Design and Implementation of Intelligent Singer Recognition System

Can Ge¹, Juanjuan Li¹*, Lina Liu¹, Gukai Li¹, Guoxing Yang¹, Hui Tang¹, Hanlin Yang¹, Yifan Xia¹, Rui Bao¹ and Haiyu Zhang²

¹School of Electronic and Information Engineering, Soochow University, Suzhou, China
²WenZheng College of Soochow University, Suzhou, China

*Corresponding author

Abstract. An intelligent singer recognition system was designed to identify the singer. The scheme established a song library at first, then used MATLAB to extract Mel Frequency Cepstral Coefficients (MFCC) from each song in the song library, moreover, set up characteristic parameters pattern base and trained the pattern base by Vector Quantization (VQ) to obtain the final codebook base. Finally, it can correctly classify the singer based on Dynamic Time Warping (DTW) matching reference characteristic parameters pattern with test pattern. Test results showed that the system's recognition rate is up to 90%.

1. Introduction
With the rapid development of computer and network technology, the application of speech technology is everywhere in people's life, such as speech recognition, speech communication, text conversion and so on. In recent years, speech recognition has made rapid development, and has been very popular in different fields. According to different functions, speech recognition has developed many different branches, where the search and recognition of the music is also an important branch. The intelligent singer recognition system uses speech recognition technology, which has a wide range of applications in music recognition, classification and search. In real life, when people hear a familiar song, if they want to know the singer of the song, they need to recall according to people's memory and relevant experience, then they can find the corresponding singer. This paper designs an intelligent singer recognition system, which can recognize the singer information of songs quickly and accurately. Compared with the traditional singer recognition through people's own feelings, the intelligent singer recognition system has the advantages of wide recognition range, fast recognition speed and high recognition accuracy.¹

2. System Design Scheme
First of all, select 50 singers, record a different song respectively. Then, the recording file should be saved in the training folder, and a song library should be established. Secondly, the training song library used Mel Frequency Cepstral Coefficients (MFCC) parameters to extract the features of the recording files, established the characteristic parameter pattern base, and then used Vector Quantization (VQ) to train the pattern base to get the final codebook base. Finally, users opened any song in the test folder or any multiple songs at the same time. The system extracted the MFCC characteristics of the song at first, and then used Dynamic Time Warping (DTW) algorithm to match the MFCC characteristic parameters of the one song or multiple songs with the MFCC characteristic

¹Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
parameters in the codebook base, and finally obtained the recognition results.\cite{2}\cite{3} As shown in Figure 1.

![Figure 1](image1.png)

**Figure 1.** General system formula.

### 3. System Design and Implementation

#### 3.1. MFCC Characteristic Parameters Extraction

MFCC is a characteristic parameter that combines the auditory perception characteristics of the human ear with the way of speech generation.\cite{4} This characteristic parameter is not related to the input signal, and it can play a good anti-noise performance, and MFCC characteristic parameter can compensate for convolution channel distortion.\cite{5} Mel scale describes the nonlinear characteristics of human ear frequencies, and its relationship to frequency can be expressed by the following formula:

$$\text{Mel}(f) = 2595 \times \lg(1 + f/700)$$  \hspace{1cm} (1)

The MFCC parameter extraction process is as shown in Figure 2.

![Figure 2](image2.png)

**Figure 2.** MFCC parameter extraction flow chart.

#### 3.1.1. Pre-emphasis

Pre-emphasis is a step of pre-processing the speech signal before feature extraction. Normally, the speech is passed through a high-pass filter, and its transfer function is as follows.

$$H(z) = 1 - \alpha z^{-1}, \quad 0.9 < \alpha < 1.0$$ \hspace{1cm} (2)

Assume that the speech sample value for n-time is $x(n)$, after pre-emphasis processing, the result is $y(n) = x(n) - \alpha x(n-1)$, taking $\alpha = 0.98$ \cite{6}.

The purpose of pre-emphasis is to aggravate the high-frequency parts of the speech, thereby eliminating the effects of vocal cord and lips during the speaking. Pre-emphasis can also compensate and enhance the high-frequency parts of the speech signal controlled by the pronunciation system, highlight the formants of the high-frequency. \cite{7}

#### 3.1.2. Framing

Speech signal is a kind of signal with unstable state. But in a relatively short period of time, it has the characteristics of short-term stability. For the convenience of analyzing the speech signal that has been read, the speech signal is often processed by framing.

#### 3.1.3. Windowing

During the process of framing, the spectrum leakage of speech signal is serious, it has been proved by a large number of experiments that the spectrum leakage can be weakened by adding windows. Windowing is to use a finite length window sequence $\omega(n)$ to intercept the speech signal into a small segment. Then analyze the speech sequence, and keep on analyzing the remaining by shifting right. The windowing expression of speech signal is as follows:

$$y(n) = \sum_{n=0}^{N-1} x(n)\omega(n)$$ \hspace{1cm} (3)
N represents the window length, \( x(n) \) is the input signal.

### 3.1.4. Fast fourier transformation.
After adding window, the spectrum of each frame is obtained by Fast Fourier Transformation (FFT) for signal of each frame. The power spectrum of speech signal is obtained by modularizing and taking the square of the spectrum of speech signal. The DFT of speech signal is as follows:

\[
X_n(k) = \sum_{n=0}^{N-1} x(n)e^{-\frac{j2\pi kn}{N}}, 0 \leq k \leq N
\]  

### 3.1.5. Mel filter.
Set a number of band-pass filters \( H_m(k) (0 \leq m < M) \) within the scope of spectrum of speech, and M is the number of filters. Each filter has the characteristic of triangular filtering, its mid-frequency is \( f(m) \). In the scope of Mel frequency, these filters are of equal band-width. The transfer function of each band-pass filter is as follows:

\[
H_m(k) = \begin{cases} 
0, & k < f(m-1) \\
2[k - f(m-1)][f(m+1) - f(m-1)]^{-1}[f(m) - f(m-1)]^{-1}, & f(m-1) \leq k \leq f(m) \\
2[f(m+1) - k][f(m+1) - f(m-1)]^{-1}[f(m) - f(m-1)]^{-1}, & f(m) \leq k \leq f(m+1) \\
0, & k \geq f(m+1) 
\end{cases}
\]  

### 3.1.6. Operation of logarithms.
The output of logarithmic energy of each filter bank is as follows:

\[
s(m) = \ln\left(\sum_{k=0}^{N-1} |X_n(k)|^2H_m(k)\right), 0 \leq m \leq M
\]  

### 3.1.7. DCT(Discrete cosine transform).
MFCC coefficients were obtained by DCT:

\[
C(n) = \sum_{m=0}^{N-1} s(m)\cos\left[\pi n\left(m - 0.5\right)M^{-1}\right], n = 1, 2, ..., L
\]  

The Mel Frequency Coefficients of order L (Order L refers to the order of MFCC coefficients, usually 12-16, and M is the number of triangular filters) are obtained by substituting the above logarithmic energy into DCT. The Mel Frequency Coefficient is the characteristic of the frame of speech of speech.

### 3.2. DTW (Dynamic Time Warping) Algorithm
Based on the idea of dynamic time warping, DTW algorithm can measure the similarity of two time series with inconsistent length, and calculate the similarity between two time series by extending and shortening the time series. If Q series, \( Q = \{q_1, q_2, \cdots, q_n\} \), whose length is n and C series, \( C = \{c_1, c_2, \cdots, c_m\} \), whose length is m, are given, the similarity between Q and C series can be calculated with one series used as reference model and the other as test model. First, construct a matrix \( D \) of n by m whose element of matrix is \( d_{i,j} = \text{dist}(q_i, c_j) \) to align the two series. The element of matrix \( d_{i,j} \) is the similarity between each point of Q series and C series. The smaller the distance is, the higher the similarity is. \( \text{dist} \) represents the function of calculation of distance. It adopts Euclidean Distance. Then, search for the shortest path from \( d_{1,1} \) to \( d_{n,m} \) in matrix D. The shortest path from \( d_{1,1} \) to \( d_{n,m} \) in matrix D is the similarity between the time series Q and C.

### 3.3. Vector Quantization (VQ)
VQ is a process of coding points in a vector space with a finite subset. The analog signal is continuous, but the computer can only process discrete digital signals. When the analog signal is converted into digital signal, replace an interval with a value of the interval. For example, replace all values of \([0, 1)\) with 0, replace all values of \([1, 2)\) with 1, and so on. This is a VQ process. VQ codebook designed based on LBG algorithm is used in the intelligent singer recognition system. \( X_k \) is the training
sequence and B is the codebook. The specific implementation process of VQ in the intelligent singer recognition system is as follows:

3.3.1. The centroid of feature vector of all extracted frames is taken as the first keyword $B_1$.

3.3.2. Split the current codebook $B_m$ according to the formula (8) to form $2m$ codewords.

$$
\begin{align*}
B^ +_m &= B_m (1 + \varepsilon) \\
B^-_m &= B_m (1 - \varepsilon)
\end{align*}
$$

Among them, the value of $m$ is changed from 1 to the quantity of codewords of the current codebook. In this system, the value of $\varepsilon$ is 0.01.

3.3.3. Classify all the training sequence according to existing codebook. Calculate the sum of training vector quantization distortion $D^{[n]}$ according to the formula (9). Calculate the relative distortion according to the formula (10). If the relative distortion is less than a certain threshold $\varepsilon$, the iteration ends and go to step 3.3.5. Otherwise, go to the next step.

The sum of distortion:

$$
D^{[n]} = \sum_{k=1}^{K} \min d (X_k, B)
$$

The relative distortion:

$$
\left| D^{(n-1)} - D^n \right| + \varepsilon
$$

3.3.4. Recalculate the new centroid of each region to get the new codebook and go to step 3.3.3. to transform.

3.3.5. Cycle step 3.3.2., 3.3.3. and 3.3.4. until the formed codebook reaches the number of all required codewords.\(^{[15]}\)

3.4. System Design

The design of the system mainly includes the training part and the recognition part. At first, 50 different songs were recorded by 50 singers, and were saved in the training folder. Training part was that to read each song in the speech corpus at first, and perform speech preprocessing, including pre-emphasis, windowing and frame, and then perform FFT on the achieved signal to obtain the processed spectrum. Pre-emphasis was to emphasize the high-frequency part of the speech, remove the influence of lip radiation, and increase the high-frequency resolution of the speech. Because of the short-term stability of the speech signal, Windowing and framing can divide the speech signal into short frame for processing. Then calculate the energy spectrum and short-term zero-crossing rate, perform the endpoint detection, and square the amplitude of the spectrum.\(^{[16]}\) The energy spectrum was passed through a triangular filter banks and was band-pass filtered by using a triangular filter banks. DCT was performed on the signal which was output by the band-pass filter to obtain the MFCC characteristic parameters. As a result, a characteristic parameter pattern base was established. Next step, the pattern base was trained by VQ to get the final codebook base. Codebook was formed through the MFCC extraction and vector quantization of the voice message. The intelligent singer recognition system uses a vector quantization algorithm based on LBG to design a VQ codebook. The number of codebooks depends on the number of singers. As shown in Figure 3.
Figure 3. System training flow chart.
In the recognition part, the system used VQ to calculate the average distortion measure. After the users opening one song or several songs in the test folder, the system pre-processed the song, calculated short-term energy and short-term zero-crossing rate and detected endpoints. Afterwards, the system extracted the MFCC characteristic parameters, and then used the DTW algorithm to match the MFCC characteristic parameters of the song with those of the codebook base. Meanwhile, the minimum distance was Calculated. When calculating the distance, the system used Euclidean Distance to calculate the lowest difference of the distance between characteristic parameters. Finally, the singer recognition result was obtained. As shown in Figure 4 and Figure 5.

Figure 4. Single song recognition flow chart. Figure 5. Multiple song recognition flow chart.

4. Test and Result Analysis
Click the button 'select training folder' and the select folder dialog box pops up. Select the folder where the training songs are located and the selected path is displayed in the system text box. Then click the 'train' button to train files for analysis and Figure 7 displays the generated codebook files. In this design, a total of 50 songs sung by 50 people were trained, and a total of 50 codebooks were generated as the pattern library. As shown in Figure 6 and Figure 7.
The system can read a song for recognition, or read more than one song for recognition.

4.1. Single Song Recognition
Click the 'select recognition folder' button, the select folder dialog box pops up, and select the folder where the songs need to be recognized. The system displays the selected path in the text box, and enters 1 in 'the number of songs' edit box (the system defaults to 1). Then click the 'recognize' button, and the system recognizes the singer. The result is shown in Figure 8.

4.2. Identification of Multiple Songs
For example, this design identified 7 songs. Click the 'select recognition folder' button, and the dialog box pops up. Select the folder where the 7 songs need to be recognized. The system displays the selected path in the text box. Enter 7 in 'the number of songs' edit box (system default is 1). Then click the 'recognize' button, and the system recognizes 7 songs at the same time. As shown in Figure 9.

This system recognized 50 different songs and calculated the recognition rate according to the formula (11).
Recognition rate = \( \frac{\text{The number of correctly recognized songs}}{\text{The total of recognized songs}} \) (11)
The recognition result of single song is as shown in Table 1.
Table 1. Recognition rate of intelligent singer recognition system.

| The total of recognized songs | The number of correctly recognized songs | Recognition rate |
|-----------------------------|------------------------------------------|------------------|
| 50                          | 45                                       | 90%              |

As can be seen from Table 1, the number of recognized songs in this experiment is 50, among which 45 are correctly recognized and 5 are wrong, so the recognition rate is 90%.

It is worth mentioning that the recognition result of a single song remains unchanged in the recognition of multiple songs, so the recognition rate of multiple songs is the same as that of a single song.

The recognition rate is related to the quantity of training samples, the more samples, the higher recognition rate. However, due to the limitations of DTW algorithm, the recognition time is long. Environmental noise interference will also affect the recognition rate, so speech filtering and enhancement technologies need to be further studied.

5. Conclusion
The system applies speech recognition technology to singer recognition, and thus designs an intelligent singer recognition system. The system can recognize one or more song files, but be sure to modify "the number of sons" on the software interface, and the file name in the song recognition folder must start with s1.wav. The system automatically recognizes singers according to the input songs, and the system recognition rate is high from the test results. Because the size of codebook base is limited and the system takes the recognition of songs without background music, which results in the limitation of the system application. The system still needs to be further optimized, such as expanding the codebook base and improving the system recognition rate.

Acknowledgments
A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (No.GD11900118).

References
[1] Arul Valiyavalappil Haridas, Ramalatha Marimuthu, Vaazi Gangadharan Sivakumar (2018). A critical review and analysis on techniques of speech recognition: the road ahead[J]. International Journal of Knowledge-based and Intelligent Engineering Systems, 22(1), 39-57.
[2] Maarten Van Segbroeck, Ruchir Travadi, Shrikanth S. Narayanan(2015). Rapid Language Identification [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing,23(7), 1118-1129.
[3] Jinpu XuYeping ZhuPengmin XuDexin Ma (2018). Agricultural price information acquisition using noise-robust Mandarin auto speech recognition [J]. International Journal of Speech Technology, 21, 681-688.
[4] Jihyuck Jo, Hoyoung Yoo, In-Cheol Park (2016). Energy-efficient floating-point MFCC extraction architecture for speech recognition systems[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 24(2), 754-8.
[5] JianFeng WU, HuiBin QIN, YongZhu HUA, LingYan FAN (2018). Pitch Estimation and Voicing Classification Using Reconstructed Spectrum from MFCC [J]. IEICE TRANS. INF. & SYST., E101–D (2), 556-559.
[6] N.J.Nalini, S.Palanive (2016). Music emotion recognition: The combined evidence of MFCC and residual phase [J]. Egyptian Informatics Journal, 17(1),1-10.
[7] Thomas G. Poder, Jean-François Fisette, Véronique Déry (2018). Speech Recognition for Medical Dictation: Overview in Quebec and Systematic Review [J]. Journal of Medical Systems (2018) 42, 89.
[8] Adam Switonski, Tomasz Krzeszowski, Henryk Josinski, Bogdan Kwolek, Konrad Wojciechowski(2018). Gait recognition on the basis of markerless motion tracking and DTW transform [J]. IET Biometrics,7(5), 415-422.
[9] Ahmed Kamil Hasan Al-Ali, David Dean, Bouchra Senadji, Vinod Chandran, Ganesh R. Naik (2017). Enhanced Forensic Speaker Verification Using a Combination of DWT and MFCC Feature Warping in the Presence of Noise and Reverberation Conditions [J]. IEEE Access, 5(2), 15400-15413.

[10] Manabu Okawa (2019). Template Matching Using Time-Series Averaging and DTW With Dependent Warping for Online Signature Verification [J]. IEEE Access, 7, 81010-81019.

[11] Abhishek Sharma, Suresh Sundaram (2018). On the Exploration of Information From the DTW Cost Matrix for Online Signature Verification [J]. IEEE Transactions on Cybernetics, 48(2), 611-624.

[12] Tao Sun, Hongbo Liu, Hong Yu, C. L. Philip Chen (2017). Degree-Pruning Dynamic Programming Approaches to Central Time Series Minimizing Dynamic Time Warping Distance [J]. IEEE Transactions on Cybernetics, 47(7), 1719-1729.

[13] Yu-Tao Liu, Yong-An Zhang, Ming Zeng (2019). Adaptive Global Time Sequence Averaging Method Using Dynamic Time Warping [J]. IEEE Transactions on Signal Processing, 67(8), 2129-2142.

[14] Jialong He, Li Liu, and G"unther Palm (1999). A Discriminative Training Algorithm for VQ-Based Speaker Identification [J]. IEEE Transactions on Speech and Audio Processing, 7(3), 353 - 356.

[15] Peyman Rahmani; Gholamhossein Dastghaibyfard (2018). Two reversible data hiding schemes for VQ-compressed images based on index coding [J]. IET Image Processing, 12(7), 1195 - 1203.

[16] Danyang Cao, Xue Gao, Lei Gao (2017). An Improved Endpoint Detection Algorithm Based on MFCC Cosine Value [J]. Wireless Pers Commun, 95:2073-2090.