotion	Happy	**50**	0	6	8	7
	Sadness	3	**48**	8	0	3
	Neutral	8	4	**54**	0	13
	Anger	16	0	0	**64**	2
	Fear	14	4	7	7	**37**

features. This is verified for both Chinese and German emotional speech corpus. Therefore, we can state that the NLD-1 and NLD-2 features extracted in this paper can effectively characterize the emotional information in speech signals.

5.2.3. Experiment Scheme 3: Speech Recognition of Mixed Language Emotion. Prosodic features, MFCC features, NLD-1 features, and NLD-2 features are used to recognize the cross-emotional speech from the Berlin-DB and CASIA in two languages. The recognition results are shown in Table 7. This further validates the efficiency of the extracted features for recognition of emotional states from speech.

From Table 7, we can draw the following conclusions: from the average recognition results with single use of four feature types (prosodic features, MFCC, NLD-1, NLD-2, and NLD-2), the average recognition rate is the highest for NLD-2 and the lowest for the prosodic features. We can conclude that prosodic features are superior for the task of recognition of emotional speech in a single language. Evaluating the results for each individual emotion, we observe that MFCC has a better discriminative power for detecting sadness; NLD-1 can better differentiate neutral emotions; However, NLD-2 provides better distinction between happiness,

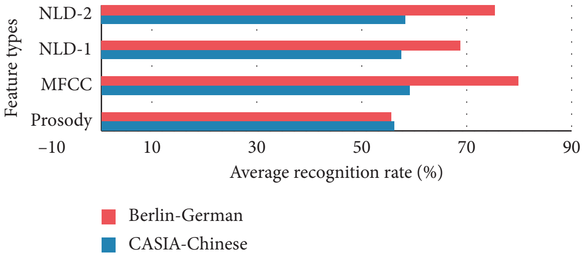

FIGURE 7: Comparison of different features for single language emotional speech recognition.

TABLE 7: Four types of features used to obtain a confusion matrix for the mixed language emotional speech recognition task.

Feature type	Emotional state	CASIA-Chinese	Berlin-German	Average
Prosody	Happiness	42.65	66.67	54.66
	Sadness	52.94	70.00	64.47
	Neutral	48.53	69.23	58.88
	Anger	69.12	62.96	66.04
	Fear	44.12	17.39	33.76
	Average	51.47	57.50	54.49
MFCC	Happiness	55.88	54.17	55.03
	Sadness	52.94	80.00	66.47
	Neutral	75.00	76.92	75.96
	Anger	69.12	85.19	77.16
	Fear	38.26	34.78	36.52
	Average	58.24	66.67	62.46
NLD-1	Happiness	50.00	54.17	52.09
	Sadness	47.06	60.00	53.53
	Neutral	76.47	80.77	78.62
	Anger	73.53	77.78	76.65
	Fear	44.12	69.57	56.85
	Average	55.29	69.17	62.23
NLD-2	Happiness	52.94	70.83	61.89
	Sadness	48.53	80.00	64.27
	Neutral	44.12	69.23	56.68
	Anger	79.41	85.19	82.30
	Fear	54.11	73.91	64.01
	Average	55.88	75.83	65.86
Prosody + MFCC + NLD	Happiness	75.00	72.06	73.53
	Sadness	76.09	72.06	74.08
	Neutral	77.78	73.53	75.66
	Anger	86.96	79.41	83.19
	Fear	79.17	70.59	74.88
	Average	79.17	73.53	76.35

anger, and sadness. Therefore, we can state that NLD demonstrates a better distinction between different emotions of great intensity, such as sadness, neutral, happiness, and anger. From the perspective of feature fusion, we observe that addition of NLD features effectively compensates the chaotic characteristics of the emotional speech signals compared to traditional acoustic linear features. In addition, we also observe that it is partial to using NLD features for characterizing the emotional difference in speech signals. This is because NLD features are obtained by treating the speech signals as a one-dimensional time series and completely ignoring the acoustic features of the emotional speech signals. Therefore, when the NLD features and acoustic features are combined, the effective information in the emotional speech signals can be better described.

6. Conclusion and Further Study

In this paper, based on the chaotic characteristics in the nonlinear generation mechanism of speech signal, aiming at the deficiency of linear feature parameters in speech signal and the limitation of existing time-domain and frequency-domain attribute features in characterizing the integrity of speech information, a nonlinear feature extraction method based on phase space reconstruction theory is proposed, and the chaotic characteristics of speech signal are verified from three aspects: power-spectrum analysis, principal component analysis, and phase space reconstruction. The nonlinear dynamic model is applied for the extraction of speech features. This paper also extracts and evaluates the contribution of NLD features from speech signals. The speech recognition experiments are designed to combine the features of traditional linear acoustics with the NLD features to verify whether this combination can improve the performance of the recognition. From the experimental results for the recognition of isolated words, the addition of nonlinear dynamical features is able to effectively compensate for the chaotic features neglected by the traditional acoustic features. This proves that merging NLD features with acoustic features can better describe the effective information contained in speech signals. From the recognition results of emotional speech, we can observe that while the performance of nonlinear features alone is ideal, we can obtain better recognition rates through feature fusion. For the experimental designed in this paper, the recognition network was developed by combining NLD features with acoustic features. Through our experiments, we demonstrate that while NLD features efficiently compensate for the chaotic characteristics of the emotional speech signals, they are also biased to represent the differences in the emotional speech alone. In future research, we would like to explore the research direction of integrating NLD and acoustic features to generate the strongest combination of the features. Additionally, in view of the high efficiency of NLD features for emotion recognition in mixed languages, the study of cross-database emotion recognition using NLD features is another research direction that needs to be further explored.

Data Availability

The databases used in the manuscript can be found in the following two links, from which you can download: http://emodb.bilderbar.info/docu/#home and http://www.chineseldc.org/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no.61371193), research project supported by the Shanxi Scholarship Council of China (Grant no.201925), and project supported by the http://doi.org/10.13039/501100004480 Natural Science Foundation of Shanxi Province, China (Grant no.201901D111096).

References

[1] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, “Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis,” *Computer Speech & Language*, vol. 41, pp. 116–127, 2017.

[2] A. Mencattini, E. Martinelli, G. Costantini et al., “Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure,” *Knowledge-Based Systems*, vol. 63, no. 3, pp. 68–81, 2014.

[3] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, “Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model,” *Journal of Voice*, vol. 30, no. 6, pp. 757.e7–757.e19, 2016.

[4] J.-C. Wang, C.-H. Lin, E.-T. Chen, and P.-C. Chang, “Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition,” in *Proceedings of the Signal and Information Processing Association Annual Summit and Conference*, vol. 76, pp. 1–14, Hong Kong, China, 2015.

[5] M. Sarria-Paja and T. H. Falk, “Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification,” *Computer Speech & Language*, vol. 45, pp. 437–456, 2017.

[6] C. Thompson, A. Mulpur, V. Mehta, and K. Chandra, “Transition to chaos in acoustically driven flows,” *The Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 2097–2108, 1991.

[7] M. D. Zbancioc, “Using the lyapunov exponent from cepstral coefficients for automatic emotion recognition,” in *Proceedings of the 2014 International Conference and Exposition on Electrical and Power Engineering*, pp. 110–113, IEEE, Iasi, Romania, October 2014.

[8] S. G. Firooz, F. Almasganj, and Y. Shekofteh, “Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals,” *Computers & Electrical Engineering*, vol. 58, pp. 215–226, 2016.

[9] K. López-de-Ipiña, J. Solé-Casals, H. Eguiraun et al., “Feature selection for spontaneous speech analysis to aid in Alzheimer’s disease diagnosis: a fractal dimension approach,” *Computer Speech & Language*, vol. 30, no. 1, pp. 43–60, 2015.

[10] K. López-de-Ipiña, J. B. Alonso-Hernández, J. Solé-Casals et al., “Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of Alzheimer’s disease,” *Neurocomputing*, vol. 150, pp. 392–401, 2015.

[11] L. Xiang and N. Tan, “Method of applying speech multi-features to detect driver fatigue,” *Chinese Journal of Scientific Instrument*, vol. 34, no. 10, pp. 2231–2237, 2013.

[12] J. A. Gómez-García, J. I. Godino-Llorente, and G. Castellanos-Dominguez, “Non uniform embedding based on relevance analysis with reduced computational complexity: application to the detection of pathologies from biosignal recordings,” *Neurocomputing*, vol. 132, no. 7, pp. 148–158, 2014.

[13] T. Lin and Y. G. Yang, “Chaotic time series analysis and its application research,” *Journal of Wuhan University of Technology*, vol. 32, no. 19, pp. 189–192, 2010.

[14] F. Takens, *Detecting Strange Attractors in Turbulence. Dynamical Systems and Turbulence, Warwick 1980*, Springer, Berlin, Germany, 1981.

[15] Z. Jiao, *Improved Feature Extraction Algorithm for ZCPA Speech Recognition*, Taiyuan University of Technology, Taiyuan, China, 2005.

[16] J. Tao, J. Yu, and Y. Kang, “An expressive mandarin speech corpus,” in *Proceedings of the Conference of Oriental COCOSDA*, Kyoto, Japan, 2005.

[17] H. S. Kim, R. Eykholt, and J. D. Salas, “Nonlinear dynamics, delay times, and embedding windows,” *Physica D: Nonlinear Phenomena*, vol. 127, no. 1-2, pp. 48–60, 1999.

[18] F. Burkhardt, “A database of german emotional speech,” in *Proceedings of the INTERSPEECH 2005—Eurospeech, European Conference on Speech Communication and Technology*, pp. 1517–1520, Lisbon, Portugal, September 2005.

[19] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, pp. 30–150, Cambridge University Press, Cambridge, UK, 2004.

[20] G. Zhao and Y. Shi, "Computing fractal dimension and the Kolmogorov entropy from chaotic time series," *Chinese Journal of Computational Physics*, vol. 16, no. 3, pp. 310–315, 1991.

[21] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, "Determining Lyapunov exponents from a time series," *Physica D: Nonlinear Phenomena*, vol. 16, no. 3, pp. 285–317, 1985.

[22] H. E. Hurst, R. P. Black, and Y. M. Simaika, "Long-term storage: an experimental study," *Journal of the Royal Statistical Society*, vol. 129, no. 4, pp. 591–593, 1965.

[23] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing*, vol. 84, no. 3, pp. 65–75, 2012.

[24] Y. Sun, H. Yao, and X. Zhang, "Feature extraction of emotional speech based on chaotic characteristics," *Journal of Tianjin University*, vol. 48, no. 8, pp. 681–685, 2015.

[25] S. S. Vijayarajsolomon, V. Parthasarathy, and N. Thangavelu, "Exploiting acoustic similarities between Tamil and Indian English in the development of an HMM-based bilingual synthesiser," *Iet Signal Processing*, vol. 11, no. 3, pp. 332–340, 2017.

[26] R. M. Kiran and K. Sreenivasa, "Robust pitch extraction method for HMM-based speech synthesis system," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1133–1137, 2017.

[27] L. Wang, S. Nakagawa, and Z. Zhang, "Spoofing speech detection using modified relative phase information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 660–670, 2017.

[28] Z. Ma, Y. Liang, and J. Zhu, "An optic-fiber fence intrusion recognition system using mixture Gaussian hidden Markov models," *IEICE Electronics Express*, vol. 14, no. 5, 2017.

[29] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Age and gender classification from speech and face images by jointly fine-tuned deep neural networks," *Expert Systems with Applications*, vol. 85, pp. 76–86, 2017.

[30] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[31] E. Zarrouk, Y. Benayed, and F. Gargouri, "Hybrid SVM/HMM model for the recognition of Arabic triphones-based continuous speech," in *Proceedings of the 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13)*, Hammamet, Tunisia, 2013.

[32] W. Zhang, D. Zhao, and Z. Chai, "Deep learning and SVM based emotion recognition from Chinese speech for smart affective services," *Software Practice & Experience*, vol. 47, no. 8, pp. 1127–1138, 2017.