Journal of Network Intelligence
Taiwan Ubiquitous Information

Volume **10**, Number **2**, May **2025**

Research on English Voice Conversion Technology Based on GMM and Deep Learning

Xiao-Xing Liu*

Department of Foreign Language Education and Teaching
Hebei Finance University, Baoding 071051, P. R. China
xiaoxingliu2023@126.com

Mei-Ling Woo

Saint Paul University, Tuguegarao City, 3500, Cagayan, Philippines
rd8250@163.com

*Corresponding author: Xiao-Xing Liu
Received March 19, 2024, revised August 03, 2024, accepted December 19, 2024.

Abstract. *The voice conversion technology refers to a technology that converts the voice of one speaker (source speaker) into the voice of another speaker (target speaker) without changing the content of the voice, and in recent years, with the proposal of deep learning algorithms, the artificial intelligence technology has developed rapidly, and the voice conversion technology has also improved the fidelity of the transformed voice greatly due to the use of deep learning. The voice conversion technology also adopts the deep learning model to enhance the standard of the converted voice. In this paper, we study the related technology of English voice conversion and propose an English voice conversion model based on Gaussian mixture model (GMM) and deep learning. For the problems of unclear and low naturalness of the converted voice, this paper improves the feature vector by using multi-layer RBF neural network, which optimizes the too smooth phenomenon in the conversion process of voice features. The joint probability density method is utilized for training the GMM model, and the feature parameters of the source speaker and the target speaker are regularized dynamically in time. The experimental findings indicate that the naturalness and clarity of the converted voice are significantly improved after the improvement, and the conversion results with high similarity are obtained. The research results of this paper will provide theoretical support and practical guidance for the improvement of English voice conversion technology, encourage the utilization of voice synthesis technology in the field of intelligent voice assistant, virtual presenter, etc., and make a positive contribution to the enhancement of human-computer interaction experience.*
Keywords: English voice conversion; GMM; RBF neural network; Deep learning

1. **Introduction.** Speech conversion refers to the process of converting one sound into another, with the requirement that the sources of these two sounds are different and the meaning conveyed by the speech cannot be changed [1]. Voice conversion technology belongs to a branch of natural language processing in the area of artificial intelligence, and the research of voice conversion technology involves knowledge in many fields such as linguistics, acoustics, voice signal processing, etc., which is a cross-disciplinary research technology. Voice conversion technology can change the timbre of voice [2, 3]. Voice conversion technology was first proposed in the 1960s, but at that time did not receive much attention, until the last few decades due to the deep learning algorithms, scholars have

found that the deep learning model can significantly enhance the efficiency of voice conversion, and thus the voice conversion technology has entered a rapid development track. Voice conversion technology consists of two parts, the learning stage and the conversion stage, the learning stage will extract the feature parameters of the source speaker's voice and the voice of the target speaker for training to get the corresponding conversion rules [4, 5], the conversion phase is to convert the feature parameters of the source speaker to get the feature parameters of the target speaker by using the above-mentioned conversion rules obtained in the learning phase. Finally, the vocoder synthesizes the feature parameters into the target voice. Currently, research in this field is focused on two main directions: voice conversion technology, one is to realize high-quality converted voice in the case of a small amount of corpus. The second is real-time voice conversion. Currently, most speech conversion technologies require a lot of training corpora, and the training process is time-consuming and laborious. Therefore, if the above two points are realized, the application of voice conversion technology can be taken to the next level. In order to make the converted voice more closely resemble the target voice, scholars engaged in voice conversion research have done a lot of research experiments, and put forward a voice conversion method based on neural networks and deep learning and other cutting-edge technologies [6, 7].

1.1. **Related work.** Since the 1990s, voice conversion techniques have been widely researched and achieved at home and abroad. The earliest method is Abe et al. [8] in 1988, which uses vector quantization to realize speaker conversion. However, the method is discontinuous and the converted voice's quality is not good. After that, many other conversion techniques were proposed, but the results were unsatisfactory. Stylianou et al. [9] proposed a voice transformation system based on GMM in 1998, which was also the first time in this field that continuous statistical model was introduced into the voice conversion system to realize the conversion of spectral features. This approach led to a research boom in the next decade or so in realizing spectral mapping with statistical models. These results make GMM-based voice conversion a basic control method in the field of voice conversion. However, the GMM-based conversion function is based on statistical averaging, which causes the spectrum to appear over-smoothed, with loss of spectral details, resulting in degradation of the quality and naturalness of the converted voice.

In order to provide auditory quality, conversion methods based on spectral distortion have been proposed one after another [10, 11, 12], which convert voice with high naturalness and have better subjective auditory quality of voice. However, the transformation of the features of the spectrum is not accurate enough, and the confirmation of the converted voice is not high. Helander et al. [13] proposed to apply the partial least squares method to the GMM, which avoids the problem of overfitting. On the basis of JD-GMM, Toda et al. [14] proposed the ML-GMM conversion method, which estimates the maximum likelihood of the trajectory of the voice conversion and introduces dynamic features and total variation, which effectively reduces the conversion spectral distortion and improves the listening quality.

In recent years, there have been studies proposed to obtain the final conversion function by nearest neighbor matching of feature vectors [15], joint temporary GMM conversion and cyclic iterative updating, which achieved better conversion results. So far, there are still latest research results published continuously. Erro et al. [16] proposed a GMM voice conversion function modified in the cepstrum domain, which not only performs voice conversion, but also makes the parameters in the conversion intelligible. Such a conversion function can be applied not only to synthesize high-quality converted voice, but also to compare the differences between the voice.

With deep neural network technology shining in various fields of information processing, Chen et al. [17] proposed a new model GTDNN based on deep neural network for voice conversion, which trains a four-layer deep neural network layer-by-layer by a cascade of one Bernoulli bi-directional associative memory and two restricted Boltzmann machines. Experimental results demonstrate the method achieves significant results and effectively reduces oversmoothing. Nirmal et al. [18] proposed a fast voice conversion method using a generalized regression neural network. The method uses vocal tract shape, excitation signals and long-term rhythmic properties for voice conversion. Wu et al. [19] proposed a template-based sparse representation for voice conversion. The method uses a weighted sum of templates to recover the converted voice, which well preserves the spectral details with good auditory quality, and at the same time restricts the sparsity of the weights. Then, a spectral compression factor and a residual compensation technique are introduced to improve the listening quality.

1.2. **Motivation and contribution.** In summary, voice conversion techniques relying on GMM and deep learning models have gained recognition and endorsement from numerous scholars, leading to the introduction of various enhancements [20, 21, 22]. Combined with the current technological development of voice conversion and the possibility of future applications, the use of GMM and deep learning models to further study voice conversion will become a major trend.

RBF neural network can model complex nonlinear mapping function very effectively, while GMM can capture the probability distribution of input data well. By combining them, the mapping function and its probability distribution between the source audio and the target audio can be jointly modeled more accurately, thus improving the quality and naturalness of speech conversion. This paper mainly focuses on the current English voice conversion problems such as unclear converted voice and low naturalness, and the primary contributions are as follows: this paper improves the feature vector by using multi-layer RBF neural network, and then optimizes the phenomenon of over-smoothing in the process of converting voice features. The experimental findings demonstrate a substantial enhancement in the naturalness and clarity of the converted voice post the improvements, resulting in conversion outcomes of high similarity.

2. Relevant theoretical analysis.

2.1. **Voice conversion fundamentals.** Voice conversion primarily focuses on exploring the transformation of speaker characteristics between distinct individuals. The fundamental principle is to guarantee that the converted voice retains the semantic information characteristics of the source voice while encompassing the identity information features of the target voice throughout the conversion process, so that the converted voice closely resembles the identity of the target voice. The voice conversion system mainly establishes the mapping relationship of voice features, and then completes the corresponding voice feature conversion according to the mapping relationship, in which the main voice feature parameters include: spectral envelope, resonance peaks, pitch and fundamental frequency, etc. The voice conversion system consists of learning and conversion phases, as shown in Figure 1.

The main purpose of the training stage is to get the conversion rules, the first thing to do in the training phase is to extract the characteristic parameters of the source and target voice, the general method used to extract the feature parameters in this phase is MFCC, and then use the DTW algorithm to align the characteristic parameters, and then put the parameters into the training model for training, and finally get the conversion rules. After that, the feature parameters obtained in the previous step are put into the training

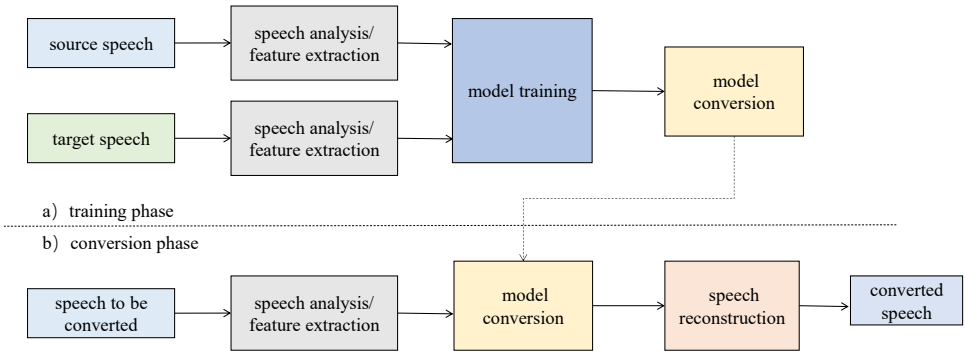


Figure 1. Voice conversion system

model for training, and finally the conversion rules corresponding to the source and target speakers' voices are obtained. In the training stage, different training models can lead to different conversion effects, because the performance of the training models can be differentiated between high and low. A good training model can complete the training in a shorter time, and the corresponding conversion rules can make the converted voice clearer and smoother. Therefore, it is especially important to select a suitable conversion model in this stage.

The main purpose of the conversion stage of voice conversion technology is to obtain the spectral characteristic. In conversion stage, the feature parameters are firstly extracted by the feature extraction method, subsequently, the feature parameters of the source voice are transformed into those of the target voice utilizing conversion rules obtained in the training stage. Finally, the parameters of the target voice can be synthesized into the target voice by putting them into the vocoder.

2.2. **Gaussian mixture model.** GMM is the use of Gaussian probability density function to accurately quantify things [23], Gaussian mixture model was proposed in 1886, Gaussian mixture model has the ability to use a small number of parameters to simulate the data, and it is the fastest training algorithm among all the mixture learning algorithms. Therefore, Gaussian mixture model plays an important role in voice recognition, voice synthesis, and voice conversion, etc. Gaussian mixture model has many similarities with K-means algorithm, for example, both need to specify the value of K, both need to use the EM algorithm to solve, and both can only converge to the local optimum. Its advantage over K-means is that K-means can only classify each sample into one class, while GMM can give the probability of a sample for all classes. GMM can be used not only for clustering, but also for estimating the probability density, and can also be used to generate new samples. Understand the generative process: there are K Gaussian distributions, each distribution is given a weight, and whenever a piece of data is generated, a distribution is randomly selected in proportion to the weight, and the data is generated according to that distribution. A single Gaussian model can be used to represent a single piece of data. It was mentioned above that a Gaussian mixture model can decompose any one thing into a model consisting of one or several functions based on Gaussian probability densities, which means that no matter what kind of distribution has what kind of regularity that the data that we want to test obeys, we can simulate it by one or more single Gaussian models. Generally speaking, the distribution of Gaussian function is called Gaussian distribution or normal distribution, which is a probability distribution that plays a crucial role in the domains of engineering, mathematics, computers, statistics, and so on.

From the above description we already know the concept of a single Gaussian distribution, which is defined as a random variable X obeying a Gaussian distribution with mathematical expectation μ and variance σ^2. This is denoted as $\mathcal{N}(\mu, \sigma^2)$. The mathematical expectation μ refers to the mean (arithmetic mean), and σ is the square standard deviation (squaring the variance to obtain the standard deviation). The Gaussian distribution's probability density function is:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) \tag{1}$$

2.3. **RBF neural network.** To address the phenomenon of oversmoothing in the spectral conversion process of Gaussian mixture model, this paper adopts the RBF to train the mean vector μ_i^y so as to improve the oversmoothing spectral conversion. The RBF is a feedforward network proposed by Broomhead and Lowe [24]. Where the role of the input layer is to assign the feature vector to the hidden layer, this layer does not perform any transformation, the hidden layer uses radial basis function and the sublayer transforms the input feature vector into the space of the hidden layer. The output layer calculates the total output weights from the hidden layer. Moreover, the RBF neural network can facilitate both online and offline training approaches, dynamically adjusting the data center, expansion constants, network structure, and hidden layer units, so the RBF model learns very fast, and is more time-efficient than the traditional conversion model. The RBF obtains the conversion function by converting the source speaker's voice acoustic features to the acoustic features of the target speaker's voice. Let us assume that $\hat{y}_j$ represents the output of vector x mapped by RBF neural network, which can be expressed as:

$$\hat{y}_j = \sum_{i=1}^{N} w_{ij}\, \Phi_i(x), \quad 1 \le j \le m \tag{2}$$

where N denotes the number of radial basis functions, w_{ij} denotes the weights of the output layer, m represents the dimension of the output feature vector. $\Phi_i(x)$ denotes the Gaussian function, which can be expressed as follows:

$$\Phi_i(x) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma_i^2}\right), \quad 1 \le i \le N \tag{3}$$

where c_i and σ_i^2 denote the center and width of the hidden layer of the RBF network, respectively.

The mean vector in GMM reflects the mean value of acoustic characteristics on each Gaussian mixture component. Using RBF network to train these mean vectors is equivalent to integrating the nonlinear mapping process into the parameter estimation of GMM, which can more accurately correspond to the nonlinear transformation from source features to target features, thus avoiding the over-smoothing problem caused by only using GMM.

3. English voice conversion method based on GMM and deep learning.

3.1. **English voice conversion method based on GMM modeling.** The English voice conversion process using Gaussian mixture model is shown in Figure 2, from which we can see that in the voice conversion experiments, we first need to carry out feature extraction on the source and target voice, and then carry out dynamic time regularization on the feature parameters to get the mixed feature parameter alignment results, and then send the mixed feature parameter alignment results into the GMM model for training, and the output results use the Expectation Maximization (EM) algorithm to obtain the

accurate values. Finally, the parameter model is obtained. Next, this paper will introduce the above voice conversion process in detail.

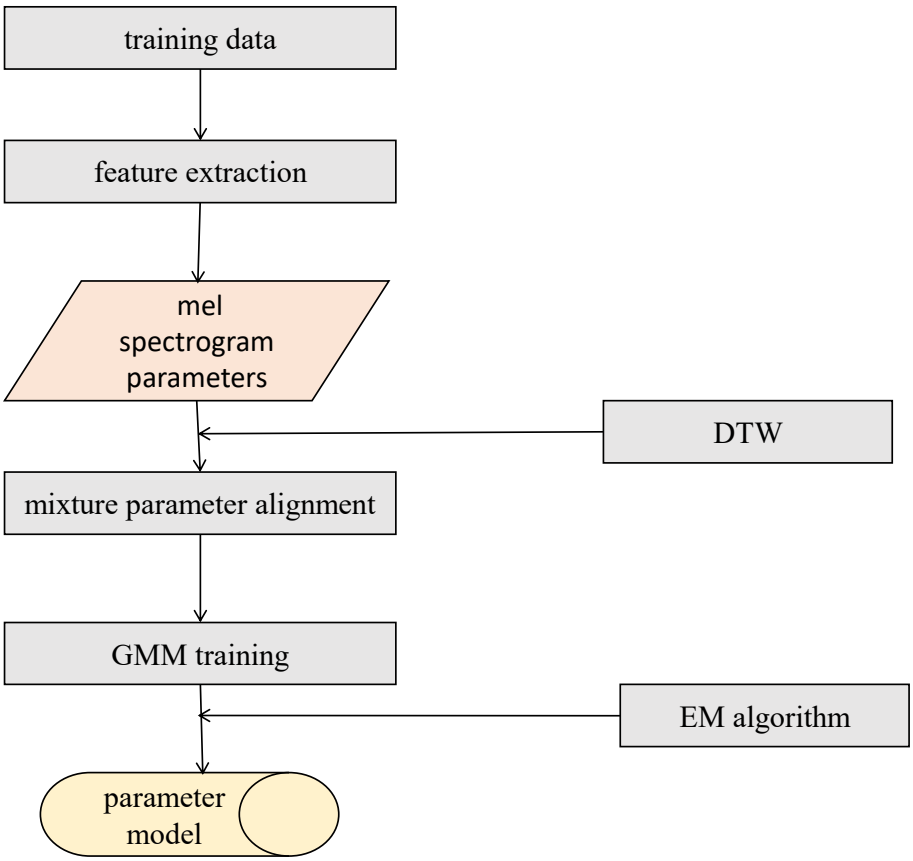


Figure 2. English voice transcription process based on Gaussian mixture modeling

3.1.1. *Dynamic time adjustment.* In this paper, the joint probability density method is utilized for training the GMM model. This is due to the fact that the joint probability density method takes into account both the source and target speakers' feature parameters. This is better in terms of classification than a single probability density function that only extracts the target speaker's feature parameters. Before the conversion of the corresponding model, the first step is to carry out dynamic time regularization on the feature parameters of the source and target speakers, where the dynamic time regularization algorithm is used in this paper.

Dynamic Time Warping (DTW) is an algorithm that calculates the similarity between two time series, especially for those with varying lengths and rhythms. DTW automatically warps the time series (i.e., locally scales them on the time axis), so that the two sequences are as consistent as possible in the time domain, thus obtaining the maximum possible similarity. Nowadays, DTW is extensively utilized in pattern matching, information retrieval, and so on. In the field of voice conversion, DTW can be used to align two sequences of different lengths.

The method of dynamic time regularization is as follows:

(1) For two data sequences which need to be matched $A = [A_1, A_2, \ldots, A_n]$ and $B = [B_1, B_2, \ldots, B_m]$, we form the set of warping indices $W = [W_1, W_2, \ldots, W_x]$, where

$\max(|A|,|B|) \le x \le |A|+|B|$. Each W_t is a pair (i,j), where i denotes the i coordinate in A and j denotes the j coordinate in B.

(2) Starting from $(i,j) = W(a_i, b_j)$, the next node must be chosen from $W(a_{i+1}, b_j)$, $W(a_i, b_{j+1})$, or $W(a_{i+1}, b_{j+1})$, such that the cumulative distance to (i,j) is minimized. This is computed via dynamic programming by considering transitions from $(i-1,j)$, $(i,j-1)$, and $(i-1,j-1)$.

(3) Based on the above two steps, one obtains the shortest distance between the two sequences and the warping path that realizes this distance.

3.1.2. *Transformation of GMM.* Let the dynamic time-adjusted source speaker feature parameter sequence be $x_1, x_2, \ldots, x_n$ and the target speaker feature parameter sequence be $y_1, y_2, \ldots, y_n$, where n represents the number of feature-parameter vectors. Combine these two L-dimensional vectors x_n and y_n into a new expanded vector $Z_n : Zn = [x_n^T y_n^T]^T$ and then mathematically model the joint probability density of x_n and y_n, thus obtaining the mathematical model of the probability density of the parameter space $\{Z_n\}$ of the feature parameter Z_n. The specific probability density function is as follows:

$$p(z_n \mid \lambda^{(e)}) \;=\; \sum_{m=1}^{M} \alpha_m \mathcal{N}\big(z_n;\, \mu_m^{(e)},\, \Sigma_m^{(e)}\big) \tag{4}$$

where $\lambda^{(e)}$ is the parameter set $\{\alpha_m, \mu_m^{(e)}\}$ of the Gaussian mixture model, $\mu_m^{(e)}$ denotes the mean vector of the m-th Gaussian component, and α_m the mixture weight. We can derive the mean vector and covariance matrix of Z_n through decomposition:

$$\mu_m^{(e)} \;=\; \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix} \tag{5}$$

$$\Sigma_m^{(e)} \;=\; \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix} \tag{6}$$

where $\mu_m^{(x)}$ and $\mu_m^{(y)}$ are the mean vectors of x_n and y_n, respectively; $\Sigma_m^{(xx)}$ and $\Sigma_m^{(yy)}$ are the self-covariance matrices of x_n and y_n; and $\Sigma_m^{(xy)}$ and $\Sigma_m^{(yx)}$ are the mutual covariance matrices between x_n and y_n.

The transformation function is:

$$\begin{aligned} \hat{y}_n &= E[y_n|x_n] \\ &= \textstyle\sum_{m=1}^{M} p(m|x_n\ , \lambda^{(z)}) \cdot [\mu_m^{(y)} + \sum\nolimits_m^{(yx)} (\sum\nolimits_m^{(xx)})(x_n - \mu_m^{(x)})] \\ &= \ = \sum\nolimits_{m=1}^{M} \beta_m [\mu_m^{(y)} + \sum\nolimits_m^{(xx)} (\sum\nolimits_m^{(xx)})^{-1}(x_n - \mu_m^{(x)})] \end{aligned} \tag{7}$$

where $\mathbb{E}[\cdot]$ is the mathematical expectation and β_m denotes the posterior probability of the m-th Gaussian component for x_n.

3.1.3. *Problems with the GMM model.* The transformation function of a Gaussian mixture model can be morphed as follows:

$$F(x) = \sum_{q=1}^{Q} P_q(x)\, \mu_q^y \;+\; \sum_{q=1}^{Q} P_q(x) \big(\Sigma_q^{xy}\, \Sigma_q^{xx}\big)^{-1} (x - \mu_q^x) \tag{8}$$

According to the superposition nature of the function, the above equation can be written as:

$$F(x) = A(x) + B(x) \tag{9}$$

$$A(x) = \sum_{q=1}^{Q} P_q(x)\,\mu_q^y \tag{10}$$

$$B(x) = \sum_{q=1}^{Q} P_q(x)\left(\Sigma_q^{xy}\,\Sigma_q^{xx}\right)^{-1}(x - \mu_q^x) \tag{11}$$

where $A(x)$ is the mean term and $B(x)$ is the correlation term, from which it can be seen that the discretization of the eigenvalues of the conversion function is too small, leading to the phenomenon that the conversion of the characteristic parameters is too smooth.

3.2. **English voice conversion method based on GMM model and RBF neural network.** The RBF neural network possesses robust input and output mapping functions, with theoretical affirmation of its effectiveness in accomplishing mapping tasks within the forward network, so this paper adopts RBF neural network to improve the features of voice conversion. Firstly, the feature vectors needed for voice conversion are converted by mapping network using RBF neural network, and then the converted vectors are re-converted by GMM network to finally get the converted parameters. One point to note is that the RBF neural network needs to determine the corresponding mapping function when converting the attributes of the source and target voice, and the mapping of the RBF neural network needs to be obtained from the relationship between the extracted datasets during the training phase. The specific flowchart of the hybrid voice conversion of the RBF+GMM is shown in Figure 3.

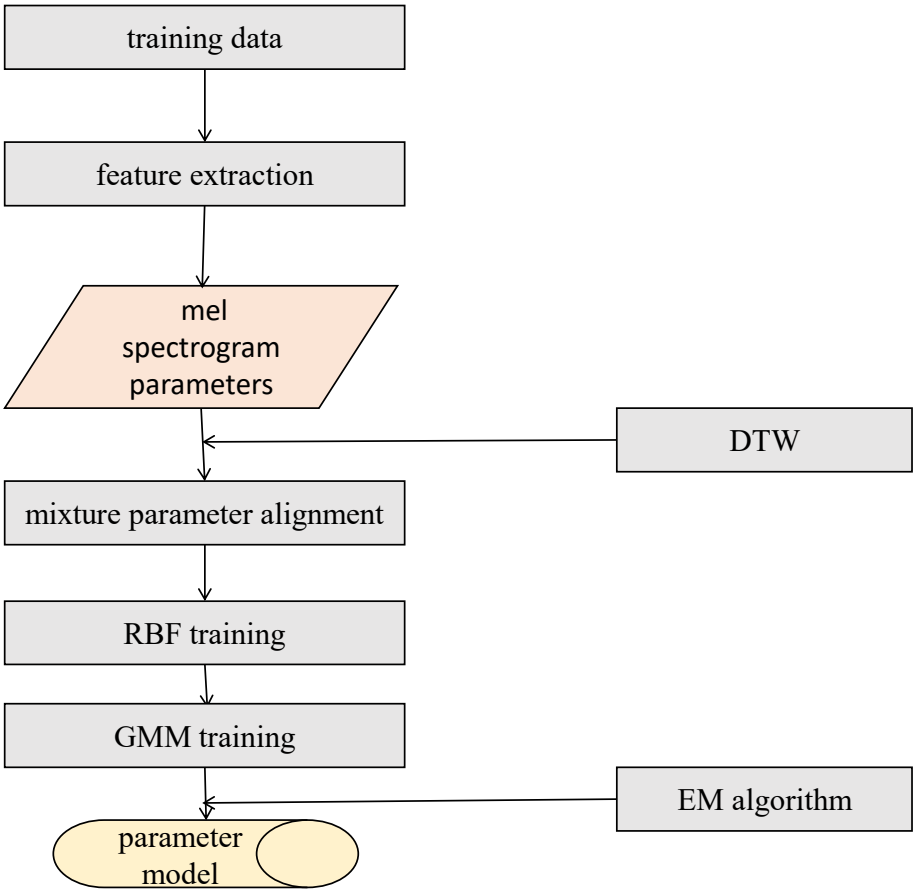


Figure 3. RBF+GMM hybrid voice conversion specific flowchart

When using the RBF–GMM model to convert voice, it is also necessary to extract the spectral features of the source and target voice through the feature extraction stage. In this stage, the experiment also adopts the MFCC as the feature parameter of the conversion, and then adopts the DTW method to align the features of the acquired Mel inverted spectral coefficients, and then adopts the RBF neural network to convert the feature parameters. After the RBF neural network conversion, the improved voice feature parameters can be obtained, and finally the feature parameters are imported into the trained GMM model for conversion. The training phase and conversion phase of the RBF–GMM model–based voice conversion technique is similar to the GMM model–based voice conversion model, and the specific processes are as follows:

Training phase:

(1) According to the set of joint source and target vectors Z mentioned in the previous section, we use the EM algorithm to determine the GMM parameter sequence $(\alpha_i, \mu_i, \Sigma_i)$. Thus, the GMM mapping function is determined.

(2) Extract the dataset required by the RBF neural network, and the extraction formula is:

$$x_{\text{new}} = \mu_i^x \tag{12}$$

$$y_{\text{new}} = \mu_i^y \tag{13}$$

(3) Calculate the mapping function F_{RBF} of the RBF neural network based on the input x_{new} and the output y_{new}.

With the above three steps we can determine the required mapping function F_{RBF} for the RBF neural network.

Conversion phase:

(1) Based on the test vector X', the EM algorithm is used to estimate the GMM parameter sequence μ_i.

(2) Use the F_{RBF} function obtained in the training phase to convert to the new mean vector μ_i^{new}, the conversion formula is:

$$\mu_i^{\text{new}} = F(\mu_i) \tag{14}$$

(3) Replace the new mean vector μ_i^{new} obtained in (2) with the original mean vector μ_i.

(4) Calculate the converted vector features of the new mean vector μ_i^{new} obtained from the computation using the GMM mapping function.

4. **Experiment.**

4.1. **Experimental data set.** In this paper, the recorded English voice data are split into two parts: the training set and the test set, 20 data from the total corpus are randomly selected as the test set, the same test set is used for all experiments, and the remaining 7936 utterances from each speaker are used as the training set. A total of 31744 utterances without repetitions were collected from all training sets, and the total voice duration was 42 hours.

To evaluate the performance of the GMM-based voice transition model and the RBF–GMM-based voice transition model, this paper conducts several sets of comparison experiments, firstly, the two models are separately verified for the effect of voice transition between the same gender, and this paper sets up 2 sets of experiments for comparison, which are: the voice transition experiment (MTM) from male voice (mon_M1) to male voice (mon_M2), and the voice transition experiment (FTF) from female voice (mon_F1) to female voice (mon_F2). Then the conversion effect between opposites is verified for the two models, and this paper also sets up 2 sets of experiments for comparison, which are: the voice conversion experiment (MTF) from male voice (mon_M1) to female voice (mon_F1), and the

voice conversion experiment (FTM) from female voice (mon_F1) to male voice (mon_M2). The experimental data of the two conversion models are shown in Table 1.

Table 1. Experimental data

Experiment type	Source speaker utterance	Target speaker utterance
MTM	3836	2744
FTF	3406	2931
MTF	3913	2512
FTM	2968	2037

By observing the training error during the experimental training process, the training error has converged to a better level after 200 000 training iterations using the GMM model and the RBF–GMM model.

4.2. **Evaluation criteria.** The experiment evaluates the voice conversion effect through subjective and objective tests, the subjective evaluation includes the average opinion score test as well as the similarity evaluation, and the objective evaluation refers to the time domain waveform graph.

(1) Sound quality evaluation.

MOS (Mean Opinion Score) is the subjective opinion score of the voice to be evaluated by human beings, and it is the most commonly used subjective evaluation standard in the field of voice conversion, and the MOS value can be evaluated by different aspects of the voice to be evaluated to obtain the score. Generally speaking, the clarity, naturalness and fluency of the voice to be evaluated are evaluated. Since MOS is highly subjective, the MOS in different experiments are not comparable, and the scoring range of MOS is from 1 to 5, with larger values indicating better voice quality and smaller values indicating worse voice quality. The specific scoring criteria are shown in Table 2. During the evaluation of the MOS, it is necessary to ensure that each evaluator hears the same voice in the same environment, and that all evaluators follow the same evaluation criteria. The MOS scores of all the evaluators are summarized, and the average value is the final score of MOS, which is calculated by the formula:

$$MOS = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} score_{i,j} \tag{15}$$

where M represents the total number of evaluators, N signifies the quantity of voice sentences assessed, and $score_{i,j}$ denotes the rating given by the i-th evaluator for the j-th voice.

Table 2. MOS scoring criteria

Score	Quality level	Degree of distortion
5	excellent	no obvious distortion
4	good	slight distortion
3	fair	obvious distortion
2	poor	high distortion
1	very poor	severe distortion

(2) Similarity evaluation.

Similarity assessment is a test used to assess the level of resemblance between the converted voice and the target voice, which is generally denoted by ABX. In the ABX test,

X, A and B all have the same voice content. The experimental tester is asked to listen to the experimentally converted sample X first, after which the experimental tester is allowed to listen to a random sample of either A or B. The experimental tester is then asked to listen to a random sample of either A or B. It was up to the tester to decide whether the converted sample X was more like the A sample or more like the B sample. Where X is the converted target voice sample, A represents the voice sample of the source speaker, while B corresponds to the voice sample of the target speaker.

(3) Time domain waveform.

The time domain waveform illustrates the fluctuation of the voice signal over time in the form of waveforms, in the voice conversion experiment can be seen through the observation of the time domain waveform of the voice signal to see some important characteristics of the voice signal, such as the time domain waveform will show the starting position of each tone, the amplitude of the waveform can be recognized by watching the amplitude and periodicity of the difference between the different factors, but also through the observation of the waveform whether the waveform is more gentle The waveform can also be used to determine the sound quality of the experimental voice and the intensity of the noise.

4.3. **Analysis of experimental results.** (1) Average opinion rating test.

The average opinion rating is firstly used as subjective evaluation, there are 4 groups of experiments for 2 source speakers to 2 target speakers respectively, each group of experiments has 10 sentences of converted voice, and 4 sentences are randomly selected as the evaluated voice. The results are shown in Table 3:

Table 3. voice conversion experiment MOS value

Model	MTF	FTM	MTM	FTF
GMM	2.81	2.32	3.32	3.48
RBF-GMM	3.16	2.68	3.61	3.89

According to Table 3, the MOS values obtained by the RBF-GMM model can be observed for converted voice are higher than those obtained by the GMM model for converted voice. The experimental results prove that the addition of RBF neural network and the improvement of the mean term in the conversion process can improve the quality of the converted voice to a certain extent. According to the above group experiments, it can also be observed the Mean Opinion Score value of the converted voice from the same gender is typically higher than that of the voice from the opposite gender, which is due to the difference in the spectral characteristics of male and female voices, and the converted voice of the same sex is closer to the target voice due to the small difference in spectral characteristics, while the converted voice of the opposite sex is difficult to reach the similarity with the target voice due to the large difference in spectral characteristics.

(2) Similarity Measurement.

A similarity assessment is used to measure experimental result, in this assessment, A and B represent the voice of the source speaker and the target speaker respectively, and X represents the converted voice. The similarity scale is obtained by calculating the corresponding percentages of the test results, as shown in Table 4:

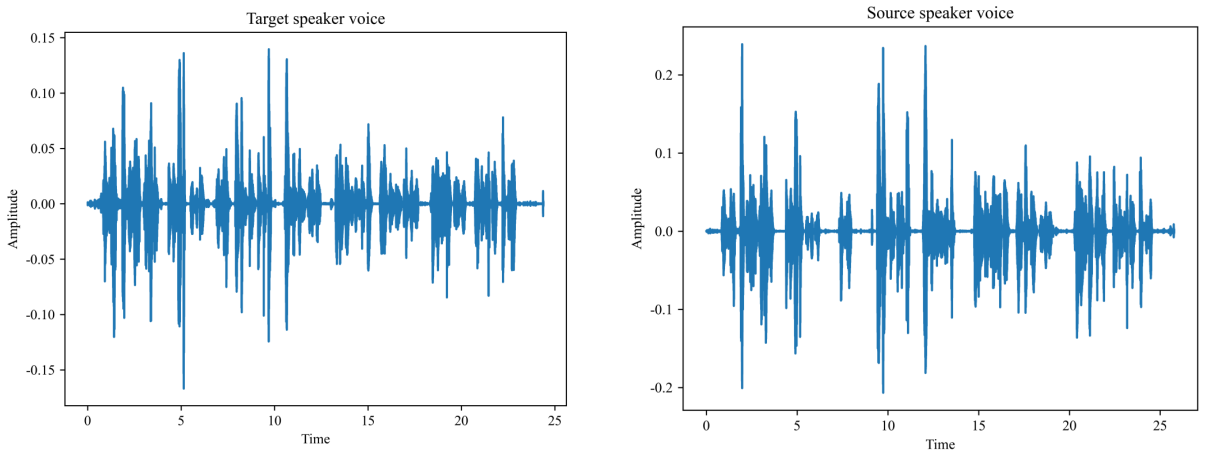
Table 4. voice conversion ABX evaluation results

Model	MTF	FTM	MTM	FTF
GMM	67%	68%	90%	93%
RBF-GMM	69%	71%	92%	95%

From the above table, it can be seen that the RBF-GMM model has a higher percentage of similarity than the GMM model in all the four experiments, representing that the converted voice is more similar to the target voice.

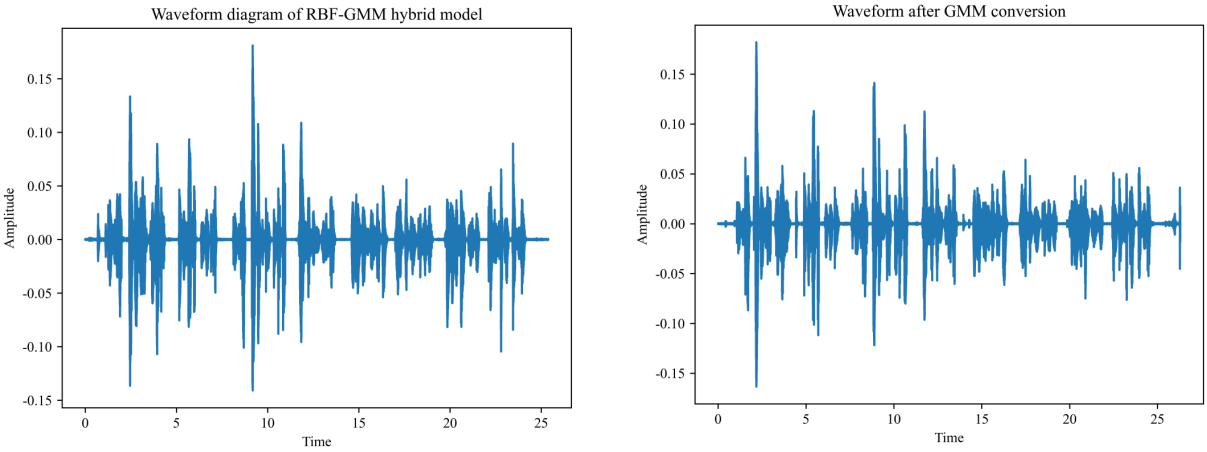
(3) Time domain waveform.

From the time-domain waveforms, the starting position of each tone can be obtained, and the differences of different phonemes can be observed through the waveform amplitude and periodicity. The time-domain waveforms of the source voice and the target voice are shown in Figure 4 and Figure 5 for the traditional GMM-based and RBF+GMM-based models.



(a) (b)

Figure 4. Original waveform



(a) (b)

Figure 5. Waveform after same-sex conversion

From the above figures, it can be concluded that the time-domain waveforms of the voice after the RBF+GMM hybrid voice conversion model in this paper are more closely resembling the target speaker's voice than the traditional GMM model, and the distribution of the voice energy in the time-domain period is also closer to the target voice.

5. **Conclusion.** This paper proposes an English voice conversion technique based on RBF-GMM, that is, the English voice of the source speaker is converted to the English

voice of the target speaker. This paper finds that the conversion is too smooth after studying the traditional voice conversion based on the GMM model; in order to deal with this situation, this paper adds the RBF in the process of the conversion, and then goes through the GMM conversion, and then obtains the final voice. After comparative experiments, it is found that the voice converted by RBF-GMM based English voice conversion technology is better than the voice converted by traditional GMM based English voice conversion technology in terms of clarity, accuracy and similarity. The research content of this paper will contribute to the field of English voice conversion.

REFERENCES

[1] N. Swetha and K. Anuradha, "Text to speech conversion," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 2, no. 6, pp. 269–278, 2013.

[2] T. K. Patra, B. Patra, and P. Mohapatra, "Text to speech conversion with phonematic concatenation," *International Journal of Electronics Communication and Computer Technology*, vol. 2, no. 5, pp. 223–226, 2012.

[3] N. Bi and Y. Qi, "Application of speech conversion to alaryngeal speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 2, pp. 97–105, 1997.

[4] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[5] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A spectral convolutional neural network model based on adaptive Fick's law for hyperspectral image classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19–46, 2024.

[6] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121–2133, 2022.

[7] F. Zhang, T.-Y. Wu, Y. Wang, R. Xiong, G. Ding, P. Mei, and L. Liu, "Application of quantum genetic optimization of LVQ neural network in smart city traffic network prediction," *IEEE Access*, vol. 8, pp. 104555–104564, 2020.

[8] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan*, vol. 11, no. 2, pp. 71–76, 1990.

[9] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[10] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2011.

[11] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[12] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556–566, 2012.

[13] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[14] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[15] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2009.

[16] D. Erro, A. Alonso, L. Serrano, E. Navas, and I. Hernaez, "Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations," *Computer Speech & Language*, vol. 30, no. 1, pp. 3–15, 2015.

[17] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[18] J. Nirmal, M. Zaveri, S. Patnaik, and P. Kachare, "Voice conversion using general regression neural network," *Applied Soft Computing*, vol. 24, pp. 1–12, 2014.
[19] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.
[20] K. Tokuda, T. Kobayashi, and S. Imai, "Adaptive cepstral analysis of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 6, pp. 481–489, 1995.
[21] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.
[22] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation," *Acoustical Science and Technology*, vol. 28, no. 3, pp. 140–146, 2007.
[23] H. Mizuno and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt," *Speech Communication*, vol. 16, no. 2, pp. 153–164, 1995.
[24] N. A. Fox, "If it's not left, it's right: Electroencephalograph asymmetry and the development of emotion," *American Psychologist*, vol. 46, no. 8, p. 863, 1991.