Research on Crane Sound Clustering of MFCC Based on HHT

Qinjian Fu, Danjue Lv, Yan Zhang, Haifeng Xu, Wang Yao, Jiali Zi, Jiang Liu
Southwest Forestry University, Kunming 650000, China
fuqinjian@swfu.edu.cn

Abstract. Due to the uniqueness of the sound mechanism of birds, they have typical non-stationary and nonlinear characteristics. This paper proposed a new acousic feature, HHT-MFCC, combined the HHT transformation and MFCC method, aiming at the dynamic instantaneousness of bird sounds. This method, firstly, uses the ensemble empirical mode decomposition EEMD to decompose the bird sounds into a number of intrinsic modal functions IMFs, and then adopts the Hilbert transform to obtain the Hilbert marginal spectrum of each IMF, at last applies the mel-scale filter to complete the feature extraction of HHT-MFCC. The experiment extracts HHT-MFCC from 9 kinds of cranes in China to cluster. Three indexes of cluster evaluation are used to evaluate the feature HHT-MFCC and MFCC. The results show that the HHT-MFCC feature is 10% higher in RI index than MFCC, 9% higher in JC index, and 4% higher in FMI index.

1. Introduction
MFCCs (Mel Frequency Cepstral Coefficients), proposed by Davis and Mermelstein in 1980, is a feature widely used in automatic speech and speaker recognition. The premise of It was proposed by Davis and Mermelstein in 1980. The traditional MFCC is that the sound is assumed to only reflect the static in a short time characteristics of speech parameters. If this premise is not guaranteed, MFCC features will not perform well[1]. In order to reflect the dynamic characteristics of the human ear to speech, Zhang Zhen et al[2]. introduced the frame energy parameters of the time domain characteristic information to improve the MFCC, and conduct simulation experiments on the speaker’s speech. The result of the improved MFCC method is better than that of the traditional method, but using the improved method, the amount of calculation will be very huge. Li Zhizhong et al[3]. proposed an improved average Mel frequency cepstral coefficient (AMFCC), using support vector machine classification and recognition, the results show that the improved AMFCC recognition rate is better than LPCC and traditional MFCC. Kazi FI et al[4]. Used high-order spectrum to obtain phase entropy from musical instrument signals. Combining high-order spectral features with MFCC can improve classification accuracy. Gao Ming et al[5], according to mfcc, the feature recognition effect of noisy audio extraction is poor, a method based on improved Mel Cepstral Coefficient (MFCC) is proposed, through multi-window spectrum estimation and first-order, second-order First-order difference method improves recognition performance. The experimental results prove that in the case of pure speech and added signal noise, the recognition accuracy of the improved method is increased.

The voice of crane can emit a high-frequency beep due to the particularity of its whistling tube[6]. The characteristics of bird sounds are similar to the morphological characteristics of birds. Both of them have species specificity and are of great significance in the recognition and classification of birds[7]. Wang Enze et al, proposed a dual GMM model recognition method based on MFCC in tweet and singing. The results show that it has a higher recognition rate than a single singing model[8].
reconstruction of the Fast Fourier transform (FFT) in the MFCC process makes the method have a certain anti-noise effect, but FFT lacks the function of positioning the frequency and cannot highlight the dynamic and instantaneous nature of speech [9].

Hilbert-Huang Transform can handle nonlinear signals well [10]. It was proposed by Norden E. Huang [11]. This method has relatively ideal results for non-stationary and nonlinear data analysis [12]. The speech signal is decomposed by EMD, and the obtained IMFs are then subjected to MFCC feature extraction, and the recognition effect is improved under the Berlin emotional speech database [13] However, the IMF obtained by EMD decomposition has modal aliasing, which causes the IMF components to be inaccurate. Aiming at the modal aliasing caused by the discontinuity of the signal, N.E. HUANG et al. proposed the Ensemble Empirical Mode Decomposition, EEMD [14, 15]. The EEMD method can effectively overcome the modal aliasing existing in EMD.

Due to the non-linear non-stationary and dynamic transientness of the Crane tone signal, this paper puts forward a HHT-based MFCC extraction algorithm. Through cluster analysis experiments, the results show that HHT-MFCC 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

![Block diagram of MFCC extraction based on HHT](image)

The extracting feature of HHT-MFCC is shown in Fig. 1. The dashed box part of the figure is the EEMD, avoiding the modal mixing of the IMF components and the Hilbert transformation process, obtaining the instantaneous spectrum.

2.2. Cluster evaluation index

This paper uses the RI, JC, and FMI indicators to evaluate HHT-MFCC features of 9 types of cranes, to reflect the differences of the 9 types of crane sounds. The clustering performance measurement index [16] is an evaluation standard for the clustering results. The corresponding range of the three indicators is between [0 1]. The closer to 1, the better the clustering effect. The calculation formula is as follows:

\[ a = \frac{SS}{SS} = \begin{cases} \lambda_j = \lambda_j, & \lambda_i^* = \lambda_j^*, & i < j \end{cases} \] (1)

\[ b = \frac{SD}{SD} = \begin{cases} \lambda_j = \lambda_j, & \lambda_i^* \neq \lambda_j^*, & i < j \end{cases} \] (2)

\[ c = \frac{DS}{DS} = \begin{cases} \lambda_i \neq \lambda_j, & \lambda_i^* = \lambda_j^*, & i < j \end{cases} \] (3)

\[ d = \frac{DD}{DD} = \begin{cases} \lambda_i \neq \lambda_j, & \lambda_i^* \neq \lambda_j^*, & i < j \end{cases} \] (4)

- Jaccard coefficient (JC)

\[ JC = \frac{a}{a + b + c} \] (5)

- FMI index

\[ FMI = \sqrt{\frac{a}{a + b} \cdot \frac{a}{a + c}} \] (6)
• Rand index, (RI) \[ \text{RI} = \frac{2(a + d)}{m(m-1)} \] (7)

2.3. Hilbert-Huang Transform (HHT)
HHT is an adaptive time-frequency analysis method suitable for non-stationary nonlinearity. This method can reflect the dynamic instantaneousness of the signal, can capture the local change of the signal, and can produce the base function according to the characteristics of the signal itself [17]. The classic HHT transform consists of two parts: The first step is to perform empirical mode decomposition EMD on the signal to obtain several intrinsic mode functions IMF; The second step is to perform Hilbert transform on each IFM to obtain the corresponding Hilbert spectrum. Summarizing all the Hilbert spectra of IMF will get the Hilbert spectra of the original signal.

2.4. Ensemble empirical mode decomposition EEMD
EMD can be widely used in nonlinear and non-stationary processes, decomposing the signal to reflect the dynamic instantaneous of the signal, but the main disadvantage is the frequent appearance of modal aliasing, which is caused by signal interruption. Interruption is an indefinite form of disturbance signal. Interruption can lead to confusion in time-frequency distribution, thereby destroying the physical meaning of IMF. Aiming at the deficiencies of the EMD method, Norden E. Huang et al. proposed a noise-assisted data analysis method, which is called ensemble empirical mode decomposition (EEMD) [7 8].

EEMD algorithm steps:
Step1: Add the normal distributed white noise \( w(t) \) to the original signal \( X(t) \) to obtain the processed signal \( X'(t) \);
\[ X'(t) = X(t) + w(t) \] (8)
Step2: The signal \( X'(t) \) added with normal distribution white noise is decomposed by EMD to obtain a series of intrinsic mode functions IMF, \( r_n(t) \) is the residual component of the decomposition.
\[ X'(t) = \sum_{j=1}^{n} c_j(t) + r_n(t) \] (9)
Step3: Repeat the first two steps, each time adding a new normal distribution white noise.
\[ X'(t) = \sum_{j=1}^{n} c_{ij}(t) + w_{in}(t) \] (10)
Step4: According to the average value of the white noise spectrum is 0, the obtained IMF integration average is obtained, and the final IMF component \( c_n(t) \) of the original signal decomposition is obtained.
\[ c_n(t) = \frac{1}{n} \sum_{i=1}^{N} c_{in}(t) \] (11)
Step5: For a given signal \( X(t) \), after EEMD decomposition, n IMF and remainder \( r_n(t) \) are obtained.
\[ X(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \] (12)

Among them, \( c_i(t) \) is each intrinsic mode function.
EEMD based on the principle of noise-assisted signal processing, the signal is balanced by adding small amplitude white noise. When the additional white noise is evenly distributed in the entire time-frequency space, at this time, the time-frequency space is divided into components of different scales by the filter bank. Because of the characteristics of zero mean noise, the noise will cancel each other after many average calculations. After EEMD decomposition, a series of intrinsic mode function and residual components with frequencies from high to low are obtained, while ensemble empirical mode decomposition EEMD generates different basis functions according to the characteristics of the signal.
itself, which reflects the adaptability of EEMD. The phenomenon of modal aliasing is effectively solved, and the real signal is retained by using the zero-mean characteristic of Gaussian white noise.

2.5. Implementation of MFCC algorithm based on HHT

The traditional MFCC is based on the short-time Fourier transform according to the characteristics of human speech utterance, which means that the signal in a sub-frame is considered to be steady-state linear. The sounding structure and sounding mechanism of birdsong of the double trachea are obviously different from that of humans, that is, the characteristics of dynamic transients are prominent. In this paper, Hilbert Huang transform HHT is used to replace FFT. This method can better deal with non-stationary and nonlinear signals; based on the local mean characteristics and time scale of the original signal itself, the EEMD method screens the signal from high to low frequency divided into the sum of a series of IMF components. Then perform Hilbert transform on the decomposed IMF components to get their respective instantaneous frequencies, and then synthesize the instantaneous spectra of all the IMF components to obtain the Hilbert spectrum. The obtained Hilbert spectrum is transformed by DCT through the Mel filter bank to achieve HHT-MFCC hybrid feature extraction. The extraction process is shown in Fig. 2.

![Fig. 2 HHT-MFCC algorithm flow chart](image)

The experimental procedure for the crane sound:
Step1 preprocess each crane audio file (endpoint detection, pre-emphasis, windowing, framing), and the window function is Hamming window;
Step2 Add n sets of normal distributed white noise to each frame X(t) to obtain X′(t);
Step3 Perform empirical mode decomposition on the noised signal to obtain n groups of IMFs, and obtain the average value of n groups of IMFs to obtain a group of IMF components;
Step4 Perform Hilbert transform on several IMFs obtained in step 3 to synthesize Hilbert marginal spectrum;
Step5 The marginal spectrum obtained in step 4 is passed through the Mel filter bank to transform the Crane sound into the Mel domain. Then logarithmic operation is performed, and finally DCT transformation is performed to realize HHT-MFCC feature extraction.
3. Experiment and the analysis of result

3.1. The introduction of experimental data

The experimental data in this paper are the audio files of 9 types of cranes in China obtained from the Chinese bird website [18] and the world bird database website [19] through Web Crawler.

The nine types of cranes are: white crane, black-necked crane, red-crowned crane, white-headed crane, gray crane, demoiselle crane, sandhill crane, white-naped crane and Sarus crane. The sampling frequency of these 9 types of crane audio files is 16000 (Hz); the audio files of these nine types of cranes and the duration of each audio file are shown in Tab.1:

| Audio_name                  | time(s) |
|-----------------------------|---------|
| Sarus Crane_0.wav           | 33.70   |
| Sarus Crane_2.wav           | 22.33   |
| Sarus Crane_3.wav           | 50.02   |
| White-naped Crane_0.wav     | 32.62   |
| White-naped Crane_1.wav     | 27.35   |
| Sandhill Crane_0.wav        | 118.23  |
| Sandhill Crane_1.wav        | 29.15   |
| Sandhill Crane_2.wav        | 65.04   |
| Sandhill Crane_3.wav        | 50.60   |
| Demoiselle Crane_0.wav      | 20.16   |
| Demoiselle Crane_1.wav      | 11.56   |
| Demoiselle Crane_2.wav      | 11.15   |

The data in Table 2 are the experimental data of cluster analysis, and the data in the table are the MFCC and HHT-MFCC characteristics of the nine types of crane sounds. Among them, mfcc (selected_seg) and hht-mfcc (selected_seg) features are extracted by manually selecting segments with better sound quality after audio endpoint detection; mfcc and hht-mfcc features are extracted from 9 types of Heming sounds. This paper uses cluster analysis experiments to verify the effectiveness of the HHT-MFCC method and the MFCC method, and the noise immunity of the HHT-MFCC method when the audio file contains environmental noise.

| Data               | Sample | Attributes | Category |
|--------------------|--------|------------|----------|
| mfcc               | 12612  | 13         | 9        |
| hht-mfcc           | 12612  | 13         | 9        |
| mfcc(selected_seg) | 1809   | 13         | 9        |
| Hht-mfcc(selected_seg) | 1809 | 13         | 9        |
3.2. Experimental parameter settings
The algorithm parameters of the experiment in this paper are shown in Table 3.

| parameters  | Parameters value |
|-------------|------------------|
| frameSize   | 256              |
| overlap     | 0                |
| Nstd        | 0                |
| NE          | 1                |

3.3. Experimental results
The experiment uses the HHT and MFCC fusion method to extract the features of these 9 types of crane audio, and obtain the HHT-MFCC features and the traditional MFCC features. The characteristic result is shown in Fig.3:

![Cepstral Coefficients](image1)

(a) White-naped Crane hhtmfcc  
(b) Demoiselle Crane hhtmfcc

![Cepstral Coefficients](image2)

(c) Sarus Crane hhtmfcc  
(d) Sandhill Crane hhtmfcc

Fig.3 HHT-MFCC corresponding to the four cranes

The experiment compared the cepstrum of HHT-MFCC and mfcc, and the result is shown in Fig.4. The results show that the cepstrum of the MFCC of the same Crane sound is quite different from the cepstrum of HHT-MFCC, and HHT-MFCC has a good reduction characteristic for the envelope change of the original signal.
In this paper, K-means clustering is used to compare and analyze the clustering effect of HHT-MFCC and MFCC through three indicators of RI, JC and FMI. The experimental results are shown in Tab.4. Experimental results show that the three indexes of HHT-MFCC feature contrast clustering are 7% higher than the traditional MFCC index on average. Especially the RI index increased by nearly 10%.

Tab.4 Comparison of clustering results

| data             | RI   | JC    | FMI   | avg  |
|------------------|------|-------|-------|------|
| mfcc(selected_seg)| 0.7199 | 0.708 | 0.847 | 0.7583 |
| mfcc             | 0.754 | 0.7238| 0.8682| 0.782 |
| hht-mfcc(selected_seg)| 0.8263 | 0.8162| 0.9086| 0.8504 |
| hht-mfcc         | 0.8227 | 0.8126| 0.907 | 0.8474 |
The data in Tab4 are the three index values and average values obtained by clustering the MFCC and HHT-MFCC features of nine kinds of crane sounds, \textbf{mfcc}(selected\_seg) and \textbf{hht-mfcc}(selected\_seg) are manually selected after obtaining all the features.

![Fig.5 Result of K-means](image)

\textbf{Fig.5 Result of K-means}

4. Conclusion
Due to the uniqueness of the bird's vocalization mechanism, it has typical non-stationary and nonlinear characteristics. Aiming at the dynamic and instantaneousness of the Crane sound signal, the MFCC feature extraction method based on the HHT transform use K-means clustering to cluster the audio features of Crane sound extracted from 9 types of cranes.

Based on the results and discussions presented above, the conclusions are obtained as below:

(1) The results show that the Heming sound feature of MFCC based on HHT is 7% higher than the three average indexes of traditional MFCC feature clustering.

(2) Through cluster analysis of \textbf{hht-mfcc} (selected\_seg) and \textbf{hht-mfcc}, the average values of the three indicators are 0.8504 and 0.8474 respectively. The experiment data demonstrates that the HHT-MFCC method has the ability of anti-noise to some extent.

Acknowledgments
This research was funded by National Natural Science Foundation of China (Grant No. 31860332, No.61462078) and Major scientific and technological projects in Yunnan Province(Grant No. 202002AA10007).

References
[1] Miao Jin. Research on Speech Recognition Method of Unspecific Isolated Words Based on Echo State Network [D]. Lanzhou University, 2012.
[2] Zhang Zhen, Wang Huaqing. Improved algorithm of Mel cepstrum MFCC in speech signal feature extraction [J]. journal6, 2008, 44(022): 54-55.
[3] Li Zhizhong, Teng Guanghui. Poultry vocal feature extraction method based on improved MFCC [J], 2008, 11(24): 202-205.
[4] F-I Kazi, Bhalke D-G. Musical instrument classification using higher order spectra and MFCC[A]/2015.
[5] Gao Ming, Sun Rencheng. Speaker feature parameter extraction algorithm based on improved MFCC [J]. Journal of Qingdao University (Natural Science Edition), 2019, 032(001): 61-65, 73.

[6] Yang Zhaofen, Wang Qishan. Taxonomy and Tracheal Evolution of Crane Birds [J]. Wuyi Science Journal, 2004, (1): 130-135.

[7] Lei Fumin, Wang Gang. The effect of bird song behavior on its species differentiation and new species formation [J]. Acta Zootaxonomica Sinica, 2002, (03): 235-242.

[8] Wang Enze, He Dongjian. Bird recognition method based on MFCC and dual GMM [J]. Computer Engineering And Design, 2014, 35(005): 1868-1871.

[9] H-U Zhengquan, Yuming Zeng, Yuan Zong, et al. Improvement of MFCC parameters extraction in speaker recognition[J]. Computer Engineering & Applications, 2014.

[10] Shen Yi, Shen Zhiyuan. A Non-Linear and Non-Stationary Adaptive Signal Processing Method—An Overview of Hilbert-Huang Transform: Development and Application [J]. Automation Technology and Application, 2010, (05): 6-10.

[11] Wu Wenjun. Non-stationary signal analysis based on HHT and machine learning [D]. 2019.

[12] T-U Binbin, Fengqin Y-U. Speech emotion recognition based on improved MFCC with EMD[J]. Computer Engineering & Applications, 2012.

[13] Semion Kizhner, Blank Karin, Flatley Thomas, et al. On certain theoretical developments underlying the Hilbert-Huang transform[A]/2006.

[14] W-U Zhaohua, Huang Norden-E. ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD[J]. Advances in Adaptive Data Analysis, 2009.

[15] Wu ZHAOHUA, HUANG NORDEN-E. ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD[J]. Advances in Adaptive Data Analysis, 2011, 1(01).

[16] Zhou Zhihua. Machine learning [M]. Tsinghua University Press, 2016: 425.

[17] Huang Meng, Wang Chuanfei, et al. Application of EMD Denoising and Wavelet Transform in Bearing Fault Diagnosis [J]. Journal of Academy of Armored Force Engineering, 2013, 27(003): 39-43.

[18] https://birdnet.cn/

[19] https://avibase.bsc-eoc.org/