Hindawi
Complexity
Volume 2021, Article ID 8876005, 10 pages
https://doi.org/10.1155/2021/8876005

Research Article

Articulatory-to-Acoustic Conversion of Mandarin Emotional Speech Based on PSO-LSSVM

Guofeng Ren, Jianmei Fu, Guicheng Shao, and Yanqin Xun

Department of Electronics, Xinzhou Teachers University, Xinzhou 034000, China

Correspondence should be addressed to Guofeng Ren; renguofeng926@sina.com

Received 13 September 2020; Revised 4 November 2020; Accepted 21 January 2021; Published 31 January 2021

Academic Editor: Wei Wang

The production of emotional speech is determined by the movement of the speaker's tongue, lips, and jaw. In order to combine articulatory data and acoustic data of speakers, articulatory-to-acoustic conversion of emotional speech has been studied. In this paper, parameters of LSSVM model have been optimized using the PSO method, and the optimized PSO-LSSVM model was applied to the articulatory-to-acoustic conversion. The root mean square error (RMSE) and mean Mel-cepstral distortion (MMCD) have been used to evaluate the results of conversion; the evaluated result illustrates that MMCD of MFCC is 1.508 dB, and RMSE of the second formant (F2) is 25.10 Hz. The results of this research can be further applied to the feature fusion of emotion speech recognition to improve the accuracy of emotion recognition.

1. Introduction

In recent decades, the development of artificial intelligence (AI) has been very rapid, and human-computer interaction technology needs that there would be harmoniously communicated relationship between the human being and the intelligent machine [1, 2]. All of emotional speech processing methods must use the human pronunciation mechanism as the basement; thus speech is pronounced with success by the movements of articulators, for example, the tongue, lips, and jaw [3]. The mapping between kinematics and acoustics has been taking form through the gathering of vast pronounced experience.

At the same time, since the notion of "emotion calculator" was proposed by the MIT Media Lab, many physiological signals have been successively applied as characteristic information in the research of emotional speech recognition, so as to help computers better analyze the emotional state of speakers from their speech signals [1]. However, as an important part of emotional speech generation, the kinematic data of articulators have not been widely used in the research of speech emotion recognition [4]. The reasons are mainly as follows: data collection of vocal organs is difficult [5]; the relationship between the articulatory features of the articulators and the mature acoustic features is unclear, and the fusion of multiple types of features is difficult [6]. There is no standard bimodal emotional speech database combining articulatory and acoustic data [7]. However, the articulatory data of the articulators are not susceptible to noise, and their features have the advantage of being more robust than the acoustic features [8].

In order to integrate articulatory features into the feature network of speech emotion recognition to improve the rate of speech emotion recognition, the domestic and foreign researchers have studied the conversion method of speech articulatory-to-acoustic features. Common conversion methods can be divided into statistical mapping and data-driven methods [9–12]. Among them, the typical method of statistical mapping is codebook-based method, which builds codebook to store the mapping pair of acoustic-articulatory features and applies the algorithm to find the optimal mapping pair to confirm the relationship between articulatory and acoustic features. This method was first proposed in 1996 by Hogden et al. [13], who constructed the mapping relationship between acoustic features and articulatory features, and the method used vector quantization to encode

the features and needs huge data to realize accurate mapping.

The idea of data-driven method is to construct a model through training data to realize the conversion of articulatory and acoustic features, and typical examples of such method are neural networks and Gaussian mixture models (GMM). As early as 2002, Richmond [9] successfully mapped acoustic features to articulatory features using neural network, constructed a mixed density network containing multilayer perceptions, and applied the data of MOCHA-TIMIT database to study the acoustic-to-articulatory inversion and finally realized the inversion effect with root mean square error value of 1.4 mm. As long as the data are sufficient, the method can achieve a more ideal effect. However, the number of hidden layers and the number of nodes in each hidden layer involving the structure of the network need special attention. Meanwhile, the method has not fully considered the effect of constraints such as phoneme and time. In addition, GMM-based inversion methods have problems of oversmoothing and overfitting, which affect the effect of inversion [14, 15].

The least squares support vector machine (LSSVM) is a new type of support vector machine (SVM), which is often used to solve the problem of pattern classification and function estimation. The particle swarm optimization (PSO) is a kind of optimization algorithm based on animal swarm motion research, which is commonly used for path planning and parameter optimization. Fairee et al. [16] implemented acoustic-to-articulatory inversion using PSO algorithm in 2015, and Yogesh et al. [17] used a hybrid PSO algorithm to study the emotional and pressure recognition of speech and achieved good results. With the development of deep learning method, BiLSTM-CNN-word attention theory is adopted to realize the articulatory-to-acoustic conversion in Mandarin and realize the recognition effect of RMSE of F2, 22.10 Hz, but the conversion effect under different emotions has not been studied [18].

Thus, combining PSO optimization and LSSVM, articulatory-to-acoustic conversion on emotional speech in Mandarin has been studied in this paper. The paper is organized as follows. Firstly, this paper reviews related work on articulatory-to-acoustic conversion and LSSVM and PSO in Section 2. Secondly, the detailed method we proposed is elaborated in Section 3, and Section 4 illustrates our experiments and results. Section 5 discusses and concludes the work of this paper.

2. Related Work

For exploring the articulatory-to-acoustic conversion and improving the conversion effect, we proposed PSO-LSSVM model to achieve the conversion, which includes LSSVM and PSO. Meanwhile, Gaussian mixture model (GMM), as a classical conversion model [8], has been adopted to comprising the conversion effect with LSSVM and PSO-LSSVM in our study. In this section, we will give a brief introduction about GMM, LSSVM, and PSO.

2.1. Articulatory-to-Acoustic Conversion Based on GMM Used in Emotional Speech. GMM has been used for the calculation of the joint probability density of acoustic and articulatory characters to achieve the conversion. The conversion can be described by the following equation:

$$y_t = \sum_{i=1}^{M} p(\lambda_i|t_i)\, p(y_i|x_i, \lambda_i). \tag{1}$$

Here, M represents the amount of Gaussian mixture element, $p(\lambda_i|x_i)$ shows the probability of acoustic feature vector x_i, and $p(y_i|x_i, \lambda_i)$ shows full covariance matrix of conditional Gaussian distribution. $x = (x_1, x_2, ..., x_M)$ and $y = (y_1, y_2, ..., y_M)$ are considered as articulatory and acoustic characters, respectively. For the articulatory characters of frame i, the first-order dynamic features can be shown as follows:

$$\Delta x_i = -0.5x_{i-1} + 0.5x_{i+1}. \tag{2}$$

The articulatory features and the first-order dynamic features are spliced as the input character vector $X_i = [x_i^M, \Delta x_i^M]^M$, and then output vector $Y_i = [y_i^M, \Delta y_i^M]^M$ will be obtained. Thus, the joint probability distribution of input and output vectors can be shown as follows:

$$P(Z_i|\vartheta) = P(X_i, Y_i|\Theta) = \sum_{j=1}^{N} \partial_j \mathbb{N}\left(Z_i, \mu_j^2, \sum_j^2\right), \tag{3}$$

$$Z_i = \begin{bmatrix} X_i \\ Y_i \end{bmatrix}, \mu_j^Z = \begin{bmatrix} \mu_j^X \\ \mu_j^Y \end{bmatrix}, \Sigma_j^Z = \begin{bmatrix} \Sigma_j^{XX} & \Sigma_j^{XY} \\ \Sigma_j^{YX} & \Sigma_j^{YY} \end{bmatrix}. \tag{4}$$

Here, $Z_i = [X_i^M, Y_i^M]$ shows the joint vector of articulatory and acoustic characters, and N represents the number of Gaussian elements, $\theta = \alpha_j, \mu_j^Z, \Sigma_j^Z | j = 1, 2, ..., N$ shows the parameters of GMM, and α_j, μ_j^Z, and Σ_j^Z represent weight, mean, and covariance of Gaussian element j, respectively. Among them, model parameter Θ can be estimated relying on maximum likelihood estimation algorithm (MLEA) [6].

During the conversion, input articulatory features are considered as $X = [x_1, x_2, ..., x_M]$, and output acoustic feature are considered as $Y = [y_1, y_2, ..., y_M]$; y^* can be calculated relying on the MLE, which is shown in the following equation:

$$y^* = \arg\max_y p(Y|X, \Theta) \text{s.t.} Y = Wy, \tag{5}$$

where W is dynamic window coefficient matrix.

If we only focus on Gaussian element, the Gaussian element will be obtained through Maximum Posterior Probability, as shown in the following equation:

$$j^* = \arg\max p(j|X, \Theta). \tag{6}$$

It is supposed that the frames are independent; for frame i,

$$p(j|X_i,\Theta)=\frac{p(j|\Theta)p(X_i|j,\Theta)}{p(X_i|\Theta)}=\frac{\alpha_j\mathbb{N}\left(X_i|\mu_j^X,\Sigma_j^{XX}\right)}{\Sigma_n^N\alpha_n\mathbb{N}\left(X_i|\mu_n^X,\Sigma_n^{XX}\right)}, \tag{7}$$

$$p\left(Y_i|j^*,X_i,\Theta\right)=\mathbb{N}\left(Y_i|\mu_{j^*,i}^{y|x},\Sigma_{j^*}^{y|x}\right). \tag{8}$$

Among them, $\mu_{j^*,i}^{y|x}$ and $\Sigma_{j^*}^{y|x}$ can be considered as mean and covariance matrix, respectively, and they are calculated using the two following equations:

$$\mu_{j^*}^{y|x}=\mu_{j^*}^{Y}+\Sigma_{j^*}^{YX}\Sigma_{j^*}^{YX^{-1}}\left(X_i-\mu_{j^*}^{X}\right), \tag{9}$$

$$\Sigma_{j^*}^{y|x}=\Sigma_{j^*}^{YY}-\Sigma_{j^*}^{YX}\Sigma_{j^*}^{XX}\Sigma_{j^*}^{XY}. \tag{10}$$

Thus, we can obtain the output vector using maximum likelihood criterion, as shown in the following equation:

$$y^*=\left(W^T\Sigma^{y-1}W\right)^{-1}W^T\Sigma^{y-1}\mu^y. \tag{11}$$

Here, Σ^y denotes square matrix and μ^y can be calculated through $\mu_{j^*,t}^{y|x}$.

2.2. LSSVM. Support vector machine (SVM) algorithm is a data-driven method of machine learning-based statistical theory proposed by Vapnikli [19] in the 20th century, the thought of which is to map the input vector to a high-dimensional feature space by nonlinear mapping method and solve the nonlinear features and regression problems of the original space using linear machine learning methods. The original SVM algorithm can be used to solve the classification problem in the pattern recognition. Later, the SVM regression algorithm can be used to solve the nonlinear regression problem by defining the insensitive loss function.

Support vector machine (SVM) can transform feature prediction problems into linear programming or quadratic programming problems to solve local minimum problems for achieving global optimization. The kernel function is used to replace the inner product of the high-dimensional feature space, so that the high-dimensional space problem can be solved properly. At the same time, the SVM method can balance the weight relation between the fitting ability and the generalization ability by adjusting the penalty parameters through the capacity, which has the advantages of simple structure and good sparsity. SVM can not only achieve the goal of structural risk minimization but also effectively deal with nonlinear, high-dimensional, local minimization and overlearning problems and has been successfully applied to speech recognition, speech feature prediction, and so forth.

In order to solve the disadvantages of SVM, Suykens [20] proposed the least squares support vector machine (LSSVM) theory. Compared to SVM, LSSVM is more suitable for parameter prediction and pattern recognition with small sample size [21].

It is supposed that training sample set is $D=\{x_i,y_i\}_{i=1}^N$, where N is the number of training samples, $x_i\in R^m$ is m dimensions input, and $y_i\in R$ is target output size of 1 dimension. The nonlinear function estimated problem can be transferred to the linear function estimation in high-dimensional feature space using the following equation:

$$f(x)=w^T\varphi(x)+b. \tag{12}$$

Here, $w=[w_1,\ldots,w_n]^T$ is coefficient vector of weight, and $\varphi(\bullet)=[\varphi_1(.),\ldots,\varphi_n(.)]^T$ is mapping function. This regression problem may represent an optimizing problem about equality constraint, the optimization aim of which is shown in the following equations:

$$\min_{w,\lambda,e}J(w,e)=\frac{1}{2}w^Tw+\frac{\lambda}{2}\sum_{i=1}^{n}e_i^2, \tag{13}$$

$$\text{s.t.}\,y^i=w^T\varphi(x_i)+b+e_i,i=1,\ldots,n. \tag{14}$$

Then, the Lagrange method is used to solve the above optimization problems:

$$L(w,b,e,a)=J(w,e)-\sum_{i}^{N}a_i\left(w^T\varphi(x_i)+b-e_i-y_i\right), \tag{15}$$

where a_i $(i=1,\ldots,n)$ is a Lagrange multiplier.

The partial derivative of equation (15) is obtained according to the optimization conditions:

$$\begin{aligned}
&\frac{\partial L}{\partial w}=0\longrightarrow w=\sum_{i=1}^{N}a_i\varphi(x_i),\\
&\frac{\partial L}{\partial b}=0\longrightarrow \sum_{0}^{N}a_i=0,\\
&\frac{\partial L}{\partial e_i}=0\longrightarrow a_i=\gamma e_i,\\
&\frac{\partial L}{\partial w}=0\longrightarrow w^T\varphi(x_i)+b+e_i-y_i=0.
\end{aligned} \tag{16}$$

Then, the kernel function is defined according to the Mercer condition:

$$k(x_i,y_i)=\varphi^T(x_i)\varphi(y_i). \tag{17}$$

It can be obtained from equations (16) and (17), which is shown as follows:

$$\begin{bmatrix}
0 & 1 & \ldots & 1\\
1 & k(x_1,x_1)+\dfrac{1}{\gamma} & \ldots & k(x_1,x_n)\\
\ldots & \ldots & & \ldots\\
1 & k(x_n,x_1) & \ldots & k(x_n,x_n)+\dfrac{1}{\gamma}
\end{bmatrix}\times\begin{bmatrix}b\\a_1\\\ldots\\a_n\end{bmatrix}=\begin{bmatrix}0\\y_1\\\ldots\\y_n\end{bmatrix}. \tag{18}$$

Finally, the nonlinear model of LSSVM is obtained as follows:

$$f(x) = \sum_{i=1}^{N} a_i k(x, x_i) + b. \tag{19}$$

2.3. PSO. PSO is an evolutionary computation proposed by Dr. Eberhart and Dr. Kennedy in 1995. On the basis of observing the activity behavior of animal clusters, the algorithm uses the shared information of the particles in the group to analyze the movement of each particle and generates the evolution process from disorder to order in the solution space, so as to obtain the optimal solution.

PSO algorithm can initialize the problem to a group of random particles and then find the optimal solution through iteration. In each iteration, the particle updates itself by tracking individual and global extremes, where the individual extreme value can be expressed by the following equation:

$$P_i(t) = \left(P_{i1}(t), P_{i2}(t), ..., P_{id}(t)\right). \tag{20}$$

The global extreme value can be expressed as follows:

$$P_g(t) = \left(P_{g1}(t), P_{g2}(t), ..., P_{gd}(t)\right). \tag{21}$$

In addition, if a small part of the population is regarded as the nearest neighbor of the particle instead of the whole, then the extremum found in all the nearest neighbors of the particle becomes the local extremum, which is expressed as

$$P_l(t) = \left(P_{l1}(t), P_{l2}(t), ..., P_{ld}(t)\right). \tag{22}$$

When the particle iterates for t times, its position information is represented by equation (23) and its velocity is represented by equation (24).

$$X_i(t) = \left(X_{i1}(t), X_{i2}(t), ..., X_{id}(t)\right), \tag{23}$$

$$V_i(t) = \left(V_{i1}(t), V_{i2}(t), ..., V_{id}(t)\right). \tag{24}$$

Equations (25) and (26) are used to update the velocity and position of particles iteratively.

$$\begin{aligned} V_{ik}(t+1) = {} & wV_{ik}(t) + c_1 r_1 \left(P_{ik}(t) - X_{ik}(t)\right) \\ & + c_2 r_2 \left(P_{gk}(t) - X_{gk}(t)\right), \end{aligned} \tag{25}$$

$$X_{ik}(t+1) = X_{ik}(t) + X_{ik}(t+1). \tag{26}$$

In both of these equations, t is the number of iterations of the particle at this time, ω is the inertia weight coefficient of the particle, c_1 and c_2 are learning factors, which illustrate acceleration weight from particle to individual extremum and global extremum, respectively, and r_1 and r_2 are random digits between 0 and 1; $k = 1, , 2, ..., d$.

In this paper, RMSE of the model calculation result has been adopted to the adaption function of PSO, obtaining optimized parameters γ and σ.

3. Methods

3.1. Articulatory-to-Acoustic Conversion Model Based on LSSVM. The algorithm flow of feature conversion model based on LSSVM is shown in Figure 1, and the specific process is as follows:

(1) Articulatory features $x = (x_1, x_2, ..., x_M)$ and acoustic features $y = (y_1, y_2, ..., y_N)$ were synchronously extracted from the bimodal emotional speech database

(2) The training set and test set were randomly divided, and the characteristic data were normalized

(3) The LSSVM model was initialized, and the radial basis kernel function was selected for the kernel function in the model

(4) Parameter optimization of LSSVM model was carried out by cross validation, where γ and σ were mainly optimized parameters

(5) The articulatory features in the test set were imported into the LSSVM model for parameter prediction, and then acoustic feature probability sequence was generated

(6) Maximum likelihood estimation algorithm (MLEM) was used to estimate acoustic feature parameters

In the LSSVM model, selection of radial basis kernel function firstly requires setting the value of kernel parameter σ and penalty factorγ. The smaller the value of γ is, the stronger the generalization ability will be and the better the smoothing ability of the model will be, but the fitting ability will decrease. The greater γ is, the smoother the training model will be and the stronger the generalization ability will be. In this paper, PSO algorithm is used to optimize the above two parameters in the LSSVM algorithm.

3.2. Optimization of PSO Algorithm. Figure 2 shows the flowchart of optimizing penalty factor γ and kernel parameter σ in LSSVM model with PSO algorithm.

It can be seen from the figure that the main process of PSO algorithm optimization is as follows:

(1) Initialization of algorithm, where specific content includes the determination of fitness function, population, and movement speed.

(2) Calculating fitness function and scale.

(3) Determining the optimal solution of the algorithm by calculating the extreme value of particle position and velocity and updating the velocity and position.

(4) Acquiring output parameters σ and γ according to the optimal solution of the system.

4. Experiments and Results

4.1. Material. The data used in this paper were from the TYUT bimodal Mandarin emotional speech database. Based on the theory of experimental phonetics and emotional speech, it is recorded, screened, and preprocessed by us [22].

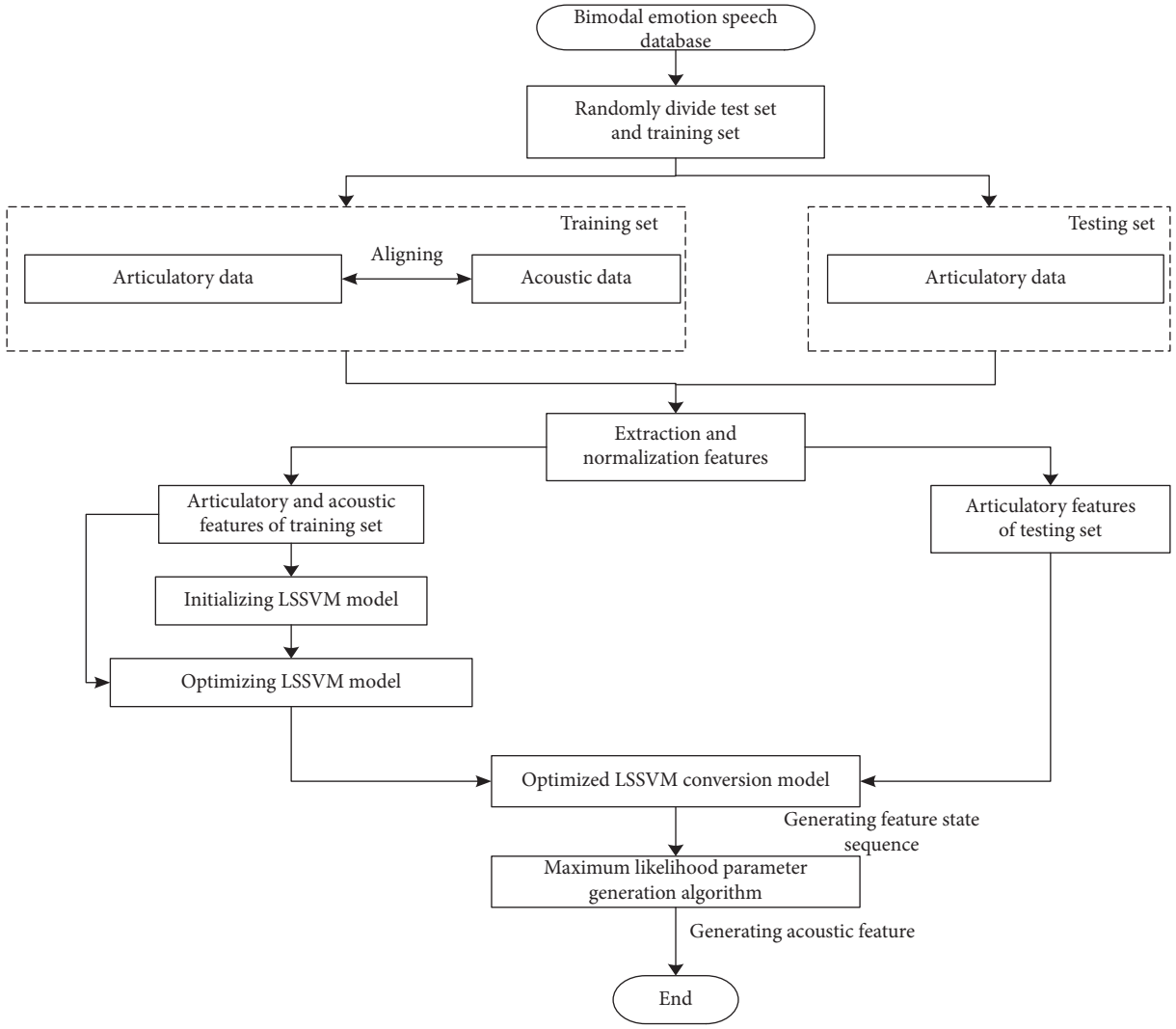

Figure 1: Articulatory-to-acoustic conversion model based on LSSVM.

The articulatory data and acoustic data of the participants in this database have been collected synchronously by the 3D electromagnetic articulography of Carstens, Germany, as shown in Figure 3 [23].

In this paper, 334 disyllable-word materials from 10 participants were selected for research, including four emotions, happiness, anger, sadness, and neutral, which are shown in Table 1.

Firstly, the articulatory data and acoustic data of emotional speech have been synchronized, and the frame length was 4 ms. Secondly, the acoustic data and articulatory data of the speech were extracted by the feature, respectively. The acoustic feature is extracted by the 12-dimensional MFCC and the second formant (F2). The articulatory features were extracted into eight dimensions, namely, the velocities of the tip of the tongue, the mid of the tongue, and the velocities of the tongue root in the up-down and forward-backward directions, as well as the lip protrusion [24] and the lip aperture [25].

Among them, lip aperture and lip protrusion are secondary features, which are determined by the position information of the four sensors attached to the upper and lower lips. Lip protrusion is the distance from the upper incisor to the upper lip calculated from the Euclidean distance between the two sensors. The lip aperture can be obtained from the following equation:

$$\text{LipAperture} = \frac{(\text{UL}_z - \text{LL}_z) - (\text{UL}_z - \text{LL}_z)_{\text{closedposition}}}{(\text{UL}_z - \text{LL}_z)_{\max}}, \tag{27}$$

where UL_z and LL_z show the z-coordinate values of the upper and lower lips, respectively, closedposition represents the distance of the upper and lower lips on the z-axis when the lips are closed, and max represents the distance of the upper and lower lips on the z-axis at the maximum lip opening.

4.2. Partition and Preprocessing of Datasets. The test set and training set have been divided using random selection method; 224 pieces of data have been selected as training

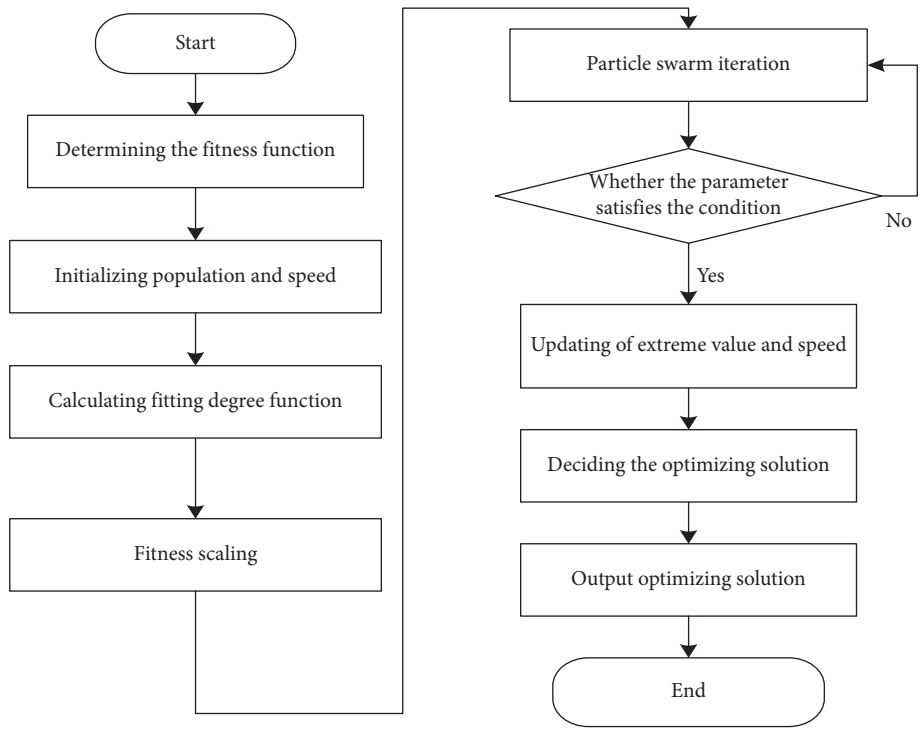

Figure 2: Flowchart of optimizing γ and σ using PSO.

Figure 3: Acquisition diagram of bimodal signal.

Table 1: Number of pieces of data in the database after selection.

Emotion type	Anger	Happiness	Neutral	Sadness
Number	67	106	78	83

data, and 112 pieces of data as test data. Then, features in test and training sets were normalized.

4.3. Model Comparison of EMA-to-F2 Conversion on Emotional Speech. In the EMA-to-F2 conversion, the GMM-based, LSSVM-based, and PSO-LSSVM-based methods have been compared. The RMSE and Collection Coefficient between the true and prediction data have been adopted to measure the conversion result. Adaption degree of PSO was used to reflect the optimization ability of PSO algorithm.

The fitness curve of PSO algorithm is shown in Figure 4.

In Figure 4, the learning factors c_1 and c_2 of PSO algorithm are both 1.5, the number of particles is 30, and the number of iterations is 300. The optimal fitness was achieved after 80 iterations of optimization.

The conversion results of the test set are shown in Figures 5 and 6. Among them, Figure 5 is the conversion result with LSSVM, and Figure 6 is the conversion result with PSO-LSSVM algorithm. In order to visually compare the conversion results of data, only 80 frames are randomly selected in the figure and the converted acoustic feature is the second formant F2 for observation.

It can be seen from the figure that the algorithm optimized with LSSVM by PSO is better than that using LSSVM algorithm. Furthermore, the error analysis of the conversion results is carried out. Here, RMSE is used as the evaluation standard for the conversion of kinematic characteristics to F2, which is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N\sigma^2}\sum_{i=1}^{N}\left(X_i - Y_i\right)^2}, \tag{28}$$

where X_i and Y_i are the true and predicted acoustic features, respectively, i represents the serial number of the frame, N represents the amount of frames in the test set, and σ^2 represents expectancy of the conversion model.

The conversion accuracy of data of various emotional characteristics is different, which is shown in Table 2. It can be seen from the table that the conversion effect of LSSVM

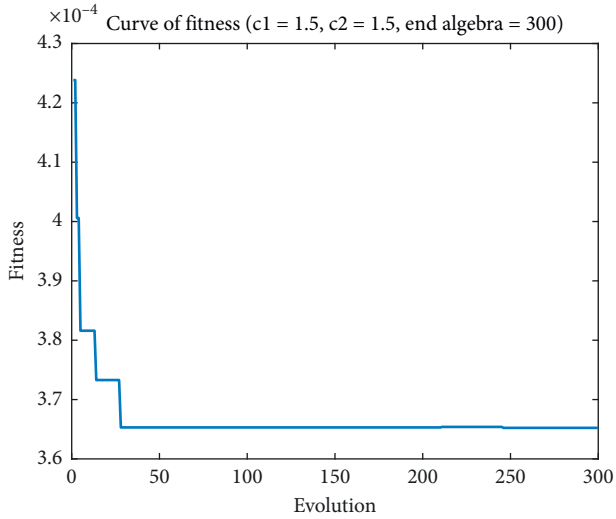

Figure 4: Fitness curve of PSO optimization.

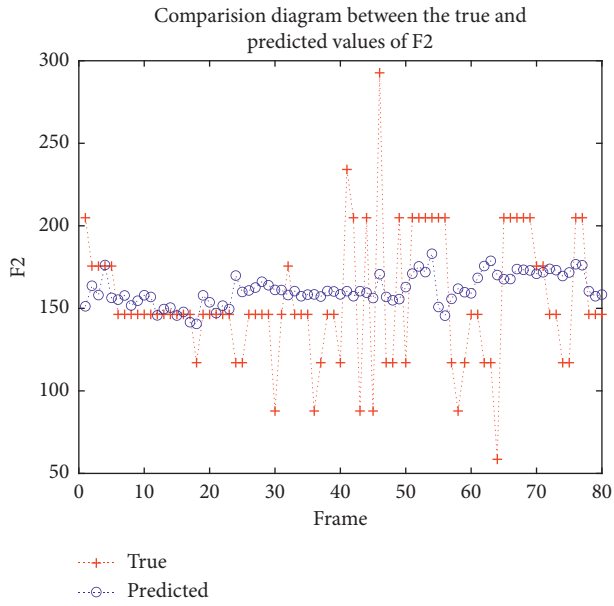

Figure 5: Conversion results using LSSVM.

algorithm optimized by PSO is significantly better than that of traditional LSSVM algorithm, while the conversion result of LSSVM algorithm is significantly better than that of GMM. The RMSE of the three conversion models are, respectively, 25.10 Hz, 30.20 Hz, and 36.52 Hz. In addition, anger and happiness are more effective emotions compared to sadness and neutral.

4.4. Model Comparison of EMA-to-MFCC Conversion on Emotional Speech. Since the converted MFCC in this paper is a multidimensional feature, MMCD [26] has been selected in this paper as the parameter to evaluate the conversion result of articulatory features to the 12-dimensional MFCC feature, which was defined as the mean of the Euclidian distance between the converted value of all frames and the real value.

Here, we compared the performances of the GMM-based, LSSVM-based, and PSO-LSSVM-based methods, and the comparison results are shown in Table 3.

It can be seen from the table that the conversion effect of LSSVM algorithm optimized by PSO is significantly better than that of traditional LSSVM algorithm, while the conversion result of LSSVM algorithm is significantly better than that of GMM. The MMCD of the three transformation models are 1.508 dB, 1.825 dB, and 2.384 dB, respectively. Moreover, both GMM and PSO-LSSVM conversion models shows that the conversion effect of anger and happiness is

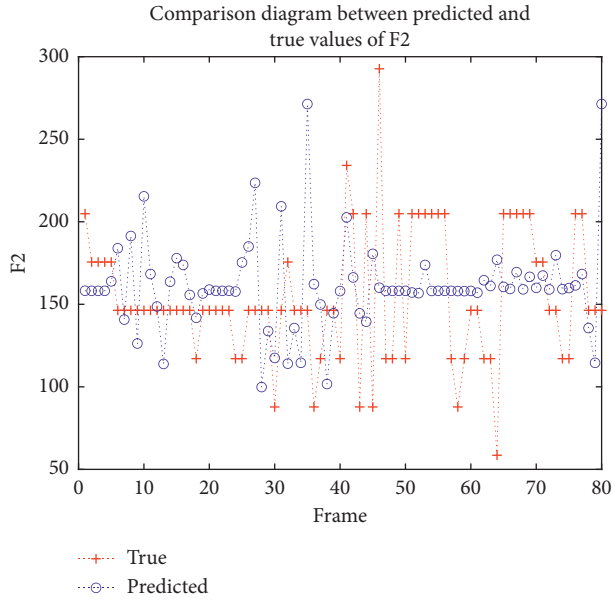

Figure 6: Conversion results of PSO-LSSVM.

Table 2: Comparison of RMSE in articulatory-to-acoustic conversion based on GMM, LSSVM, and PSO-LSSVM (Hz).

Acoustic feature	Emotion type	Number of samples in the test set	GMM	LSSVM	PSO-LSSVM
F2	Anger	29	32	28	23
	Happiness	32	31	26	22
	Neutral	28	43	35	26
	Sadness	23	42	33	31
	Average		36.52	30.20	25.10

Table 3: Comparison of MMCD in articulatory-to-acoustic conversion based on GMM, LSSVM, and PSO-LSSVM (dB).

Acoustic feature	Emotion type	Number of samples in the test set	GMM	LSSVM	PSO-LSSVM
MFCC	Anger	29	2.21	1.98	1.44
	Happiness	32	2.12	1.69	1.49
	Neutral	28	2.46	2.03	1.50
	Sadness	23	2.88	1.96	1.63
	Average		2.384	1.825	1.508

better than that of sadness and neutral, and the conversion result of sadness is slightly better than that of anger in LSSVM conversion result.

4.5. Application of the Converted Features into Speech Emotion Recognition. To verify whether the converted features obtained in this paper can distinguish the emotion type in bimodal speech, in this section, we used the converted features and true features in emotion recognition system based on SVM.

The data used in this section are exactly the same as that used in the PSO-LSSVM feature conversion system and have been derived from the bimodal emotional speech database. In the experiment, the training set and test set of data were randomly divided, and one-third of the data were selected as the test data and two-thirds as the training data. The recognition results are shown in Table 4.

As can be seen from Table 4, the average recognition rates of the converted F2 and MFCC features are 11.47% and 9.42% lower than the original articulatory features, respectively, which do not reach the recognition rate of 50%. At the same time, by comparing the converted MFCC with the MFCC features extracted from the corresponding audio data, the emotion recognition rate is 6.76% lower. Compared with F2 features extracted from audio data, the emotional recognition rate of converted F2 is 10.65% lower. The reason why the recognition rate of the transformed feature is lower than that of the real feature is the error between the transformed feature and the real feature, which requires further improvement of the conversion method to improve the conversion accuracy.

5. Discussion and Conclusion

In this paper, a kind of articulatory-to-acoustic conversion based on LSSVM has been proposed, and PSO optimization algorithm was used to optimize the model parameters, so as to realize the conversion of articulatory features in Mandarin

Table 4: Recognition results on bimodal emotional speech database based on different features (%).

Feature type	Emotion type				Average rate of recognition
	Anger	Happiness	Neutral	Sadness	
Original articulatory features	62.72	69.23	38.1	50	55.01
Converted F2	48.67	40.51	49.87	35.1	43.54
Converted MFCC (12-D)	53.07	41.2	50.25	38.23	45.69
Fused conversion features	58.43	47.22	54.11	38.1	49.47
True F2	54.21	53.91	56.54	52.11	54.19
True MFCC (12-D)	53.03	52.32	56.22	48.23	52.45
Fused true feature	55.46	52.98	59.54	56.11	56.02

bimodal emotional speech to 12-dimensional MFCC and the second formant F2. The experimental results show that the LSSVM model optimized by PSO algorithm can get high precision conversion results. In addition, in this paper, the application of conversion features to the emotion recognition of bimodal speech has been compared with the result of emotion recognition of articulatory features and real acoustic features. The average recognition rate of converted features is lower than that of real features and original features, especially the recognition rate of happiness and sadness emotions. This indicates that although the conversion accuracy of this paper has been improved to some extent compared with the previous methods, the feature conversion of emotional speech cannot meet the requirements for features of emotional speech recognition. Therefore, the rate of emotional recognition is far below 50%. Therefore, it is necessary to continue to improve the algorithm in the future conversion method research to improve the conversion effect. This also indicates that the features generated by the conversion system cannot be applied to emotion recognition as independent features but can be integrated with kinematic and acoustic features for emotion recognition.

The research of articulatory-to-acoustic conversion of emotion speech can lay a foundation for the research of bimodal emotion speech recognition. Of course, there are still some deficiencies in this paper. Firstly, the conversion accuracy of the feature can be further improved. Secondly, the content of target transformation features can be further enriched.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

Thanks are due to all the subjects in current experiment, to Guicheng Shao for technical assistance, to Jianmei Fu and Yanqin Xun for modal design, and to Jianzheng and Dong Li for assistance in data collection. This work was supported by Science and Technology Project of Xinzhou Teachers University (2018KY15 and 2018KY22), the Educational Reform Innovation Project of Shanxi Province of China (J2019174), and Academic Leader Project of Xinzhou Teachers University.

References

[1] W. J. Han, H. F. Li, and C. Y. Wang, "Research and analysis on speech emotion information visualization modeling," *Journal of Yanshan University*, vol. 34, pp. 128–132, 2010.

[2] Z.-J. Chen and W.-H. Wang, "Study on the articulation mechanism of emotional speech," *Journal of Changchun Education Institute*, vol. 30, pp. 61-62, 2014.

[3] G. E. Dahl, D. Dong Yu, L. Li Deng, and A. Acero, "Context-Dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[4] J. Kim, S. Lee, and S. Narayanan, "Detailed study of articulatory kinematics of critical articulators and dependent articulators of emotional speech," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, p. 2549, 2011.

[5] D. Erickson, A. Abramson, and K. Maekawa, "Articulatory characteristics of emotional utterance in spoken English," in *Proceedings of the ICSLP*, Beijing, China, October 2000.

[6] A. Ji, J. J. Berry, and M. T. Johnson, "Vowel production in Mandarin accented English and American English: kinematic and acoustic data from the Marquette University Mandarin accented English corpus," *Speech Communication*, vol. 19, p. 060221, 2013.

[7] L. Xu and M. X. Xu, "Chinese emotional speech database for the detection of emotion variation," in *Proceedings of the NCMMSC*, Urumchi China, August 2009.

[8] A. Ji, "Speaker independent acoustic-to-articulatory inversion," *Doctor, Electrical and Computer Engineering*, Marquette University, Milwaukee, WI, USA, 2015.

[9] K. Richmond, "Preliminary inversion mapping results with a new EMA corpus," in *Proceedings of the INTER SPEECH*, Brighton, UK, September 2009.

[10] H. Lachambre, L. Koenig, and R. André-Obrecht, "Performances of unsupervised HMM in acoustic-to-articulatory inversion," in *Proceedings of the ICASSP*, Vancouver, BC, Canada, May 2013.

[11] H. Behbood, S. A. Seyyedsalehi, H. R. Tohidypour, M. Najafi, and S. Gharibzadeh, "A novel neural-based model for acoustic-articulatory inversion mapping," *Neural Computing and Applications*, vol. 21, no. 5, pp. 935–943, 2011.

[12] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, 1978.

[13] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: new conclusions based on human data," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1819–1834, 1996.

[14] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using Neural Networks," in *Proceedings of the INTER SPEECH*, San Francisco, CA, USA, September 2016.

[15] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation," *Speech Communication*, vol. 99, pp. 161–172, 2018.

[16] S. Fairee, B. Sirinaovakul, and S. Prom-On, "Acoustic-to-articulatory inversion using particle swarm optimization," in *Proceedings of the International Conference on Electrical Engineering/electronics*, Mexico, May 2015.

[17] C. K. Yogesh, M. Hariharan, N. Ruzelita et al., "A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal," *Expert Systems with Applications*, vol. 69, pp. 149–158, 2017.

[18] G. Ren, G. Shao, and J. Fu, "Articulatory-to-Acoustic conversion using BiLSTM-CNN word-attention-based method," *Complexity*, vol. 2020, Article ID 4356981, 10 pages, 2020.

[19] Y. J. Zhang, W. Xu, L. B. Tang, and etal, "Modelling of base station energy consumption system based on sliding window PSO-LSSVM," *Journal of Hunan University (Natural Sciences)*, vol. 44, pp. 122–128, 2017.

[20] W. Qiu, Q. Tang, H.-J. Tang, and X. Shao, "Study on weighing sensor's nonlinear compensation of the moisture instrument based on PSO-LSSVM," *Chinese Journal of Scientific Instrument*, vol. 38, pp. 757–764, 2017.

[21] C. Jin and G.-Y. Chen, "Emotion recognition of multimodal physiological signals based on optimized LSTSVM," *Electronic Technology and Application*, vol. 44, pp. 112–116, 2018.

[22] X.-Y. Zhang, T. Zhang, and Y. Sun, "Emotional speech database optimization and quantitative annotation based on PAD model," *Journal of Taiyuan University of Technology*, vol. 48, pp. 470–475, 2017.

[23] G.-F. Ren, X.-Y. Zhang, D. Li, and J. Z. Yan, "Design and evaluation of Mandarin bi-modal emotion speech database," *Modern Eletronics Technique*, vol. 41, pp. 182–186, 2018.

[24] H. Saito, "Lip movements for an unfamiliar vowel: Mandarin front rounded vowel produced by Japanese speakers," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 6, pp. s1558–s1565, 2016.

[25] E. E. Ma and C. R. Wang, "Study of lip feature point location in visual speech synthesis," *Computer Engineering and Application*, vol. 46, pp. 190–192, 2014.

[26] B. Erika, Z. Frank, A. Bistra, and M. Bernd, "Mel-cepstral distortion of German vowels in different information density contexts," in *Proceedings of the INTER SPEECH*, Stockhom, Sweden, August 2017.