Research on Ethnic Original Ecological Music Clustering of MFCC Based on HHT

Qinjian Fu¹, Danju Lv²*, Yan Zhang³, Haifeng Xu⁴, Wang Yao⁵, Jiali Zi⁶, Jiang Liu⁷

¹ College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, Yunnan, 650233, China
²* College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, Yunnan, 650233, China
³ Institute of mathematical, Southwest Forestry University, Kunming, Yunnan, 650233, China
⁴ College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, Yunnan, 650233, China
⁵ College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, Yunnan, 650233, China
⁶ College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, Yunnan, 650233, China
⁷ College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, Yunnan, 650233, China

*Corresponding author’s e-mail: lvdanjv@swfu.edu.cn

Abstract. In view of the characteristics of ethnic original ecological music and the limitation of human ear’s understanding of hearing. This paper presents a MFCC feature extraction method based on HHT transformation. In this method, the ethnic original ecological music signals were decomposed into several inherent mode functions (IMFs) by EEMD and Hilbert transform, in which the Hilbert marginal spectrum of each IMF was used to complete the feature extraction of HHTMFCC through the MEL scale filter. The experiment collected 13 kinds of original ecological music of ethnic minorities in Yunnan. Based on the feature extraction results of HHTMFCC, Kmeans clustering analysis method was adopted to compare and analyze the clustering effect of HHTMFCC. The experimental results showed that the HHTMFCC feature was 0.49 higher than that of MFCC in the Purity index, 1.2 lower than that of Entropy index, and 3% higher than that of F index. The experimental results show that the HHTMFCC features extracted from ethnic original music are better than the traditional MFCC features.

1. Introduction

MFCCS (Mel Frequency Cepstral Coefficients) was proposed by Davis and Mermelstein in 1980. It is a widely used feature in speech recognition. MFCC can effectively reflect the static characteristics of speech parameters[1]. In order to optimize the speech feature extraction method, Shao Mingqiang et al.
[2] damaged the sound signal while removing the noise in the existing algorithm. A method of speech feature extraction based on MFCC_P is proposed. The filter banks are arranged neatly without overlap, and the noise is effectively eliminated. Li Zhizhong et al. [3] proposed the mean Mel frequency cepstrum coefficient (AMFCC), which was used for classification and recognition by support vector machine, and the results showed that the improved AMFCC had a better recognition rate than LPCC and traditional MFCC. Kazi Fi [4] et al. used high-order spectrum to obtain phase entropy from musical instrument signals, and combining high-order spectral features with MFCC could improve classification accuracy. Cheng Long et[5] reasonable eigenvalues of chirp signal are extracted, this paper proposes a based on empirical mode decomposition (EMD) improved MFCC algorithm, through the study of the empirical mode decomposition of birdsong signal, FFT computation again, after the frequency synthesis by Mel filter, take its logarithm energy DCT transform, improved MFCC parameters, and then by using the gaussian mixture model (GMM) for the identification of birdsong. The experimental results show that the improved MFCC is superior to the traditional MFCC.

Due to the characteristics of folk music itself and the limitations of the human ear in auditory understanding, music is not speech, and it is very different from speech in signal characteristics and structure [6]. Grosal Ariit et al. [7] adopted MFCC as the classification feature, combined with support vector machine and RANSAC and Random Samples Consensus to classify the collected instrumental and vocal music genres, and the experimental results showed that RANSAC had a better effect on data processing with many changes. Elias Pam Palkout et al. [8] developed a music clustering system based on SOM(self organizing), which adopts the type judgment method of template matching and carries out music clustering by calculating the Euler distance between template vector and feature vector. In the process of MFCC, Fourier transform is used to reconstruct the frequency spectrum, so that this method has a certain anti-noise effect, but FFT lacks the function of frequency positioning and cannot highlight the dynamic instantaneity of speech [9]. Chen Shu et al. [10] proposed an improved feature parameter extraction method based on cochlear cepstrum coefficient (MFCC), aiming at the problem that the speech recognition rate of Mayer frequency cepstrum coefficient (MFCC) parameter decreases in the noise environment. The improved linear discriminant analysis (LDA) algorithm is used to transform the extracted characteristic parameters to obtain more discriminative characteristic parameters and diagonalized covariance matrix to meet the requirements of HMM. Experimental results show that the proposed method can effectively improve the recognition rate and robustness of speech recognition system in noisy environment.

The Hilbert-Huang Transformm can handle double non- signal signals very well[11]. Norden E.Huang et al [12] proposed the Hilbert-Huang Transformm (HHT) method to deal with the unsatisfactory double non-signal processing method, and this method has relatively ideal results for double non-data analysis. Speech signals are decomposed by EMD, and the obtained IMFS is then extracted by MFCC, and the recognition effect is improved under the Berlin Emotional Language Database [13]. However, the IMF obtained by EMD decomposition has the phenomenon of mode mixing, resulting in the inaccurate IMF component [14]. Mode aliasing is caused by the discontinuous characteristics of the signal, N.E.Huang et al.proposed Ensemble Empirical Mode Decomposition (EEMD), EEMD method can effectively overcome the modal aliasing existing in EMD [15 16].

Due to the characteristics of ethnic original ecological music and the limitation of human ear's understanding of hearing, this paper proposes the MFCC extraction algorithm based on HHT. Through Kmeans clustering analysis, the experimental results show that HHTMFCC parameters are better than MFCC characteristic parameters.
2. MFCC extraction principle based on HHT

2.1. Principle block diagram of MFCC extraction based on HHT

Figure 1 is the extraction principle block diagram of HHT-MFCC. The dashed box part is the realization process of replacing EMD with EEMD. IMF is obtained after the signal is decomposed by EEMD, and the mixed HHT-MFCC features are extracted through MFCC.

2.2. Cluster evaluation index

In this paper, HHT-MFCC of 13 categories of Yunnan ethnic original ecological music was evaluated by three clustering indexes, including Purity, Entropy and F-measure($F_\beta$). The clustering performance metric [17] is a kind of evaluation standard for clustering results. The corresponding range of Purity and F-measure indicators is [0 1]. The closer to 1, the better the clustering effect, and the smaller the Entropy index, the better the clustering result.

Table 1. confusion matrix.

| fact | Forecast results | negative example |
|------|------------------|------------------|
|      | positive example | negative example |
| positive example | TP | FN |
| negative example | FP | TN |

Accuracy ratio $P$ and recall ratio $R$ are respectively:

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R} \quad (3)$$

$\beta > 0$ measures the relative importance of recall ratio to precision ratio. $\beta = 1$ degenerates to standard F1; $\beta > 1$ has a greater impact on recall rate; $\beta < 1$ has a greater effect on the accuracy.

The calculation formula of Purity[18]:

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j| \quad (4)$$

$N$ represents the total number of samples; $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ represents cluster after clustering, $C = \{c_1, c_2, \ldots, c_k\}$ is the correct category; $\omega_k$ represents all samples in the KTH cluster after clustering, $c_j$ represents the true sample in category $j$. The value range of $P$ is [0,1], and the larger the value, the better the clustering effect.

The calculation formula of Entropy [19]:

[Further content not shown]
\[
P_{ij} = \frac{m_{ij}}{m_i} \quad (5)
\]
\[
e_i = -\sum_{j=1}^{K} P_{ij} \log_2 P_{ij} \quad (6)
\]
\[
e = \sum_{i=1}^{m} \frac{m_i}{m} e_i \quad (7)
\]

\(P_{ij}\) is the probability that a member of cluster I belongs to class J, \(m_{ij}\) is the number of all the members in the cluster I, \(L\) is the number of classes, \(K\) is the number of clusters, \(M\) is the number of members involved in the whole cluster division.

2.3. HHT Hilbert-Huang Transform (HHT)
Hilbert transform (ht) refers to the complex envelope of a signal modulated by a real carrier. The Hilbert transform is used to describe the envelope of amplitude or phase modulation. The instantaneous frequency and instantaneous phase will make the analysis more simple. It has important theoretical significance and application value in communication system. hht is an adaptive time-frequency analysis method suitable for nonstationary nonlinearity. This method can reflect the dynamic instantaneous of the signal and capture the local change of the signal. The basis function [17] can be generated according to the characteristics of the signal itself. hht compared with ht, one more empirical mode decomposition. EMD is to decompose complex signals into several inherent modal functions from high frequency to low frequency (imf). Imf needs to meet two conditions: The number of polar points of the signal is equal to zeros or the difference is 1; The local mean of the upper envelope defined by the maximum and the lower envelope defined by the minimum is 0(that is, the upper and lower symmetry of the envelope).

EMD is the process of extracting the highest frequency component of the signal in each local in turn and obtaining a series of imf, so that each imf component can be hilbert transformed and the hilbert edge spectrum of each component can be obtained.

2.4. EEMD Ensemble empirical mode decomposition EEMD
The empirical mode decomposition (emd) method is huang proposed, As a new time-frequency analysis method, it is an adaptive time-frequency localization analysis method: imf is related to sampling frequency; It is based on changes in the data itself. emd is better than Fourier transform method, it gets rid of the limitation of Fourier transform. emd more important drawback is modal aliasing, which is eemd proposed by huang in order to better solve this problem[15].

EEMD algorithm steps:
Step1: add normal white noise to the original signal \(X'(t)\).
\[
X'(t) = X(t) + w(t) \quad (8)
\]
Step2: signal with white noise as a whole, then decompose the emd to get imf components.
\[
X'(t) = \sum_{j=1}^{n} c_j(t) + r_n(t) \quad (9)
\]
Step3: repeat steps 1 and 2, add a new normally distributed white noise sequence each time.
\[
X'(t) = \sum_{j=1}^{n} c_{ij}(t) + w_{in}(t) \quad (10)
\]
Step4: each time the imf is integrated and processed as the final result.
\[
c_n(t) = \frac{1}{n} \sum_{i=1}^{N} c_{in}(t) \quad (11)
\]
\[ X(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \]  

Among them, \( X(t) \) is the original signal, \( X'(t) \) is a noisy signal, \( r_n(t) \) is the remainder component of the decomposition, \( c_i(t) \) represents each intrinsic modal function.

EEMD the influence of signal extremum imf, if the distribution is not uniform, modal aliasing will occur. eemdir the essence of the method is a multiple empirical mode decomposition of superimposed Gao Si white noise, using the statistical characteristics of Gao Si white noise with uniform frequency distribution, The extreme point characteristics of the signal are changed by adding different white noise of the same amplitude each time, and then the corresponding imf obtained by multiple emd are averaged to counteract the added white noise, thus effectively suppressing the generation of modal aliasing.

3. Implementation of MFCC algorithm based on HHT

The mfcc feature is a human sound perception model based on inner ear frequency analysis. mfcc set provides a perceptual, smooth estimation of speech spectrum over time. The fft transformation in mfcc feature extraction mfcc feature extraction is replaced by Hilbert yellow transform hht, because the double non-signal can be processed well. eemdir the white noise is introduced into the signal to be analyzed, the spectrum of the white noise is uniformly distributed, and the white noise makes the signal automatically distributed to the appropriate reference scale. Because of the characteristic of zero mean noise, the noise will cancel each other after many average calculations, so the calculation result of integrated mean can be regarded as the final result directly. The difference between the calculated results of the integrated mean and the original signal decreases with the increase of the number of integrated mean. The obtained hilbert spectrum is transformed by mel filter banks, logarithm and dct to obtain hht-mfcc mixed characteristic parameters. The improved mfcc feature parameter extraction process based on eemd is shown in figure.2.

![Figure 2. HHTMFCC algorithm flow chart.](image)

The experimental steps of feature extraction for ethnic original ecological music audio are as follows:
Step1: The pretreatment of each ethnic original music audio file;
Step2: Add normally distributed white noise to the original signal;
Step3: The signals added with white noise were taken as a whole, and then EEMD decomposition was carried out to obtain each IMF component;

Step4: Repeat steps 2 and 3, adding a new normally distributed white noise sequence each time;

Step5: The IMF obtained each time will be integrated and averaged as the final result;

Step6: The marginal spectrum obtained in Step 5 goes through MEL filter bank, logarithmic operation, and finally DCT transformation is performed to obtain HHT-MFCC features.

4. Experiment and the analysis of result

4.1. The introduction of experimental data

Table 2 audio data of ethnic original music.

| data | time (s) |
|------|----------|
| Bai  | 199      |
| Yi   | 99       |
| Bulang | 289    |
| Deang | 186      |
| Pumi | 115      |
| Dai  | 83       |
| Dulong | 100    |
| Jingpo | 89      |
| Wa   | 99       |
| Lisu | 79       |
| Naxi | 135      |
| Hani | 175      |
| Nu   | 111      |

The experimental data in this paper are derived from tutor's project team. A total of 40 WAV files of original ecological music of 13 categories of ethnic minorities in Yunnan, China, were obtained, including Bai, Yi, Bulang, Deang, Pumi, Dai, Dulong, Jingpo, Wa, Lisu, Naxi, Hani and Nu. Table 2 shows the total duration corresponding to each type of ethnic original ecological music.

Table 3 Cluster analysis of experimental data.

| data   | sample | attributes | category |
|--------|--------|------------|----------|
| HHT-MFCC | 90709  | 13         | 13       |
| MFCC   | 90709  | 13         | 13       |

Table 3 shows the experimental data of cluster analysis. The data in the table are the MFCC and HHT-MFCC characteristics of 13 categories of ethnic original music audio.

4.2. Experimental parameter settings

Table 4 Experimental parameter table.

| parameters | Parameters value |
|------------|------------------|
| frameSize  | 256              |
| overlap    | 0                |
| Nstd       | 0                |
The algorithm parameters of the experiment in this paper are shown in Table 4.

4.3. Experimental results
The experiment uses the HHT and MFCC fusion method to extract the features of these 13 categories of ethnic original music audio, and obtain the HHTMFCC features and the traditional MFCC features. Fig. 3 is the HHTMFCC feature diagram of ethnic original music audio, from which it can be seen that the cepstrum coefficients of each order are quite different.

Figure 3. Hhtmfcc corresponding to the four ethnic original music audio.
The cepstrum mfcc in figure 4 is quite different from the cepstrum of the hht-mfcc, but the hht-mfcc has good reduction characteristics for the envelope change of the original signal.

This paper uses Kmeans clustering to compare and analyze the clustering effect of HHT-MFCC and MFCC through three indexes: Purity, Entropy, and F-measure ($F_\beta$). The experimental results are shown in Table 5.

| data      | Purity | Entropy | $F_\beta$ |
|-----------|--------|---------|-----------|
| HHT-MFCC  | 0.695  | 1.272   | 0.169     |
| MFCC      | 0.209  | 2.463   | 0.131     |

The data in Table 5 are three index values obtained by clustering the MFCC and HHT-MFCC characteristics of 13 types of native music. The traditional MFCC and HHT-MFCC feature parameters are clustered by K clustering method, and the two types of features are clustered by evaluating the external clustering performance indicators Purity, Entropy, and F-measure.

From Table 5, it can be seen that the HHT-MFCC feature is better than the single MFCC feature in Purity, Entropy, and F-measure. It can be concluded that the HHTMFCC features are 0.49 higher in the Purity index than in the MFCC, 1.2 lower in the Entropy index, and 3% higher in the F index., where the purity index and the F indicator are better, and the smaller the Entropy indicator. In summary, the improved HHTMFCC features are superior to the traditional MFCC characteristics.
5. Conclusion

Ethnic original ecological music in Yunnan is a treasure of national culture and art, which is of great value to the classification and recognition of ethnic original ecological music. In view of the characteristics of original ecological music of 13 kinds of nationalities in Yunnan, this paper proposes a feature extraction method of MFCC based on HHT transform. K-means clustering is used to cluster the features extracted from 13 categories of ethnic original music audio, and the results show that the features of ethnic original music based on improved HHT are better than those of traditional MFCC. The next step can be classified and recognized according to the HHTMFCC characteristics of ethnic original ecological music.

Acknowledgments

This research was funded by National Natural Science Foundation of China (Grant No. 31860332) and Major scientific and technological projects in Yunnan Province (Grant No. 202002AA10007), Science Research Fund of Yunnan Education Department(2019J0220).

References

[1] Miao Jin.( 2012) Research on Speech Recognition Method of Unspecific Isolated Words Based on Echo State Network . Lanzhou University, TN912.34.
[2] Shao Mingqiang, Xu Zhijing. (2017)Speech Recognition Algorithm Based on Improved MFCC Feature. Microcomputers and Applications, 21:52-54+57.
[3] Li Zhizhong, Teng Guanghui. (2008) Poultry vocal feature extraction method based on improved MFCC. Transactions of the CSEA, 11:202-205.
[4] Kazi F I , Bhalke D G . (2015) Musical instrument classification using higher order spectra and MFCC. In: International Conference on Pervasive Computing, Pune.
[5] Cheng Long, Zhang Huaqing. (2017) Research on Bird Singing Recognition Method Based on Improved MFCC. Journal of Communication University of China (Natural Science Edition), 2017(3):41-46.
[6] Whitman, Brian A. (2005) Learning the meaning of music. In: Massachusetts Institute of Technology. Massachusetts Avenue. pp. 99-104.
[7] Ghosal A, Chakraborty R, Dhara B C, et al. (2011) Song Classification: Classical and Non-classical Discrimination Using MFCC Co-occurrence Based Features. In: Signal Processing, Image Processing & Pattern Recognition-international Conference, Sip, Held As. Jeju Island. pp. 260.
[8] Rauber A. (2002) Using Psycho-Acoustic Models and Self-Organizing Maps to Create a Hierarchical Structuring of Music by Sound Similarity. Proc of Ismir Paris France.
[9] Li Zhizhong, Teng Guanghui. (2008) Poultry vocal feature extraction method based on improved MFCC. Transactions of the Chinese Society of Agricultural Engineering, 11(24): 202-205.
[10] Chen Shu, Yu Haibo. (2018) Application of an improved feature extraction method in speech recognition. Transducer and Microsystem Technologies, 037(005):154-157.
[11] Shen Yi, Shen Zhiyuan. (2010) A Non-Linear and Non-Stationary Adaptive Signal Processing Method-An Overview of Hilbert-Huang Transform: Development and Application. Automation Technology and Application, 5: 6-10.
[12] Semion Kizhner, Blank Karin, Flatley Thomas, et al. (2006) On Certain Theoretical Developments Underlying the Hilbert-Huang Transform. In: Aerospace Conference. Big Sky, MT, USA. pp. 14.
[13] Chu YY , Xiong W H , Chen W . (2014) Speech Emotion Recognition Based on EMD in Noisy Environments. Advanced Materials Research, 831:460-464.
[14] T-U Binbin, Fengqin Y-U. (2012) Speech emotion recognition based on improved MFCC with EMD. Computer Engineering & Applications, 48(18):119-122.
[15] Wu Wenjun. (2019) Non-stationary signal analysis based on HHT and machine learning. University of Electronic Science and Technology of China.
[16] Wu Zhaohua, Huang Norden-E. (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis, 1:1-41.
[17] Zhou, Z.H. (2016) Machine learning. Tsinghua University Press, Beijing.
[18] Xie Deyan. (2019) Multi-view Clustering Based on Graph Learning. Xidian University.
[19] Zhang Long, Song Chengyang, Zou Youjun, Hong Chuang, Wang Chaobing. (2020) Performance degradation assessment of bearings based on renyi entropy and k-medoids clustering. Journal of Vibration and Shock, 39:24-31.