Hindawi
Complexity
Volume 2020, Article ID 8810901, 9 pages
https://doi.org/10.1155/2020/8810901

Research Article

A Multiscale Chaotic Feature Extraction Method for Speaker Recognition

Jiang Lin,[1] **Yi Yumei**,[1] **Zhang Maosheng**,[2] **Chen Defeng**,[1] **Wang Chao**,[3] **and Wang Tonghan**[4]

[1]*College of Computer and Information Engineering, Institute of Big Data and Internet Innovation, Hunan University of Technology and Business, Changsha 410205, China*
[2]*School of Mathematics and Statistics, Yulin Normal University, Yulin 537000, China*
[3]*National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan 430072, China*
[4]*School of Software, East China University of Technology, Nanchang 330013, China*

Correspondence should be addressed to Jiang Lin; jlcdf@163.com

Received 2 October 2020; Revised 30 October 2020; Accepted 19 November 2020; Published 3 December 2020

Academic Editor: Karthikeyan Rajagopal

In speaker recognition systems, feature extraction is a challenging task under environment noise conditions. To improve the robustness of the feature, we proposed a multiscale chaotic feature for speaker recognition. We use a multiresolution analysis technique to capture more finer information on different speakers in the frequency domain. Then, we extracted the speech chaotic characteristics based on the nonlinear dynamic model, which helps to improve the discrimination of features. Finally, we use a GMM-UBM model to develop a speaker recognition system. Our experimental results verified its good performance. Under clean speech and noise speech conditions, the ERR value of our method is reduced by 13.94% and 26.5% compared with the state-of-the-art method, respectively.

1. Introduction

Speaker recognition is a biometric recognition technique, which can identify speaker identity according to speaker personality information on a speech signal. From the existing biometric recognition, speaker recognition is one of the most convenient and accessible ones due to the abundance of mobile devices, with a microphone, allowing users to be authenticated across multiple environments and devices [1].

Research in speaker recognition has focused increasingly on enhancing robustness in adverse conditions induced by background noise. Many approaches have been proposed to address these challenges; one of the most successful being the *i*-vector technology [2] used jointly with the probabilistic discriminant analysis (PLDA) back-end [3, 4]. In addition to the new utterance level features and back-ends, robust acoustic features are developed to improve the performance of the speaker recognition system.

The cepstrum features (such as MFCC) of speech are the most distinguishing and first used [5] in speaker recognition. However, under the influence of channel distortion and background noise, the cepstral feature distribution of speech will change arbitrarily, which leads to its weak distinguish ability. Therefore, in the early 1990s, a series of feature compensation techniques were proposed to enhance the generalization ability of speech features in recognition [6–8]. The existing feature compensation is mainly including filter compensation, noise model compensation, and empirical compensation.

The main purpose of filter compensation is to reduce noise or relieve the influence of noise on features. This method is based on the fact that channel and environmental distortions are superimposed on the logarithmic spectrum and the cepstrum domain. Furui S. believe that the variation in the channel is the offset of a single coefficient in the cepstrum vector. Therefore, the cepstrum mean subtraction (CMS) method is used to relieve the influence of the channel

[9]. This method also can reduce the channel noise to a certain extent, but this method also impairs the information on the cepstral coefficient. Unlike the CMS method, the relative spectrum feature is proposed to compensate for rapidly changing channel distortions, and it uses moving average filtering to simulate the exponential decay of the mean subtraction [10]. However, this method was later confirmed to have limited improvements in channel mismatch and additive background noise.

The noise model compensation uses the prior knowledge of the noise spectrum to estimate the parameters of the pure speech through the noise model or the influence of noise on the speech. It mainly uses spectral equalization and spectral subtraction to relieve the influence of noise on features. J. Hansen et al. proposed a multidimensional equalization method to reduce the sensitivity of speech features of noise, thereby improving the distinguishing ability of speech features [11]. S.S. Bharti utilized interframe features to estimate the continuous noise spectrum, which can alleviate the problem of noise spectrum changes caused by single-frame estimation in the original spectrum subtraction, and then uses spectrum subtraction to enhance speech features and improve the robustness of speaker features [12]. The noise compensation model method mainly relies on the mathematical model of noise estimation. Because of the uncertainty of noise change, it is difficult to find a mathematical model with good performance.

Empirical compensation is a data-driven method, which is inherently random. Studies have shown that this method is better than the previous two [13]. This method directly uses spectrum comparison based on experience. In the training phase, to estimate the change between the clear speech and the noisy speech, the difference in the feature vector between the two frames is calculated, and the probability distribution is modeled by adding a bias term of this difference. In the evaluation stage, the minimum mean square error prediction method is adopted, and the bias vector is used to convert the noisy test feature vector into the equivalent clear speech feature vector. Afify M et al. proposed a random mapping method [14], which uses the joint distribution of clear speech and noisy speech feature vectors to generate a Gaussian mixture model, and then uses this joint model distribution to predict clear speech. This prediction method has a significant improvement compared with the previous minimum mean square error.

In summary, the method of feature compensation is to improve the distinguishing ability of speech features for reducing the influence of noise on the features. However, as long as noise exists, this improvement is always difficult to avoid the impact on noise on the recognition performance.

In this paper, a novel multiscale chaotic feature is proposed to speaker recognition. The proposed multiscale chaotic feature is evaluated using a nonlinear dynamic model based on wavelet decomposition (multiresolution analysis (MRA)). In our method, an MRA technique is used to capture more finer spectrum information. In speech feature, harmonic feature is an important factor to distinguish different speakers. Because harmonic can represent speaker's tone, tone information is usually distributed over different frequency components. The wavelet decomposition is an adapt method to capture frequency components. Moreover, we also take into account a chaotic characteristic of speech signal. Speech signal is a nonlinear system on long time series. In addition, this nonlinear characteristic should be reflected in speech features to speaker recognition. The chaotic feature based on the nonlinear dynamic model is used widely to the speech application system. The nonlinear dynamic model has been used in various fields of speech processing area, such as speech steganalysis [15], speech synthesis [16], speech recognition [17], and speech encryption [18]. The proposed feature represents the signal chaotic characteristic at different frequency bands.

2. Proposed Speaker Recognition System

To improve the performance of speaker recognition under adverse conditions induced by background noise, we proposed a multiscale chaotic feature extraction method to enhance the robustness of the recognition system. The proposed speaker recognition system is illustrated in Figure 1.

In our proposed system, a time-domain speech signal is treated with short-time frames of N samples by windowing each frame with, e.g., the hamming window. To relieve the influence of noise, we extract a multiscale chaotic feature (MCF) in each frame, which is comprised of multiresolution analysis and chaotic feature. Multiresolution analysis is implanted by wavelet decomposition and chaotic features including the nonlinear dynamic model and acoustics features. More details will be descripted in Section 3. A Gaussian mixture model (GMM) is used to identify each speaker; here, we introduce a universal background model (UBM) for training the distribution of features that are not related to the speaker. The GMM-UBM model is used widely for speaker recognition [19–21] as a classifier; it is a generalization of the GMM model. The GMM-UBM model firstly performs a pretraining for the current speaker by collecting feature data from other speakers, which can solve the problem of recognition performance declining due to the insufficient feature data of the current train speaker. Then, the pretrained model is fine-tuned to the target speaker model by a maximum a posteriori (MAP) adaptive algorithm [22].

3. Multiscale Chaotic Feature

In the speaker recognition systems, the speech feature is vital to recognition performance. The significant discrimination of the feature can distinguish accurate speakers. The acoustic feature (such as MFCC and LPC) has a strong discrimination to speaker recognition of clear speech. However, the performance of classification will decline sharply if the speech signal is disturbed by noise. In order to reduce the disturbance due to environment noise, we introduced a wavelet decomposition and reconstruction technology to enhance the resolution of speech features on the frequency domain. Take account of the speech signal is a nonlinear system, we extract the chaotic feature by the nonlinear dynamic model

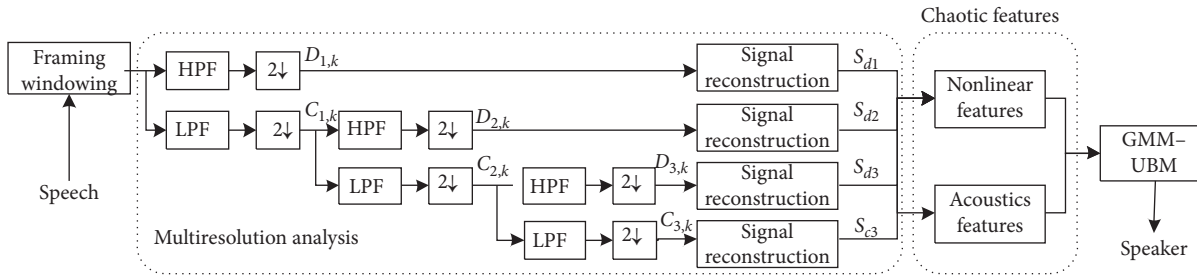

Figure 1: The proposed speaker recognition system using a multiscale chaotic feature.

to improve the recognition rate. In our proposed speaker recognition system, the feature extraction consists of two parts: multiresolution analysis and chaotic feature extraction.

3.1. Multiresolution Analysis. Multiresolution analysis (MRA) is a technique, which forms a set of basis functions through stretch and scale based on wavelets. On large scale, they expanded basis functions to search for significant features, and on smaller scales, they find more details of features. In our system, according to the literature [23] method, the speech signal is decomposed into a number of subband signals by MRA. The MRA is carried out by passing a speech signal $s(n)$ through a series of high pass and low pass filter banks. The speech signal is simultaneously passed through high pass and low pass filters with impulse response $h(n)$ and $g(n)$, respectively. The resulting outputs are the convolution of $s(n)$ with $h(n)$ and $g(n)$, respectively:

$$y_1(n) = s(n) * h(n) = \sum_{k=-\infty}^{\infty} s(k)h(n-k), \tag{1}$$

$$y_2(n) = s(n) * g(n) = \sum_{k=-\infty}^{\infty} s(k)g(n-k). \tag{2}$$

As the literature [23], we selected Daubechies 4 (db4) as the basis function for MRA in this paper. Likewise, we have also tested with other basis functions. We found that the recognition rate achieved with multiscale chaotic features using db4 basis function is higher than those results with other basis functions. Therefore, the db4 basis function is selected for MRA. The filter coefficients h_k and g_k, corresponding to the high pass and low pass filters, respectively, are computed from the following MRA equations [24]:

$$\varphi(n) = \sqrt{2}h_k\varphi(2n-k), \tag{3}$$

$$\psi(n) = \sqrt{2}\sum_k g_k\psi(2n-k), \tag{4}$$

$$s(n) = \sum_k C_{j_o,k}\varphi_{j_o,k}(n) + \sum_{j=1}^{j_o}\sum_k D_{j,k}\psi_{j,k(n)}, \tag{5}$$

where j represents the decomposition scale ($j = 1, 2, \ldots, j_o$) and k denotes the coefficient index of each decomposition scale. At the first level of decomposition ($j = 1$), the detail coefficient $D_{j,k}$ is an output of the high pass filter and the approximation coefficient $C_{j,k}$ is an output of the low pass filter. The detail coefficients and approximation coefficient captured the high frequency and low frequency information, respectively. The approximation band is further decomposed into detail and approximation bands at the next level of decomposition. The repeated decomposition can obtain multiple levels for getting better resolution. In our system, 3-level decomposition is set ($j_o = 3$) as shown in Figure 1. In order to analyse the speech signal to different resolutions, subband signals are reconstructed via each of the approximation and detail coefficients applying inverse discrete wavelet transform (IDWT) [24]. When reconstructing a subband signal, the other subband coefficients are set 0. In this paper, we obtained 4 subbands signals: S_{d1}, S_{d2}, S_{d3}, *and* S_{c3}; these subband signals will be utilized to extract chaotic features.

3.2. Chaotic Feature. In our feature extraction method, acoustic feature and nonlinear feature are extracted. Acoustic feature mainly focuses on speech spectral features (such as MFCC and LPC). Nonlinear feature represents the speech chaotic characteristics using a nonlinear dynamic model.

3.2.1. Acoustic Feature. In acoustic features, we extract Mel frequency cepstrum coefficient (MFCC) and linear prediction coefficient (LPC). MFCC is computed based on the perception characteristics of the human auditory system. In the human auditory system, Mel frequency has a nonlinear relationship to the Hz frequency. We can obtain the Mel spectral feature by the nonlinear relationship as follows:

$$f_{\text{mel}}(f) = 2595 \cdot \log\left(1 + \frac{f}{700Hz}\right), \tag{6}$$

where f denotes the Hz frequency.

LPC represents the frequency envelope of voice; its computation is based on the speech signal digital model. The vocal tract model is a key factor to distinguish different speakers. Therefore, LPC is usually used to represent the vocal tract envelope for various speech recognitions [25, 26]. According to the speech signal digital model, a speech frame signal can be equivalent to a unit pulse sequence to excite the vocal tract. The process is a linear time-invariant system and can be represented as a form of different equations:

$$x(n) = \sum_{i=1}^{p} \alpha_i x(n-i) + e(n), \tag{7}$$

where $x(n)$ is the real signal, the weighting term represents the prediction signal, and $e(n)$ denotes the prediction error. The filter coefficient α_i is calculated according to the minimum mean square error (MSE) criterion of $e(n)$.

3.2.2. Nonlinear Feature. The nonlinear dynamic model is an effective analysis method to study the chaotic characteristics of speech signals. According to this model, the nonlinear characteristics of the speech signal are obtained by processing the speaker signal as an one-dimensional time series. From the Takens embedding theorem, to reconstruct the phase space, one-dimensional speaker signals $(x(1), x(2), \ldots, x(N))$ can be mapped to high-dimensional space by selecting an appropriate minimum delay time τ and embedding dimension m with two parameters. In addition, the high-dimensional spaces after reconstruction are equivalent to the original space [27]. The reconstructed speaker speech signal becomes $X_i = (x(i), x(i+1), \ldots, x(i+(m-1)\tau))$, where $i = 1, 2, \ldots, N-(m-1)\tau$. The key point of chaotic feature extraction includes the analysis of speaker speech signal in a high-dimensional space, the extraction of nonlinear feature parameters under the voice dynamic model.

(1) The minimum delay time: the minimum delay time describes the correlation between the neighboring components of the speaker speech signal $(x(1), x(2), \ldots, x(N))$. In order to reconstruct the phase space of one-dimensional speaker speech signals, we calculate the minimum delay time τ and embedding dimension m by the C-C method [28].

(2) Maximum Lyapunov exponent: the Lyapunov exponent represents the average change rate of the local convergence or divergence of adjacent orbits in the phase space. The maximum Lyapunov exponent λ_1 denotes the speed of orbit convergence or divergence. When $\lambda_1 > 0$, the larger the value of λ_1, the greater the rate of orbital divergence and the greater the degree of chaos. We use the small data size method to compute Lyapunov exponent [28]. The calculation method is as follows:

(1) Calculate the average period P by fast Fourier transform on the time series $(x(1), x(2), \ldots, x(N))$.

(2) Calculate the minimum delay time τ and embedding dimension m by the C-C method.

(3) Reconstruct phase space of series $(x(1), x(2), \ldots, x(N))$ and denote it as $X_i = (x(i), x(i+1), \ldots, x(i+(m-1)\tau))$, $i = 1, 2, \ldots, N-(m-1)\tau$. Then, find the nearest neighbor $X_{i'}$ of each point X_i in the phase space and limit the short separation. Define the distance $d_i(0)$ from the i-th point of the nearest point $X_{i'}$ in its adjacent track:

$$d_i(0) = \left\| X_i - X_{i'} \right\| \quad \left| i - i' \right| > P. \tag{8}$$

(4) Find each point X_i in the phase space and calculate the distance next n unit times of the adjacent point pair:

$$d_i(n) = \left| X_{i+n} - X_{i'+n} \right| \quad n = 1, 2, \ldots, \min(M-i, M-i'). \tag{9}$$

(5) If the orbit, which locates on the nearest point in the neighborhood domain, diverges from an exponential rate of λ_1, then

$$d_i(n) = d_i(0) e^{\lambda_1 n T_s}, \tag{10}$$

where T_s is the sampling period. Taking the logarithm of both sides with the equation, we get

$$\ln d_i(n) = \ln d_i(0) + \lambda_1 n T_s. \tag{11}$$

Take the average of the logarithmic difference in the distance between all adjacent points, which is

$$\text{ave}\lambda(n) = \frac{1}{qT_s} \sum_{i=1}^{q} \left(\ln d_i(n) - \ln d_i(0) \right), \tag{12}$$

where q is the numbers of nonzero $d_j(i)$. Last, we use the least squares method to fit λ_1:

$$\lambda_1 = \frac{\sum_{n=1}^{\min(M-i, M-i')} n\lambda(n)}{\sum_{n=1}^{\min(M-i, M-i')} n^2}. \tag{13}$$

(3) Correlation dimension and Kolmogorov entropy: correlation dimension and Kolmogorov entropy are both nonlinear representation quantities under the nonlinear dynamic model. The correlation dimension describes the self-similar structure of the system. Kolmogorov entropy accurately describes the degree of confusion of the distribution probability of time series. We use the G-P algorithm [29] to calculate the correlation dimension and Kolmogorov entropy at the same time. The algorithm is as follows:

(1) Firstly, we calculate the correlation integral $C(r, m)$ and the $C(r, m) - r$ curve. Reconstruct the m-dimensional phase space. Then, given a critical distance r, search the phase point pair whose distance is less than r, and further calculate the ratio of all phase points. Last, we get the correlation integral function as follows:

$$C(r, m) = \frac{1}{M(M-1)} \sum_{i,j=1, i \neq j}^{N} \theta\left(r - \left\| X_i - X_j \right\| \right), \tag{14}$$

where m is the embedding dimension, M is the total number of phase points, $M = N-(m-1)\tau$, and θ is the Heaviside function, which satisfies $\theta_{(z)} = \begin{cases} 0, & z \leq 0, \\ 1, & z > 0. \end{cases}$.

(2) The correlation dimension $D(m)$ is derived by the G-P algorithm as follows:

$$D(m) = \frac{\ln C(r, m)}{\ln r}. \tag{15}$$

Draw the $\ln C(r,m) - \ln r$ curve, and take the slope of the approximate straight line part; the slope is as the correlation dimension D.

(3) The Kolmogorov entropy formula is derived by the G-P algorithm as follows:

$$K = \frac{1}{m\tau} \ln \frac{C(r,m)}{C(r,m+1)}. \quad (16)$$

(4) Hurst exponent: the Hurst exponent (H) may measure the long-term memory of the time series. It can also find the evolution trend of a time series converging on one direction. In addition, the range of H values is 0~1. If $H > 0.5$, it means that the time series has a long-term autocorrelation and means a greater correlation between the context time series. $H < 0.5$ means no autocorrelation in a time series. Extracting the Hurst exponent feature of speaker speech can reflect the level of relevance to the speaker's speech change, so this paper selects the Hurst exponent as one of the nonlinear features. Hurst proposed the H exponent and introduced the rescaled range analysis method [30] to calculate the value. This method is a nonparametric statistical method, which is not affected by the time series distribution.

4. Evaluation and Analysis

4.1. Experiment Setting. In order to evaluate the performance of the proposed speaker recognition system, we carried out 2 experiments under clear speech and noise speech, respectively. To feature selection, we set 3 combinations: (1) acoustic features, (2) nonlinear feature, and (3) chaotic features (acoustic + nonlinear feature). We selected the *i*-vector speaker recognition model in the literature [31] as the baseline system. In the *i*-vector model, MFCC feature is extracted and further obtain the supervector as the speaker feature. Currently, it is the state-of-the-art of speaker feature. The details of the experiment settings are listed in Table 1. In our experiments, the hardware environment is Intel i7 CPU and 8 GB memory. The software environment is Windows 10 OS with 64 bits, the Matlab 2016a, and Voice speech tool package as a develop tool.

4.2. Corpus. TIMIT corpus is used to evaluate our system. The corpus is composed of 630 speakers from different regions, each with 10 sentences. The length of each sentence is 3~5 s, and the sampling frequency is 16 KHz with 16 sampling bits. In the TIMIT database, 630 people are divided into 462 and 168 according to a 3 : 1 ratio, which are used to train the background model and test the recognition system, respectively. Among the voice data of the 168 speakers tested, 9 sentences of each speaker were randomly selected as training data and 1 sentence as test data. The noise library is Noise-92 noise library. White noise and babble noise are selected as experimental objects. Since the two selected noises are associated with daily life scenes, these noises will also be mixed in real life application scenes. Therefore, this noise experiment has certain representative and feasibility.

4.3. Preprocessing. Speech signal is a nonstationary time-varying signal. Preprocessing must be performed first before speech analysis and feature extraction. The preprocessing usually includes endpoint detection, pre-emphasis, windowing, and framing processing. In this paper, endpoint detection adopts a double-threshold method based on zero-crossing rate and energy. The pre-emphasis is carried out by a first-order FIR high pass filter, and the pre-emphasis coefficient is set to 0.97. The frame length is set to 20 ms with 50% overlap.

4.4. Feature Extraction. In the feature extraction phase, we extracted the acoustic features and nonlinear feature for each subband (4 subbands in one frame). Acoustic feature consists of 12-order MFCC coefficients and 4-order LPC coefficients and then calculated statistical features of each coefficient for classifying the statistical features including skewness, kurtosis, mean, variance, and median. Nonlinear feature comprises the minimum delay time, correlation dimension, K entropy, maximum Lyapunov exponent, and Hurst exponent. We also calculate its statistical characteristics, like as maximum value, minimum value, mean, median, and variance. The speaker feature is listed in Table 2.

In order to eliminate the problem of internal dependence of speech features due to different dimensionality, a mean normalization is carried out on features as follows:

$$x = \frac{x - \mu}{\sigma}, \quad (17)$$

where μ and σ denote mean and standard deviation, respectively.

4.5. Results and Analysis. In order to evaluate the discriminability of the proposed feature, we carried out 2 group experiments according to Table 1. Take account of the validity of equal error rate (EER) on speaker recognition evaluation, we selected EER as a metric to evaluate the performance of chaotic features. For EER, more less value, the better performance. In our experiments, the mixture numbers of GMM-UBM are set to 512, and the iteration of the EM train algorithm is 10 times, and the dimension of the T matrix of the *i*-vector model is set to 100. All parameters are obtained by iterative optimization.

4.5.1. Results and Analysis under Clean Speech Condition. The results are listed in Table 2 under clean speech condition. It can be seen from Table 3, if using acoustic feature alone, the ERR is 2.562%, which has a similar performance compared with the *i*-vector model. This shows LPC feature is helpful to identify different speakers because LPC can represent the vocal tract envelope. The EER of nonlinear feature is 2.833%, which is the worst performance compared with other features. However, we find the chaotic feature is the better performance, and the EER value is reduced by 14%

Table 1: The combinations of experiment settings.

Experiment group	Speech type	Noise type	Acoustic feature	Nonlinear feature	Chaotic feature (acoustic + nonlinear)
1	Clean speech	—	√		
				√	
					√
2	Noise speech	White	√		
				√	
					√
		Babble	√		
				√	
					√

Table 2: The speaker chaotic features.

Feature type	Feature ID	Feature parameter	Statistical description
Acoustic feature	1~60	12-order MFCC	Skewness, kurtosis, mean, variance, median
	61~108	4-order LPC	
Nonlinear feature	109~113	Minimum delay time	Maximum, minimum, mean, median, and variance
	114~118	Correlation dimension	
	119~123	K entropy	
	124~128	Maximum Lyapunov exponent	
	129~133	Hurst exponent	

Table 3: The EER value under clean speech condition.

Feature	Acoustic feature	Nonlinear feature	Chaotic feature	*i*-vector model
EER (%)	2.562	2.833	2.149	2.497

Table 4: The EER value under white noise speech condition.

SNR (dB)	Acoustic feature	Nonlinear feature	Chaotic feature	*i*-vector model
30	2.156	2.785	1.59	1.963
25	3.167	3.389	2.531	3.144
20	5.148	5.987	4.121	4.907
15	9.746	10.182	7.115	9.755
10	17.929	18.114	14.152	18.122
5	29.535	30.157	25.144	32.696
0	37.005	39.142	20.504	33.134
Average	14.96	15.68	10.74	14.82

compared with *i*-vector, which shows that the speech chaotic characteristic has a good discrimination.

4.5.2. Results and Analysis under Noise Speech Condition. The purpose of this group experiment is to evaluate the robustness of chaotic features with noise speech. We selected stationary noise (white noise) and nonstationary noise (babble noise) as the disturb signal and set different disturb degrees. The SNR is set to 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB, and 30 dB.

Table 4 and Figure 2 show the EER value of speaker recognition under white noise speech condition. From the results, the acoustic feature, nonlinear feature, and *i*-vector model are the similar average EER value. However, the chaotic feature obtained a better performance of recognition compared with other features. The EER values reduced by 27.53% compared with the *i* -vector model. The good performance is attributed to the chaotic characteristic. This also shows that the speech nonlinear features can relieve the disturbance of noise and improve the robustness of speaker recognition.

Table 5 and Figure 3 give the ERR results under nonstationary noise babble condition. Similar with white noise disturb, there is also a good robustness of chaotic feature.

Table 6 shows the average EER values for all experiments. From the results, compared with the *i*-vector model, the EER value reduced by 13.94% and 26.5% under clean speech and noise speech conditions, respectively. Therefore, we believe the following:

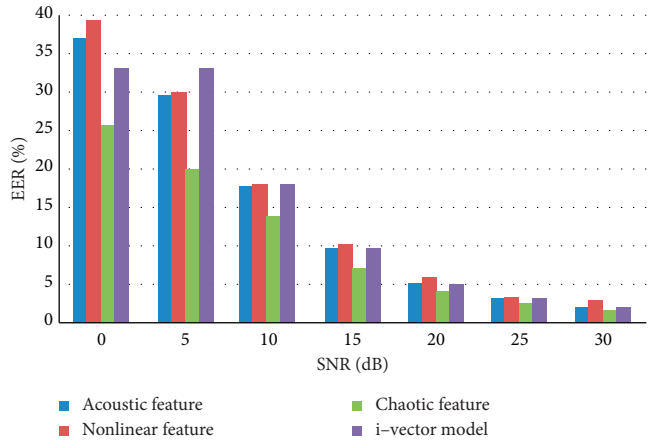

Figure 2: EER results under white noise speech condition.

Table 5: The EER value under babble noise speech condition.

SNR (dB)	Acoustic feature	Nonlinear feature	Chaotic feature	*i*-vector model
0	27.589	30.142	20.092	26.172
5	23.897	25.563	13.095	16.077
10	20.481	21.792	11.607	16.981
15	12.074	13.346	8.103	10.252
20	6.547	5.897	2.918	4.266
25	2.801	2.105	1.691	2.678
30	1.822	2.185	0.901	1.587
Average	13.60	14.43	8.34	11.14

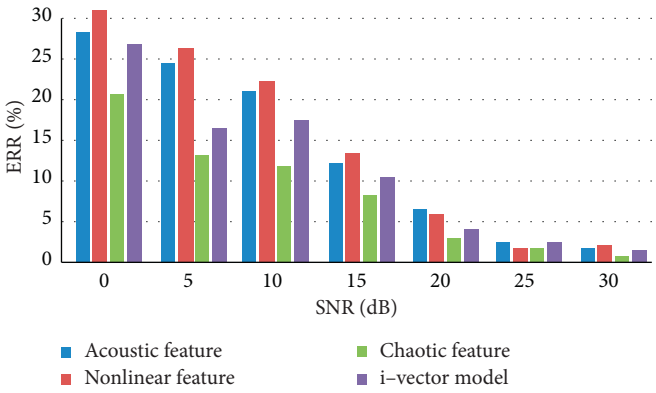

Figure 3: EER results under babble noise speech condition.

Table 6: The average EER value in all experiments.

Feature	Speech type	Acoustic feature	Nonlinear feature	Chaotic feature	*i*-vector model	Compare with *i*-vector
EER (%)	Clean	2.562	2.833	2.149	2.497	↓13.94%
	White	14.96	15.68	10.74	14.82	↓26.5%
	Babble	13.60	14.43	8.34	11.14	

(1) Acoustic feature has a good performance of recognition with clean speech. MFCC and LPC have a perfect discrimination to different speakers. However, the recognition performance will decline if the speech signal is disturbed by environment noise.

(2) Speech chaotic characteristic based on the nonlinear dynamic model has a good compensation for the acoustic feature under environment noise condition. That is, the nonlinear feature can better distinguish different speakers in a noise environment. These benefits of the multiresolution analysis techniques can better capture the frequency information of speakers.

(3) In the proposed method, we only extracted 5 nonlinear parameters, and the feature combination is not optimized. We suggest that feature optimization may improve the robustness of recognition.

5. Conclusion

In this paper, we proposed a novel multiscale chaotic feature for speaker recognition. The MRA technique is used to capture the frequency information of a speaker under environment noise condition. We extracted the nonlinear feature based on speech chaotic characteristics to improve the robustness of recognition. The experiment results show that this method is valid. Therefore, we believe the speech chaotic characteristic is a robust feature to various speech application systems, such as speech recognition and speech emotion recognition. In this paper, the proposed feature is not optimized, and the feature optimization will be the next work.

Data Availability

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

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 under grants 61762005, 61702472, and 61671335, Hunan Provincial Natural Science Foundation of China under grant 2019JJ40144, Scientific Research Project of the Hunan Province Education Department of China under grants 18A304 and 18B338, and the Key Laboratory of Hunan Province for New Retail Virtual Reality Technology under grant 2017TP1026.

References

[1] V. Vestman, D. Gowda, M. D. Sahidullah, P. Alku, and T. Kinnunen, "Speaker recognition from whispered speech: a tutorial survey and an application of time varying linear prediction," *Speech Communication*, vol. 99, pp. 62–79, 2018.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[3] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of the 11th IEEE International Conference on Computer Vision*, pp. 1–8, Rio De Janeiro, Brazil, October 2007.

[4] P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen, "From single to multiple enrollment *i*-vectors: practical PLDA scoring variants for speaker verification," *Digital Signal Processing*, vol. 31, pp. 93–101, 2014.

[5] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[6] B. A. Acero, "Acoustical and environmental robustness in automatic speech recognition," Ph. D thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 1993.

[7] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994.

[8] R. J. Mammone, X. Xiaoyu Zhang, and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, p. 58, 1996.

[9] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[10] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[11] J. H. L. Hansen and M. A. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 407–415, 1995.

[12] S. S. Bharti, M. Gupta, and S. Agarwal, "A new spectral subtraction method for speech enhancement using adaptive noise estimation," in *Proceedings of the 2016 3rd International Conference on Recent Advances in Information Technology*, Dhanbad, India, March 2016.

[13] A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Transactions on Speech & Audio Processing*, vol. 11, no. 6, pp. 568–580, 2003.

[14] M. Afify, Y. X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1325–1334, 2009.

[15] O. H. Kocal, E. Yuruklu, and I. Avcibas, "Chaotic-type features for speech steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 651–661, 2008.

[16] H. A. Patil and T. B. Patel, "Chaotic mixed excitation source for speech synthesis," in *Proceedings of the 15th Annual Conference of International Speech Communication Association*, Singapore, Asia, September 2014.

[17] V. Pitsikalis and P. Maragos, "Speech analysis and feature extraction using chaotic models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, USA, May 2002.

[18] K. Gopakumar and F. J. Farsana, "Speech encryption algorithm based on nonorthogonal quantum state with hyperchaotic keystreams," *Advances in Mathematical Physics*, vol. 2020, Article ID 8050934, , 2020.

[19] J. McLaughlin, D. A. Reynolds, and T. P. Gleason, "A study of computation speed-UPS of the GMM-UBM speaker recognition system," in *Proceedings of the European Conference on Speech Communication & Technology*, Budapest, Hungary, September 1999.

[20] Z. Xiong, T. F. Zheng, Z. Song, and W. Wu, "Combining selection tree with observation reordering pruning for efficient speaker identification using GMM-UBM," in *Proceedings of the International Conference on Acoustics Speech & Signal Processing*, Philadelphia, PA, USA, March 2005.

[21] S. Anchal, B. Mukhopadhyay, M. Parvatini, and S. Kar, "GMM-UBM based person verification using footfall signatures for smart home applications," in *Proceedings of the IEEE Global Conference on Signal and Information Processing*, Ottawa, Canada, November 2019.

[22] A. K. Sarkar and S. Umesh, "Investigation of speaker-clustered UBMs based on vocal tract lengths and MLLR matrices for speaker verification," in *Proceedings of the Speaker and Language Recognition Workshop-Odyssey*, Brno, Czech Republic, July 2010.

[23] S. Deb and S. Dandapat, "Multiscale Amplitude feature and significance of enhanced vocal tract information for emotion classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 802–815, 2019.

[24] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1995.

[25] L. R. Rabiner, M. M. Sondhi, and S. E. Levinson, "A vector quantizer combining energy and LPC parameters and its application to isolated word recognition," *AT & T Bell Laboratories Technical Journal*, vol. 63, no. 5, pp. 721–735, 1984.

[26] J. B. T. Curipe and A. H. Camacho, "Feature extraction using LPC-residual and MelFrequency cepstral coefficients in forensic speaker recognition," *International Journal of Computer and Electrical Engineering*, vol. 5, no. 1, pp. 40–43, 2013.

[27] F. Takens, "Detecting strange attractors in turbulence," in *Lecture Notes in Math*, pp. 366–381, Springer, New York, NY, USA, 1981.

[28] J. Lv, A. Lu, and S. Chen, *Chaotic Time Series Analysis and its Application*, Wuhan University Press, Wuhan, China, 2002, in Chinese.

[29] 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, 1999, in Chinese.

[30] H. E. Hurst, "Long-term storage: an experimental study," *Journal of the Royal Statistical Society*, vol. 129, no. 4, pp. 591–593, 1965.

[31] D. G. Da Silva and C. A. Medina, "Evaluation of MSR identity toolbox under conditions reflecting those of a real forensic case (forensic_eval_01)," *Speech Communication*, vol. 94, pp. 42–49, 2017.