Automated Detection of Wilson’s Disease Based on Improved Mel-frequency Cepstral Coefficients with Signal Decomposition

Zhenglin Zhang$^{1,2}$, Li-Zhuang Yang$^{1,2,3}$, Xun Wang$^{4,*}$, Hai Li$^{1,2,3,*}$

$^1$Anhui Province Key Laboratory of Medical Physics and Technology, Institute of Health and Medical Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China
$^2$University of Science and Technology of China, Hefei, China
$^3$Hefei Cancer Hospital, Chinese Academy of Sciences, Hefei, China
$^4$Institute of Neurology, Anhui University of Chinese Medicine, Hefei, China

Abstract

Wilson’s disease (WD), a rare genetic movement disorder, is characterized by early-onset dysarthria. Automated speech assessment is thus valuable in early diagnosis and intervention. Time-frequency features, such as Mel-frequency cepstral coefficients (MFCC), have been frequently used. However, human speech signals are nonlinear and nonstationary, which cannot be captured by traditional features based on the Fourier transform. Moreover, the dysarthria type of WD patients is complex and different from other movement disorders such as Parkinson’s disease. Thus, developing sensitive time-frequency measures for WD patients is needed. The present study proposes DMFCC, the improved MFCC using signal decomposition. We validate the usefulness of DMFCC in WD detection with a sample of 60 WD patients and 60 matched healthy controls. Results show that the DMFCC achieves the best classification accuracy (86.1%), improving by 13.9%-44.4% compared to baseline features such as MFCC and the state-of-art Hilbert cepstral coefficients (HCCs). The present study is a first attempt to demonstrate the validity of automated acoustic measures in WD detection, and the proposed DMFCC provides a novel tool for speech assessment.

Index Terms: Wilson’s disease, Mel-frequency cepstral coefficients, signal decomposition, automated detection, dysarthria

1. Introduction

Wilson’s disease (WD), an autosomal recessive hereditary disease caused by the mutation of the ATP7B gene located on chromosome 13q14.3, is a rare disease that commonly occurs in adolescents with a higher prevalence in Eastern countries than in Western countries [1, 2, 3, 4]. Up to 68% of patients with WD have initial neurological manifestations, with dysarthria being the most frequent symptom, a speech disorder caused by damage to the central or peripheral nervous system, resulting in paralysis or motor incoordination of speech-related muscles [3, 5, 6]. Previous studies have shown that up to 97% of neurological WD patients have varying degrees of dysarthria, with 57.6% being the initial symptom [3, 6, 7]. The voice tremor, rhythm disorder, and low tone in the speech of WD patients can be characterized by speech signal analysis, which enables the automated detection of WD.

Automated speech analysis techniques have been widely used in assessing and diagnosing movement disorders, particularly in Parkinson’s disease (PD), with encouraging and promising results [8, 9, 10]. Some of these researches are based on specific acoustic tasks, such as isolated words pronunciation and sentence reading. However, it has been found that sustained phonation carries more disease discriminatory information and is more convenient to perform than isolated words and sentences [11, 12]. Previous studies have investigated articulatory and rhythmic aspects for the sustained phonation task, where typical acoustic features include voice quality and rhythm [13, 14, 15]. However, these features do not consider the time-frequency properties of the voice [12, 16]. In this regard, some other studies have performed PD detection based on spectral and cepstral features, including linear predictive coding (LPC) and Mel-frequency cepstral coefficients (MFCC) [17, 18]. However, these features only reflect the static characteristics of the speech signal. And there are limitations in implementing algorithms that directly use the Fourier transform to process nonlinear and nonstationary speech signals, ignoring the detailed information of the voice [19, 20].

Recently, adaptive signal time-frequency analysis based on empirical mode decomposition (EMD) has been used to diagnose speech disorders [21, 22, 23, 24]. EMD is a signal decomposition method that is not subject to the Heisenberg uncertainty principle and is particularly suitable for processing nonlinear and nonstationary signals [19]. It has been demonstrated that the intrinsic mode functions (IMFs) obtained after voice signal decomposition using EMD carry information about the vocal tract and vocal folds [25]. Karan et al. [25] proposed intrinsic mode function cepstral coefficient (IMFCC) based on EMD from sustained vowels to effectively characterize the PD patients’ voice, improving the accuracy by 10% over MFCC-based features. However, traditional EMD algorithm suffers from mode mixing, end effects, sensitivity to noise, and lack of complete mathematical theory [19, 26]. To overcome these drawbacks, another study by Karan and collaborators [26] combined variational mode decomposition (VMD) with Hilbert spectrum analysis to propose Hilbert cepstral coefficients (HCCs), which can be used as a biomarker for the detection and assessment of PD. Currently, EMD has been extensively studied in PD detection. However, different from the hypokinetic dysarthria in PD, the dysarthria in WD is a mixed type [6, 27]. Whether the features and models of these studies could be generalized to the WD population remains ambiguous. To our knowledge, few relevant works have been performed in this field.
Therefore, we proposed an improved MFCC method with signal decomposition (DMFCC) based on complete ensemble EMD with adaptive noise (CEEMDAN). CEEMDAN avoids the mode mixing and reconstruction error of the traditional EMD method, while it effectively improves the computational efficiency and has well completeness compared with the VMD method, which has boundary effects, sudden signal onset, and predefined mode numbers that are sensitive to mode mixing, leading to complicated parameter tuning [28, 29]. Meanwhile, DMFCC can capture the dynamic nonlinear features in speech signals compared to standard MFCC. We performed a classification experiment using SVM on WD patients and matched healthy controls (HCs). Simultaneously, we compared the performance of different signal decomposition methods, HCCs, and standard MFCC features on the same dataset. Our results demonstrated for the first time the effectiveness of automated acoustic measurements in WD detection and that the proposed DMFCC features are promising in speech disorder assessment.

2. Dataset

We recruited 120 native Mandarin speakers, including 60 WD patients diagnosed according to the Chinese WD clinical guidelines [30] (30 males, aged 19-44, 28.93±6.34 yrs, and 30 females, aged 18-37, 28.00±5.34 yrs), and 60 gender-matched HCs (30 males, aged 22-33, 24.10±2.64 yrs, and 30 females, aged 22-40, 24.80±3.51 yrs). All subjects were free of cognitive impairment, psychiatric disorders, primary language deficits, or substance abuse and signed informed consent forms. The study was approved by the local ethics committee and performed following the Declaration of Helsinki.

A trained neurologist performed the task with a standardized recording protocol and speech battery. Subjects were asked to take a deep breath and then sustain the vowel /a/ as long and steadily as possible at a comfortable pitch and loudness, which was repeated three times. The speech was recorded through a condenser microphone (AT2035, Audio-Technica, Japan) placed approximately 10 cm in front of the subject’s mouth in a small room with low ambient background noise (less than 45 dB C-weighted). The microphone was connected to a professional sound card (Scarlett Solo 3rd Gen, Focusrite Audio Engineering, High Wycombe, UK). The audio signal was transferred to a laptop computer, with a sampling frequency of 44.1 kHz and 16-bit resolution, and saved as a mono WAV format.

3. Methodology

3.1. Complete ensemble EMD with adaptive noise (CEEMDAN)

CEEMDAN is an improved EMD method by superimposing Gaussian white noise on the original signal several times, achieving decomposition completeness with fewer iterations and thus reducing the computational cost [28]. The CEEMDAN pseudocode used in this paper is described in Algorithm 1.

As shown in Figure 1, we decomposed the synthesized vowel /a/ into IMFs based on CEEMDAN and used LPC to obtain the spectrum of the first 8 IMFs (fundamental frequency: F0 = 125 Hz; first four formant frequencies: F1 = 800 Hz, F2 = 1200 Hz, F3 = 2300 Hz, and F4 = 2800 Hz). It is observed that IMF 1 and 2 carry F2 and F3 information; IMF 3-5 have F1 and F3 information; IMF 6 and 7 contain F1, F3, and F4 information; while IMF 8 carries F3 and F4 information. Thus, CEEMDAN can decompose the speech signal into IMFs characterizing the speaker’s vocal tract and vocal fold information, making it feasible to derive features from the IMFs that may contribute to the effective identification of WD.

3.2. Features extraction

We obtained the 13-dimensional MFCC applied 26 Mel filters through a standard computational pipeline, then calculated the dynamic features corresponding to each dimension and got the global statistical features of all dimensions, including the mean, standard deviation, skewness, and kurtosis. Also, we replicated the HCCs from Karan and his collaborators’ research [26]. The standard MFCC and HCCs were used as the baseline features, and the strength of the proposed CEEMDAN-based DMFCC features for WD detection was validated. In addition, we applied three different signal decomposition methods to HCCs and DMFCC features, including traditional EMD, VMD, and CEEMDAN, to demonstrate the effectiveness of the CEEMDAN method in improving WD detection performance. For all the above features, we calculate the mean of the features for each subject’s three speech samples as the final feature.

Algorithm 1: CEEMDAN

Input: Voice signal S, the trials number I, and the amplitude \( \beta_n \) (n = 0, 1, ..., k − 1) of white noise \( \omega_i \) (i = 1, 2, ..., I).

Output: IMFs \( n = 1, 2, \ldots, k \) and the final residue r.

1: IMFs \( S + \beta_n \omega_i \) \( \forall E_n(\cdot) \) is the nth mode obtained by EMD. */
2: \( n \leftarrow 2 \)
3: while \( IMF_n \) do not fulfil the EMD stopping criteria do
4: \( IMF_n \leftarrow \sum_{i=1}^{I} E_i[|r_{n-1} + \beta_n E_{n-1}(\omega_i)|] \)
5: \( n \leftarrow n + 1 \)
6: end while
7: \( r \leftarrow S - \sum_{i=1}^{k} IMF_n \)

Figure 1: LPC spectrum of the synthesized vowel /a/ and its first 8 IMFs.
IFD are the absolute energy and frequency deviation in different frequency bands, respectively [26]:

\[
\begin{align*}
    \text{IED} &= \sum_{i=1}^{M_e} \frac{|I_{E_i} - \text{mean}(I_E)|}{M_e}, \\
    \text{IFD} &= \sum_{i=1}^{M_f} \frac{|I_{F_i} - \text{mean}(I_F)|}{M_f}
\end{align*}
\]

(1)

where \(I_{E_i}\) and \(I_{F_i}\) are the \(i\)th instantaneous energy (IE) and instantaneous frequency (IF) of an IMF signal, separately; \(M_e\) and \(M_f\) are the lengths of the sequences \(I_{E_i}\) and \(I_{F_i}\), respectively.

The final HCCs are calculated as:

\[
\text{HCCs} = \text{det}(\{(\text{IED}_1, \cdots, \text{IED}_k), (\text{IFD}_1, \cdots, \text{IFD}_k)\})
\]

(2)

where \(\text{det}(\cdot)\) is the discrete cosine transform (DCT) function and \(k\) is the number of IMFs to be determined.

Furthermore, besides keeping in line with the previous study of getting IMFs using VMD, we obtained IMFs based on both the traditional EMD and CEEMDAN to contrast the performance of different signal decomposition methods in identifying WD.

3.2.2. Improved MFCC with signal decomposition (DMFCC)

We introduce the signal decomposition method into the traditional MFCC computation flow to compensate for its shortcomings in processing nonlinear and nonstationary speech signals. Figure 2 shows the feature calculation flow as follows:

1. **Preprocessing of the speech signal.** Pre-emphasis is used to balance the spectrum and improve the signal-to-noise ratio, followed by applying a window function on the framed signal to reduce spectral leakage.

2. **Mode decomposition.** Consistent with HCCs, we used three methods to obtain the IMFs, the traditional EMD, VMD, and CEEMDAN, to verify the advantages of the CEEMDAN-based MFCC features in WD detection. Then we applied the short-time Fourier transform (STFT) to each IMF and stacked the results in order of the mode components’ frequencies to get the spectral matrix \(D(\omega)\):

\[
D(\omega) = [X_k(\omega)X_{k-1}(\omega) \cdots X_1(\omega)]
\]

(3)

where \(k\) is the number of IMFs to be determined and \(X_k(\omega)\) is the complex-valued spectral matrix of the \(k\)th IMF after the STFT with angular frequency \(\omega\) (rows = frame number of the signal, columns = \(1 + N/2\), where \(N\) is point number of the STFT).

3. **Mel filter bank.** We calculated the periodogram-based power spectrum estimate of the spectral matrix \(D(\omega)\), then applied 26 Mel filters and summed the energies in the frequency bin range of each filter to obtain the Melf spectrum \(S_{\text{mel}}\):

\[
S_{\text{mel}}(m) = \sum_{n=1}^{k(1+N/2)} \frac{|D_n(\omega)|^2 H_m(n)}{k(1+N/2)}, 0 < m \leq M
\]

(4)

where \(M\) is the filter number, and \(H_m(n)\) is the Mel-scale triangular filter function corresponding to the \(n\)th frequency point.

4. **Cepstral coefficients.** We took the logarithm of \(S_{\text{mel}}\) and performed DCT to obtain the first 13 values as the cepstral coefficients \(C\) (i.e., the number of the coefficients \(L = 13\)):

\[
C_i = 10\log(S_{\text{mel}}(m)) \cos \frac{\pi i(2n+1)}{2M}, i = \{1, 2, \cdots, L\}
\]

(5)

5. Dynamic characteristics and statistical features. The delta (\(\Delta\)) and delta-delta (\(\Delta^2\)) coefficients were calculated and appended to the original feature vector \(C\) to describe the dynamic properties of the speech. In addition, to keep the feature vectors aligned and characterize the global features, we computed the statistics of each dimension at the utterance level, including the mean, standard deviation, skewness, and kurtosis.

According to previous studies, we set the number \(k\) of IMFs obtained based on traditional EMD and VMD methods to 6 and 4, respectively [22, 25, 26]. To determine the CEEMDAN-based \(k\) value, we divided the 120-sample dataset into training and test sets by 7 to 3, then used the support vector machine (SVM), random forest, and multilayer perceptron under the scikit-learn’s default parameters to model the training set [31]. Figure 3 shows that both CEEMDAN-based HCCs and DMFCC features have the highest classification accuracy on the test set when \(k = 8\).

3.3. Classification

We split the 120-sample dataset 7 to 3 at the speaker level into training and test sets and then implemented the WD identification task using an SVM classifier. The optimal hyperparameters were obtained by grid search, and the classifier was evaluated on the training set using 10-fold cross-validation (CV). Five metrics were calculated and used as indicators to evaluate the model performance, which are accuracy (Acc), F1 score (F1), sensitivity (Sen), specificity (Spe), and the area under the receiver operating characteristic (ROC) curve (AUC). The complete code of this study is available online \(^1\).

4. Results and Discussion

Table 1 shows the classification results of WD vs. HC. Compared with the standard MFCC, the proposed DMFCC improves the accuracy from 72.2% to 75%-86.1% on the test set, and all

\(^1\)https://github.com/zlzhang1124/WD-detection
other metrics are improved to various degrees. It indicates that introducing time-frequency analysis theory based on mode decomposition overcomes the limitations of the traditional MFCC in nonlinear time-varying systems. The decomposed IMFs contain the time-frequency information of the original voice signal at different levels, which can well capture the physiological message of speech in patients with dysarthria, reflecting the pathological variation of their vocal organs [12].

For another baseline feature, the HCCs based on the VMD method (HCCs-VMD), the test set accuracy is limited to 41.7%, and the accuracy of HCCs based on both EMD and CEEMDAN is poor. On the one hand, HCCs are represented by the frequency and energy of IMFs, not measuring the dynamic properties of speech signals. On the other hand, HCCs are an improved framework proposed for PD assessment, whereas the articulatory characteristics exhibited in the WD population differ significantly from PD [6, 8], leading to HCCs might be inappropriate for the WD detection. Nevertheless, the HCCs feature set with the decomposition method replaced by CEEMDAN performed almost the best on the test set, with an accuracy of 61.1%. Therefore, it is demonstrated that the CEEMDAN approach can enhance the WD detection performance to some extent.

In addition, the different mode signals obtained by the traditional EMD method could be mixed up, leading to erroneous time-frequency distribution, which affects the accurate articulatory characteristics [20, 26]. Although the VMD approach avoids the EMD limitations to a certain extent, it introduces new issues, such as the predefined mode number sensitive to mode mixing [29]. In comparison, CEEMDAN with better completeness overcomes the mode mixing problem and effectively improves the computational efficiency [28], as evidenced by the performance of the proposed DMFCC on WD classification. The ROC curves of DMFCC features based on three different signal decomposition approaches on the test set are shown in Figure 4. Compared with EMD and VMD, the AUC performance of CEEMDAN improves from 81% and 76% to 83%. Furthermore, the CEEMDAN-based decomposition method achieves optimal performance in all other metrics, especially the sensitivity of 94.4% on the test set, which improves 16.6%-22.2%. Therefore, DMFCC-CEEMDAN is ideally suited as an indicator for WD early screening. Combined with the finding that CEEMDAN is also the optimal method on HCCs, we further demonstrate the superiority of the proposed CEEMDAN-based DMFCC features in WD detection.

The Mel-scale enables DMFCC to reflect the speech production mechanism and the nonlinear nature of auditory perception [17], while the mode decomposition allows DMFCC to adapt the nonlinear and nonstationary speech signals. Our results show the strength of the proposed DMFCC features in WD detection and validate that automated acoustic measures are a promising approach to capture the subtle differences.

5. Conclusions

The present study proposes an improved MFCC feature using signal decomposition. The CEEMDAN-based DMFCC method achieves the highest accuracy of 86.1% in the WD detection task, improving by 13.9%-44.4% compared to the baseline features, demonstrating the usefulness of DMFCC features in WD diagnosis. As far as we know, the present study is the first attempt at automated detection of WD using acoustic measures, which provides a promising efficient approach for the early diagnosis of WD. Future studies should expand cross-center collaborations with enhanced sample size and further distinguish between different movement disorders, in addition to linking abnormal acoustic feature patterns with potential neurobiological substrates using neuroimaging.

6. Acknowledgements

This work was supported by the Natural Science Fund of Hefei City (2021033), the Natural Science Fund of Anhui Province (2008085MC69), the General scientific research project of Anhui Provincial Health Commission (AHWJ2021b150), Collaborative Innovation Program of Hefei Science Center-CAS (2020HSC-CIP001, 2021HSC-CIP013), Anhui Province Key Laboratory of Medical Physics and Technology (LMPT201904).
7. References


