The Speech Recognition of Double-Syllable Chinese Words Based on the Hilbert Spectrum

Tianyang Long[1], Long Zhang[2], Tingfa Xu[3], Shuangwei Wang[1*]

[1] School of Physics, Northeast Normal University, changchun, Jilin, China
[2] Liaohe Oilfield of China National Petroleum Corporation (Chaoyang) Gas Co., Ltd., Liaoning, China.
[3] School of Optoelectronics, Laboratory of Photoelectric Imaging and Information Engineering, Beijing Institute of Technology, China.

* Corresponding author. Tel.: 0086-15526893016; email: wsw77@nenu.edu.cn
Revised manuscript submitted July 11, 2017; accepted October 12, 2017.
doi: 10.17706/jsw.12.9.732-743

Abstract: Here a Chinese lexical recognition task is studied by a small vocabulary including 40 double-syllable Chinese words. In the approach presented, the Hilbert-Huang Transform (HHT) which consists of two steps is applied to speech signal analyzing. First, the speech signals are decomposed into a set of intrinsic mode functions (IMFs) by using the empirical mode decomposition (EMD) technique. Second, the first two IMFs are retained for further Hilbert spectral analysis. Final presentation of the speech signal is an energy-frequency-time distribution designated as the Hilbert spectrum, which can be used to depict the characteristics of speech sounds. For feature extraction, the Hilbert spectrum of each speech signal is divided into a set of frequency sub-bands. The number of discrete points on the Hilbert spectrum each sub-band contained is calculated as an element of the feature vector. Feature vectors obtained are fed to Support Vector Machine (SVM) classifier for classification. The proposed method is evaluated using 3840 speech samples from 8 different speakers (4 male). The experimental result, overall recognition rate of the 40 words achieving around 97% demonstrates the effectiveness of this approach.

Key words: Speech recognition, empirical mode decomposition, hilbert-huang transform, hilbert spectrum.

1. Introduction

For non-stationary signals, there are some conventional data processing methods summarized by Huang *et al.* [1]*:* (a) *short-time Fourier analysis*; (b) *the wavelet analysis*; (c) *the Wigner-Ville distribution*, etc. Both (a) and (b) rely on the traditional Fourier spectral analysis. (a) has to assume the data to be piecewise stationary. The problem with (b) is its leakage generated by the limited length of the basic wavelet function. The difficulty with (c) is the severe cross terms for some frequency ranges. Huang *et al.* [1] developed a new method for analyzing nonlinear and non-stationary data. The key part of the method is the EMD method, with which any complicated data set can be decomposed into a finite and often small number of IMFs that admit well-behaved Hilbert transforms. This decomposition method is adaptive and highly efficient. Since the decomposition is based on the local characteristic time scale of the data, it is applicable to nonlinear and non-stationary processes. With the Hilbert transform, the IMFs yield instantaneous frequencies as functions of time that give sharp identifications of imbedded structures. The final presentation of the results is an energy-frequency-time distribution designated as the Hilbert spectrum.

Since the HHT method proposed, there are lots of practical applications for speech signals [2]-[4]. Chowdhury deals with the speech based gender identification problem using the EMD method [5]. Zao et al. f proposed a speech enhancement method with EMD [6]. Based on the EMD, Hybrid speech enhancement is implemented by El-Moneim *et al.* [7], [8] proposed a method to classify the voiced/non-voiced speech

signals using adaptive thresholding with bivariate EMD.

This paper attempts to adopt the HHT method for speech recognition of double-syllable Chinese words. In common vocabulary of mandarin, double-syllable words account for 72% [9]. 40 of the most frequently used 100 double-syllable Chinese words are chose for test. They almost contain all different initials, finals, and tones, all of which constitute Chinese phonetic transcriptions. More detailed information about these 40 words is given in ***Appendix***. The general procedure of this method is as follows: (1) the speech samples are decomposed into a finite set of intrinsic mode functions (IMFs) by the EMD. The first two IMFs are retained for the next step. (2) Hilbert transform is applied to IMF1-2 and a time-frequency spectrogram is performed. (3) Feature extraction based on the Hilbert spectrum. Our aim is not to generalize the results of this approach when it is applied to the speech signals of all double-syllable Chinese words but to tentatively show its efficiency in analyzing speech signals within a certain vocabulary. Finally the results demonstrate the effectiveness of this method.

2. Hilbert-Huang Transform (HHT)

The Hilbert-Huang Transform (HHT) consists of two steps. First, the signals are decomposed into a set of intrinsic mode functions (IMFs) by the empirical mode decomposition (EMD) technique. Second, IMFs are retained for further Hilbert spectral analysis.

2.1. The Empirical Mode Decomposition (EMD)

In EMD processing, speech signals are decomposed into a finite set of intrinsic mode functions (IMFs) based on the localization of their extrema. First, calculate the envelopes of the maximum and minimum. Second, find the average of the two envelopes. Third, a residual signal is obtained (Residue = Original signal-Mean signal). This sifting process is repeated until the residual is a constant, a monotonic slope, or a function with only one extrema. The signal s is decomposed to

$$s[n] = \sum_{i=1}^{M} c_i[n] + T_M[n] \tag{1}$$

where ci is the ith intrinsic mode of the signal s (i.e. the intrinsic mode function). TM is the residue. This decomposition is complete and the sum of all modes ci equals to the signal s.

2.2. The Hilbert Spectrum

Huang *et al.* [1] apply the Hilbert transform to each IMF as follows:

$$s_H[n] = \mathrm{HT}(s[n]) = s[n] \otimes \frac{1}{\pi n} \tag{2}$$

The analytic signal is defined as:

$$s_A[n] = s[n] + j s_H[n] \tag{3}$$

where $j^2 = -1$. s_A can be written as:

$$s_A[n] = a[n] e^{j\theta[n]} \tag{4}$$

with the magnitude $a[n] = \sqrt{s^2[n] + s_H^2[n]}$ and the phase $\theta[n] = arctan \frac{s_H[n]}{s[n]}$. Based on the Hilbert transform, the instantaneous frequency is defined by:

$$f[n] = \frac{1}{2\pi} \frac{d\theta[n]}{dn} \tag{5}$$

The instantaneous frequency is not limited to the uncertainty principle. Each IMF (i=1,…,M) can be represented by the triplet {n, $f_i[n]$, $a_i[n]$} in the time-frequency plane. At last, the Hilbert spectrum H[n,f] is defined as:

$$H(n,f) = Re\sum_{i=1}^{M} a_{i(n)}\, exp(j\sum f_i(n))\,, f_i(n) = f \tag{6}$$

3. Feature Extraction

8 participants (4 male) with standard mandarin pronunciation are recruited to read the 40 words, each of which was recorded for 12 times. At last 3840 speech samples were collected for further experiments. More details about the speech samples are presented on **Table 1**.

Table 1. Details about the Speech Samples

Gender	Speech length	Sample rates	Recording environment
Female	800ms	8kHz	Noise-free
Male	600ms	8kHz	Noise-free

3.1. Applying the EMD Method to Decompose Speech Signals into IMF Components

Fig. 1 shows an example of Word 1 and its IMFs obtained by the EMD method. As shown, the IMFs do capture the information of the original signal. However, due to the nature of EMD algorithm, as the order of the IMF increases, the relative sense of the data approaches to zero [10]. Top 5 IMFs have occupied most energy of the speech sample. We choose IMF1-2 for Hilbert spectral analysis.

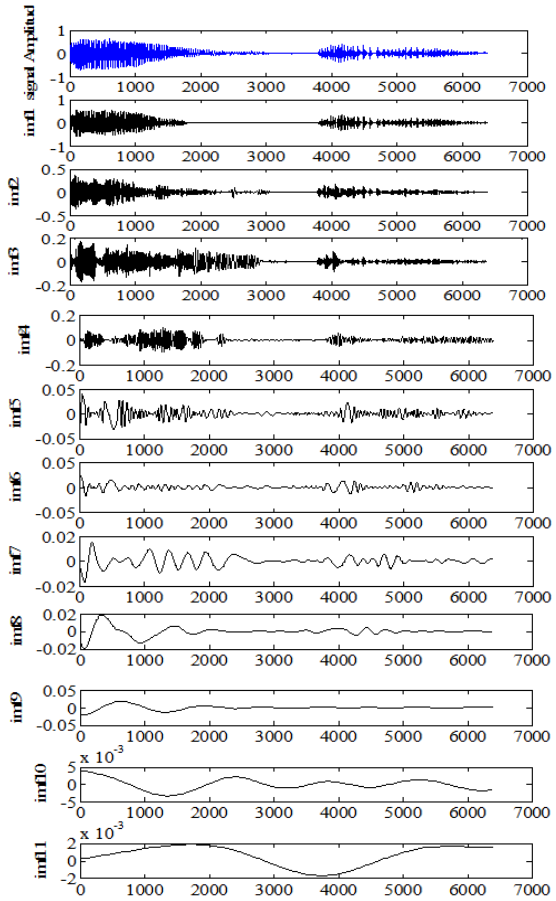


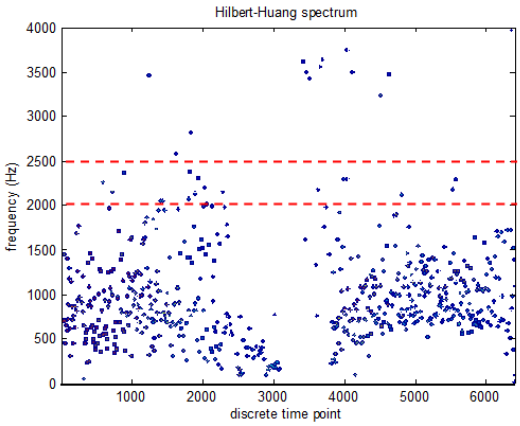
Fig. 1. The original signal of **Word 1** and its IMFs

3.2. Feature Extraction Based on the Hilbert Spectrum

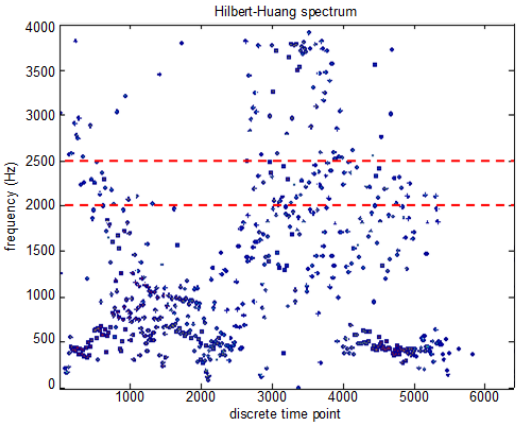
The Hilbert spectrum is a 3D image in which discrete time points are plotted along the horizontal axis, while instantaneous frequencies are plotted along the axis. The intensity gives the amplitude of speech

energy at a cross coordinate. To extract features from speech signals of different Chinese words, we exploit the Hilbert spectrum to depict the characteristics of speech sounds.

The spectrograms of the first two IMFs of **word1** and **word2** are shown in **Fig. 2**. Obviously, discrete spectro-temporal events visible on the speech spectrogram underlie the perception of individual speech sounds [11]. For feature extraction, the distribution of speech energy in time-frequency domain (*i.e.* the position of the discrete points on the spectrum) is initially focused on only, while the amplitude of the energy is not considered. It is noticed that the quantity of discrete points in the region of the frequency sub-band (2000-2500Hz, for example) enveloped with dotted lines on picture **(a)** is different from **(b)**. According to this characteristic, a banding technique in frequency domain of the Hilbert



(a) Word 1



(b) Word2

Fig. 2. The Hilbert spectrums IMF1-2 of Word 1 and Word2.

Spectrum is developed. The banding process consists of two steps.

First, determine the dimension of the feature vector, which equals to decide how many sub-bands the Hilbert spectrum divided into. At the same time make sure that the sub-bandwidth is neither too wide for it is meaningless nor too narrow to be fault-tolerant. Here the Hilbert spectrum is divided into 27 bands along the frequency axis. More details are presented on Table 2.

Table 2. 27 Frequency SUB-Bands of the Hilbert Spectrum

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1-100	101-200	201-300	301-400	401-500	501-600	601-700	701-800	801-900
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
901-1000	1001-1100	1101-1200	1201-1300	1301-1400	1401-1500	1501-1700	1701-1900	1901-2100
19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
2101-2300	2301-2500	2501-2700	2701-2900	2901-3100	3101-3300	3301-3500	3501-3700	3701-4000

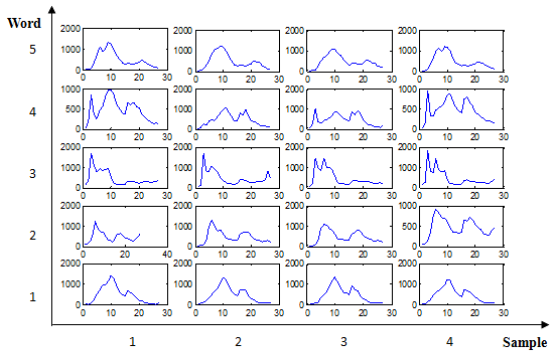


Fig. 3. Visualizations of the feature vectors belong to **Word1-5.**

Second, calculate the number of points in each sub-band as the value of each dimension of the feature vector. According to Equation (6), we define

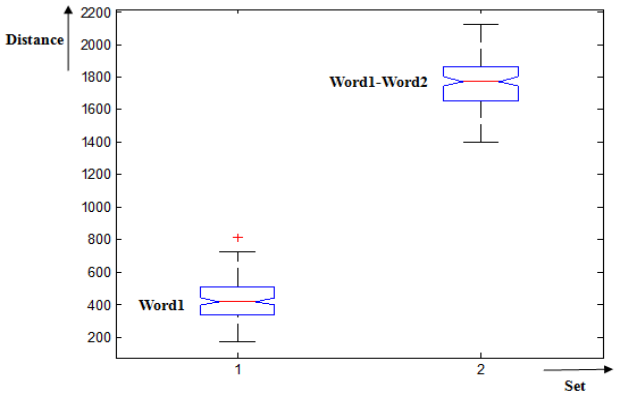
$$P = \sum sng\big(H(n,f)\big), M = 2 \tag{7}$$

P represents the number of discrete points in the frequency sub-band. Applying Equation (7), calculate the number of points in each sub-band and a 27-dimension feature vector of each speech sample is obtained. Due to space limitation, the visualizations of the feature vectors belong to Word1-5 (4 speech samples of each word) from Female1 are presented in **Fig. 3**. In each small image on **Fig. 3**, the horizontal axis represents the dimension of the feature vector, while the vertical axis represents the value of each dimension.

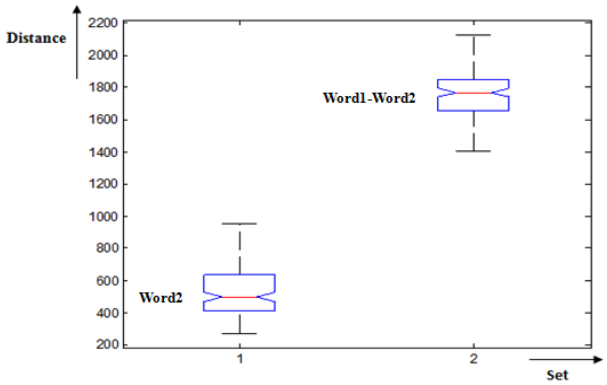
3.3. Feature Vectors Evaluation: ANOVA Test

If the Euclidean distances (the Euclidean distance between two n-dimensional vectors is defined as $d = \sqrt{(\boldsymbol{a}-\boldsymbol{b})(\boldsymbol{a}-\boldsymbol{b})^T}$) between the feature vectors of the speech samples from **one word** are far less than those from **two different words**, it means that the feature vectors collected possess higher recognition ability.

Feature vectors extracted from 960 speech samples (12 samples per word, 10 words (word1-10) per participant) are initially selected for statistical analysis. Due to space constraint, **Female1** is taken as an example for exhaustive demonstration. There are C_{12}^2 effective combinations of the feature vectors from **one word** while 144 from **two different words**. Correspondingly, we get the same quantity of Euclidean distances. The box plots of Euclidean distances between the feature vectors of the speech samples from **Word1** and those from **Word1-Word2** are shown in **Fig. 4 (a)**, those from **Word2** and **Word1-Word2** shown in **Fig. 4 (b).** As shown, the Euclidean distances between the feature vectors of speech samples from.



(a) From **Word1** and **Word1-Word2**



(b) From **Word2** and **Word1-Word2**

Fig. 4. Box plots of Euclidean distances between the feature vectors of the speech samples.

Word1-Word2 are much higher than those from **the same word** such as **Word1** or **Word2**. The Euclidean distances from two groups are quite concentrated respectively. A pairing ANOVA test is performed on **Word1-10** and the specific F-values are presented on **Table 3**. The F-value is computed as the ratio of the between-group variance in the data over within-group variance, which indicates relative discriminative power between the groups. Results that the high F-value with low p-value ($p<0.01$) suggest that there are significant differences between the two sets of Euclidean distances. There is a strong evidence for the feature vectors to be used in classifying the speech signals of double-syllable Chinese words.

Table 3. Specific F-Values Obtained by the Pairing Test on **Word1-10**

--	1	2	3	4	5	6	7	8	9	10
1	-	3856.6	3686.0	209.9	979.4	21892.4	1591.6	474.2	924.8	1801.4
2	2897.0	-	1022.3	1659.3	2635.0	6441.1	63.4	885.2	1222.6	1198.9
3	2301.6	673.2	-	1224.2	1356.6	1759.4	724.4	1125.7	672.2	588.4
4	47.5	1127.8	1403.8	-	339.6	6818.5	402.1	126.1	164.9	626.9
5	932.6	4655.4	2269.4	657.5	-	15200.8	1876.3	1256.8	817.7	2338.7
6	5695.9	1976.4	1185.2	2852.7	4133.9	-	2316.7	4846.1	2376.7	2056.0
7	199.3	297.6	103.6	5213.5	628.2	423.8	-	885.2	12.13	807.1
8	443.2	1477.0	1878.8	414.6	1266.8	29316.6	400.5	-	1012.3	1108.1
9	515.2	975.2	900.7	231.7	449.9	6908.0	355.1	466.0	-	145.3
10	1041.3	892.3	808.3	762.5	840.6	4227.6	248.7	517.0	148.0	-

4. Results and Discussion

Support vector machines (SVM) have shown great effectiveness in pattern classification. It is a learning system that separates input vectors into two classes with optimal separating hyperplane, which is optimally separated from input vectors when they are separated with no error and the distance between the nearest vectors to hyperplane is maximal [12], [13]. In this paper, C-SVC two-class SVM model is adopted to classify Chinese words.

4.1. Experiment on Word1-10

The 960 (12*10*8) speech samples from Word1-10 are initially fed to the SVM for classification. **Table 4** presents the recognition rates.

50% of 120 (12*10) speech samples from each volunteer are used as the training set, while the other 50% as the testing set. Of the 120 speech samples, every 12 samples come from the same **Word**. We divide each of the 12 samples into two groups: The first 6 samples are chosen to be the training set, while the last 6 samples to be the test set, which means there are 60 training samples and 60 testing samples (6*10) for

each person. After normalizing, the training set is used to train the SVM classifier to get the classification model to predict the testing set.

Table 4. The Recognition Rates of Word1-10 from 8 Volunteers

Female	**1**	**2**	**3**	**4**
Rate (%)	98.33	100	100	98.33
Male	**1**	**2**	**3**	**4**
Rate (%)	100	98.33	100	100

As shown in **Table 4,** the recognition rates of **Female1, Female4 and Male2** fail to reach 100%. The detailed results of these wrongly recognized words are given on **Table 5**. In **Table 5**, we summarize the experimental results from 8 individuals and find that 3 of them whose recognition rates failed to reach 100%. The overall recognition rate is calculated as follows: For one individual there are 60 testing samples (6*10, 12 recording speech samples from **one word** divided evenly into two sets). Count the number of wrongly recognized samples (N) and calculate the overall recognition by formula $R1 = (60 - N)/60 * 100\%$. The concrete classification result of **Female1** shows that one speech sample belongs to **Word2** is wrongly identified as **Word7**. Similarly, the diagram of testing set of **Female4** shows that one speech sample belongs to **Word6** is wrongly identified as **Word8**. Also, for **Male2** one speech sample belongs to **Word7** is wrongly identified as **Word2**. As shown in **Fig. 5**, we take **Female1** for example for space limitation.

Table 5. The Detailed Results of Wrongly Recognized Words
Rate--- Overall Recognition Rate; W-W---Wrongly Recognized Word

Volunteer	**Rate (%)**	**W-W**	**Wrongly recognized as Word**									
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Female1	98.33	Word2	-	-	-	-	-	-	1	-	-	-
Female4	98.33	Word6	-	-	-	-	-	-	-	1	-	-
Male2	98.33	Word7	-	1	-	-	-	-	-	-	-	-

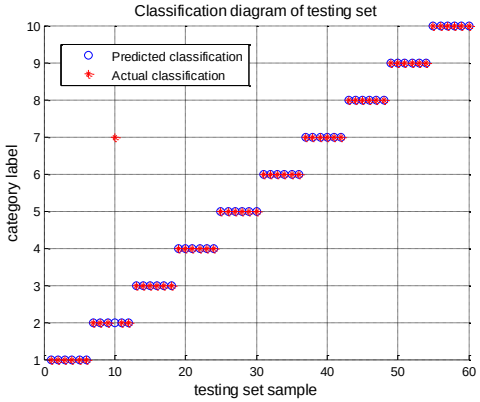


Fig. 5. The classification diagram (**Word1-10**) of **Female1.**

4.2. Experiment on Word1-40

In order to verify the effectiveness of the method further, the experiment is expanded to a larger database. 3840 (12*40*8) speech samples are studied with the same experiment procedure as before. There are 240 testing samples (6*40) for each person. The results are listed on **Table 6**. For space limitation, the detailed results of wrongly recognized words are added to ***Appendix*** to make the result concrete and clear**.** The corresponding classification diagram (**Female1**) is shown on **Fig.6.** As shown, a total of seven samples are misidentified. The overall recognition rate of the 40 words achieves 97.08%.

Table 6. The Recognition Rates of Word1-10 from 8 Volunteers

Female	**1**	**2**	**3**	**4**
Rate (%)	97.08	97.92	98.33	97.92
Male	**1**	**2**	**3**	**4**
Rate (%)	99.17	95.42	94.58	97.08

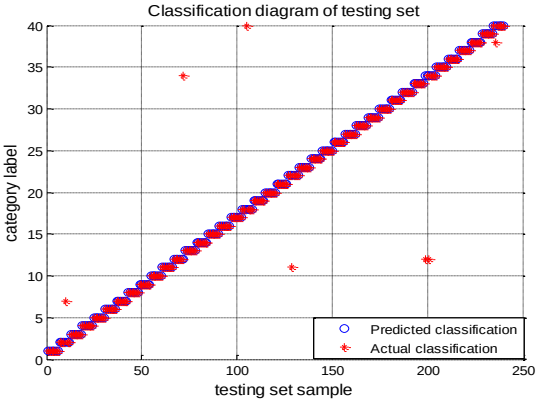


Fig. 6. The classification diagram (**Word1-40**) of **Female1.**

4.3. Discussion

The result, a relatively high recognition rate for each speaker, demonstrates great effectiveness of the proposed method. Taking **Female 1** for example, **Fig. 5** and **Fig. 6** present the concrete speech recognition results. Comparing the recognition rate of 10 words and 40 words shown on **Table 4** and **Table 5**, we find that the recognition rate decreases with the increase of capacity of the database on the whole. The recognition rate is affected by the size of database since the number of the words with similar pronunciation goes up with the increasing amount of the speech samples. This issue needs to be studied further. **Table 5**shows that the worst recognition rate is 94.58% while the best rate reaches 99.16%, both of them coming from Male speech samples, which means gender is not a major problem affecting recognition rates. The quality of the sample itself is probably the main factor affecting the recognition results.

Comparing with the recognition results of some classical signal analysis methods, Table 7 presents a summary of previous literatures. Related literatures using the same experimental material with our article have not been found. However, the results in this article may not be directly comparable to the former ones owing to different subjects, expressions, and number of features, etc. The result, as shown on Table 7, that a relatively high recognition rate with a lower vector dimension validates the method we proposed.

Table 7. A Summary of Previous Literatures

Literature	**Method**	**Feature**	**Experimental**	**Sample**	**Recognition**
Al-Assaf, Y. (2003)[14]	MRWA	---	Consonants (b,d,g)	90	97.8%
Selouani *et al.* (2007)[15]	MFCC	39	The TIMIT	---	86.0%
Lee *et al.* (2012)[16]	MFCC	39	stops, fricatives, and	54144	80.0%
Joshi, V.(2015)[17]	STFT	39	The Aurora-4	330	94.4%
Biswas, A.(2016)[18]	Wavelet	78	The BINICD	2700	98.1%
Gauthier, E.(2016)[19]	Cepstral	---	Hausa languages	1021	92.2%
Proposed method	**HHT**	**27**	**Double-syllable**	**3840**	**At around**

5. Conclusion

Experiments demonstrate the method we proposed for the speech recognition of double-syllable Chinese

words. Most recognition rates of **word1-40** are at around 98%, suggesting that the recognition effect is relatively stable.

This paper deals only with 40 common double-syllable Chinese words. In the future, other database such as four-word idioms, *etc.*, is to be studied. When extract features based on the Hilbert spectrum, we do not consider the value of speech energy. If it is able to contribute to the feature extraction is to be studied.

Appendix

The Detailed Results of 4.2 Experiment on Word1-40

Rate — Overall Recognition Rate; W-W — Wrongly Recognized Word

Volunteer	Rate (%)	W-W	Wrongly recognized as Word																
			5	7	8	11	12	16	19	23	24	25	29	33	34	37	38	39	40
Female1	98.33	Word2	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word12	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-
		Word18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
		Word22	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word34	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
		Word40	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-
Female2	97.92	Word13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
		Word19	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-
		Word23	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word24	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-
		Word30	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-
Female3	98.33	Word6	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-
		Word7	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
		Word19	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word23	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-
Female4	97.92	Word10	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-
		Word19	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word30	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
		Word34	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
		Word38	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-

Volunteer	Rate (%)	W-W	Wrongly recognized as Word																			
			2	6	7	9	12	14	15	17	19	20	21	24	27	28	31	33	34	37	38	39
Male1	99.17	Word28	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word31	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Male2	95.42	Word11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-
		Word14	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word19	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-
		Word20	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-
		Word21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-
		Word22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-
		Word25	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word33	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-
		Word38	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
		Word39	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Male3	94.58	Word4	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
		Word6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
		Word15	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word19	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-
		Word29	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word36	-	-	-	-	-	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-
		Word38	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Male4	97.08	Word12	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-
		Word19	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-
		Word39	-	-	-	-	-	-	-	-	1	-	-	2	-	-	-	-	-	-	-	-
		Word40	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-

Speech Sample Materials

Word	Chinese Pinyin[1]	phonetic notation	meaning
1 办法	bàn fǎ	´panfa	means
2 表示	biǎo shì	piɑu´ʂi	indicate
3 出现	chū xiàn	tʂhu´ɕian	appearance
4 参加	cān jiā	tshan tɕia	participate

5 代表	dài biǎo	ˊtaipiɑu	represent
6 第一	dì yī	ˊtii	first
7 而且	ěr qǐe	ɚ˴tɕhiɛ	furthermore
8 正在	zhèng zài	ˊʈʂəŋˊʦai	in process of
9 发生	fā shēng	faʂəŋ	happen
10 方面	fāng miàn	fɑŋˊmian	aspect
11 工作	gōng zuò	kuŋˊʦuo	work
12 革命	gé mìng	kɤˊmiŋ	revolution
13 活动	huó dòng	xuoˊtuŋ	activity
14 还是	hái shì	xaiˊʂi	still
15 继续	jì xù	ʨiˊɕy	continue
16 解决	jǐe jué	ʨiɛʨyɛ	solve
17 可以	kě yǐ	khɤi	passable
18 可能	kě néng	khɤnəŋ	possible
19 历史	lì shǐ	ˊliʂi	history
20 领导	lǐng dǎo	liŋtɑu	leadership
21 美国	méi guó	meikuo	America
22 没有	méi yǒu	meiiou	none
23 能够	néng gòu	nəŋˊkou	capable
24 组织	zǔ zhī	ʦuʈʂi	organization
25 其中	qí zhōng	ʨhiʈʂuŋ	Inside
26 情况	qíng kuàng	ʨhiŋˊkhuɑŋ	situation
27 人民	rén mín	ʐ̩ənmin	people
28 如果	rú guǒ	ʐ̩ukuo	if
29 上海	shàng hǎi	ˊʂɑŋxai	shanghai
30 甚至	shèn zhì	ˊʂənˊʈʂi	even
31 思想	sī xiǎng	siɕiɑŋ	mind
32 所有	suǒ yǒu	suoiou	all
33 特别	tè bíe	ˊthɤpiɛ	special
34 通过	tōng guò	thuŋˊkuo	pass
35 文化	wén huà	yənˊxua	culture
36 问题	wèn tí	ˊyənthi	question
37 学习	xué xí	ɕyɛɕi	learn
38 许多	xǔ duō	ɕytuo	many
39 研究	yán jīu	iantɕiou	study
40 一定	yí dìng	iˊtiŋ	definite

Acknowledgment

This work is supported by "Chinese Speech Recognition Based on Spectral Information and Research on Speech Enhancement Methods" Project Approval Number: 61471111

References

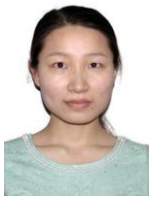
[1] Huang, N. E., Long, S. R., & Shen, Z. (1996). The mechanism for frequency downshift in nonlinearwave evolution. *Adv. Appl. Mech*., *32*, 59-111.

[2] Boudraa, A. O., & Cexus, J. C. (2006). Denoising via empirical mode decomposition.

[3] Khaldi, K., Alouane, M. T. H., & Boudraa, A. O. (2008, November). A new EMD denoising approach dedicated to voiced speech signals. *Proceedings of the* 2*nd International Conference on Signals, Circuits and Systems.*

[4] Liu, Z. F., Liao, Z. P., & Sang, E. F. (2005, August). Speech enhancement based on Hilbert-Huang transform. *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*.

[5] Chowdhury, M. H. (2014). Speech based gender identification using empirical mode decomposition (EMD) .
[6] Zao, L., Coelho, R., & Flandrin, P. (2014). Speech enhancement with emd and hurst-based mode selection. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *22(5)*, 899-911.
[7] El-Moneim, S. A., Dessouky, M. I., El-Samie, F. E. A., Nassar, M. A., & El-Naby, M. A. (2015). Hybrid speech enhancement with empirical mode decomposition and spectral subtraction for efficient speaker identification. *International Journal of Speech Technology*, *18(4)*, 555-564.
[8] Molla, M. K. I., Hirose, K., & Hasan, M. K. (2016). Voiced/non-voiced speech classification using adaptive thresholding with bivariate EMD. *Pattern Analysis and Applications*, *19(1)*, 139-144.
[9] Li, X. (2008). Lexicon of Common Words in Contemporary Chinese.
[10] Rilling, G., Flandrin, P., & Goncalves, P. (2003, June). On empirical mode decomposition and its algorithms. In IEEE-EURASIP workshop on nonlinear signal and image processing (Vol. 3, pp. 8-11). IEEER.
[11] Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *J. Acoust. Soc. Am*., *67*, 648–662.
[12] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2(2)*, 121-167.
[13] Cristianini, N., Nello, & Taylor, J. S. (2000). An Introduction to support vector machines and other kernel-based learning methods. Cambridge, UK: Cambridge University Press.
[14] Al-Assaf, Y. (2003). The application of wavelets transforms and neural networks to speech classification. *Intelligent Automation & Soft Computing*, *9(1)*, 45-55.
[15] Selouani, S. A., O'Shaughnessy, D., & Caelen, J. (2007). Incorporating phonetic knowledge into an evolutionary subspace approach for robust speech recognition. *International Journal of Computers and Applications*, *29(2)*, 143-154.
[16] Lee, S. M., & Choi, J. Y. (2012). Analysis of acoustic parameters for consonant voicing classification in clean and telephone speech. *The Journal of the Acoustical Society of America*, *131(3)*, EL197-EL202.
[17] Joshi, V., Bilgi, R., Umesh, S., Garcia, L., & Benitez, C. (2015). Sub-band based histogram equalization in cepstral domain for speech recognition. *Speech Communication*, *69*, 46-65.
[18] Biswas, A., Sahu, P. K., Bhowmick, A., & Chandra, M. (2016). Speech recognition using ERB-like Admissible wavelet packet decomposition based on perceptual sub-band Weighting. *IETE Journal of Research*, *62(2)*, 129-139.
[19] Gauthier, E., Besacier, L., & Voisin, S. (2016). Automatic speech recognition for African languages with vowel length contrast. *Procedia Computer Science*, *81*, 136-143.



Tianyang Long was born in 1993. She is a graduate student studying at Northeast Normal University. She focuses on speech recognition and signal processing.

Long Zhang was born in 1984. He has got the bachelor degree in engineering. He is majored in electronic information engineering.

Tingfa Xu the professor of Beijing Institute of Technology, He comes from School of Optoelectronics, Laboratory of Photoelectric Imaging & Information Engineering.

Shuangwei Wang is a corresponding author, the professor of Northeast Normal University. His research direction is digital signal processing & circuit and system.