International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)
ISSN: 2150-7988 Vol.1 (2009), pp.295-302
http://www.mirlabs.org/ijcisim

A New Approach to Extract Formant Instantaneous Characteristics for Speaker Identification

Limin Hou[1] and Juanmin Xie[2]

[1]School of Communication and Information Engineering, Shanghai University,
149 Yanchang Road, Shanghai, China
lmhou@staff.shu.edu.cn

[2]School of Communication and Information Engineering, Shanghai University,
149 Yanchang Road, Shanghai, China
xiejuanm@shu.edu.cn

Abstract**:** This article presents a new approach to extract formant instantaneous characteristics (FIC) parameters for speaker identification (SI). On the one hand, FIC could be derived from time-frequency description of speech signal in the Hilbert-Huang Transform (HHT). HHT is a powerful tool to analyze non-stationary signal and consists of sifting procedure of empirical mode decomposition (EMD) and the Hilbert Transform (HT). The sifting procedure of EMD is to get intrinsic mode functions (IMF), so it is significant to determine all the instantaneous information from nonlinear or non-stationary signals like speech signals. This could be achieved directly through HT yet. On the other hand, a lot of information comprised in formant is not only reflection of speech contents but also speakers' individual features, so that have to get finer formant properties. Compared with traditional methods, the FIC of extracting by HHT is able to describe fine formant instantaneous information in detail. These FIC parameters are a class of reflections of speaker's individual features from both glottal wave and vocal tract. Finally, different kinds of FIC parameters were combined to MFCC to form a plurality of experimental parameters for SI based on a Gaussian mixture model (GMM). And results show that FIC parameters play a compensating role to MFCC in SI, with one of improved relative rate up to 11.96%. Experimental utterances are Chinese mandarin under clean background recording circumstances.

Keywords**:** speaker identification, formant, HHT, instantaneous frequency, MFCC

1. Introduction

Speaker identification (SI) is a research subject studying speaker individual information contained in speech signal. In the past decades, scholars have carried on a depth study in SI. One typical system model of SI is Gaussian mixture model (GMM) together with Mel-frequency cepstral coefficients (MFCC) [1] inputting. The MFCC are just a set of magnitude parameters, utilizes only the amplitude of the incoming sound waves at different frequencies, that is totally neglected the phases of the different frequencies. However, speech is a nonlinear and noncausality signal, there is independent information between its amplitude and phase spectrum.

Recently the phase of speech signal as a function of frequency conveys meaningful information that is useful for a various speech processing tasks, such as directional hearing, localization, and tracking [2], sound signal reconstruction, human listening comprehension [3]. Alsteris [4] presented about a group delay function and instantaneous frequency application to automatic speech recognition. Hegde [5] offered a way of the modified group delay feature in speech recognition. Especially, the first formant central frequency has a positive relationship with derivative of glottal area function, and its bandwidth has a positive relationship with glottal area [6]. Glottal changes are closely associated with speaker individual features. Therefore, more accurate formant frequency, we will find the more information for SI seeking individual performances.

Meanwhile, for the wrapping of phase spectrum at multiples of 2π, any meaningful use of phase spectrum for speech involves the unique process of phase unwrapping, while Hilbert Transform (HT) has being widely utilized to extract instantaneous phase and frequency [7]. In the AM-FM theory, instantaneous frequency could be considered as modulated frequency of an AM-FM signal. According to air-acoustic model, speech signal could be thought as summation of a group of AM-FM signals. There are specific explanations for formant and harmonic frequencies in [8]. Features based on AM-FM theory of speech signals and instantaneous frequencies and obtained through a group of filters distributed in the whole frequency domain, play a more effective performance than conventional MFCC for SI [9]. That work shows the good efficacy of phase features.

Hilbert-Huang Transform (HHT), a new type of time-frequency analysis, has been proposed for nonlinear and non-stationary signals [10]. The studies of HHT have showed that it is a favorable technology to acquire signals' instantaneous information. It has being proposed to use in some other fields, such as analysis of vibration signals [11], analysis of spectral properties of short genes [12], nonlinear response of a cracked rotor [13], ocean wave signals [10], analysis for distorted power quality signals [14], and speech signal processing [15].

Traditional formant frequency estimation methods are based on spectral analysis, like linear prediction (LP) analysis algorithm. In order to improve robustness of formant frequency extraction, researches have investigated into other schemes, using such as time-varying adaptive filter [16], formant energy detector [17], gender-voicing combined detector, and even with plenty of overlapping wideband filters [18] which showed a method for fine structure spectrogram. And HHT could also be applied in formant extraction studies in some way [19].

This paper offers proposals for the formant instantaneous characteristics (FIC), that is, to extract three main formants' fine features via HT or HHT. Hence, combined parameters of FIC and MFCC in series are the right ones in SI experiments. Here GMM is considered as a basic pattern recognition technique to do SI. Experimental results show that FIC's additions are actually helpful for the improvement of identification accuracy. Slight formant structure's fluctuation really plays a compensatory function for SI.

The remainder of this article is organized as follows. In section II, basic theories of HT and HHT are summarized. FIC's extraction will be explained in section III. Section IV is the experiment description, and a total conclusion is in section V.

2. Hilbert Transform and Hilbert-Huang Transform

2.1 Hilbert Transform

HT is one of the important theoretical tools in the filed of signal analysis and processing [7]. Supposing there is a real signal $s(t) = a(t)\cos\phi(t)$, its analytic form $z(t)$ can be obtained from

$$z(t) = s(t) + j\hat{s}(t) \quad (1)$$

where $\hat{s}(t)$ represents the HT output of $s(t)$. Because the output data are complex numbers, the above equation can be rewritten as (2) in another expression.

$$z(t) = a(t)e^{j\phi(t)} \quad (2)$$

where $a(t)$ and $\phi(t)$ are the instantaneous amplitude and phase of $s(t)$, respectively. Further, one of the useful descriptions of $\phi(t)$ is its first-order time derivative, called instantaneous frequency and marked as $f(t)$ in (3).

$$f(t) = \frac{1}{2\pi}\cdot\frac{d\phi(t)}{dt} \quad (3)$$

However, HT cannot be applied alone for nonlinear and non-stationary signals though it is powerful and convenient to do. In order to make the outputs of HT have more concise physical meanings, it is best to choose narrowband signal as the input signal of HT. Then the interaction of $a(t)$, $\phi(t)$ and $f(t)$ is able to accurately express time-varying properties of $s(t)$ from different angles. Generally speaking, narrative-band filters are necessary for a multi-component signal before HT computing.

2.2 Hilbert-Huang Transform

Hilbert-Huang Transform (HHT) [10] is a new type of analysis method in time-frequency domain. HHT has better performances while analyze nonlinear and non-stationary signals. Its implementation process could be briefly explained as follows. Firstly, HHT is derived from the principles of empirical mode decomposition (EMD) and the next is utilization of HT (just as previously). There are two steps.

The purpose of EMD is to decompose original signal into a superposition of a set of nearly mono-component signals, referred as intrinsic mode functions (IMF). This is based on the idea that most of signals in nature are multi-component ones except a small mono-component section. In most cases, these IMF signals could meet those required conditions of using HT, because they can easily render narrow-band properties. Then after EMD step, with the help of HT, outputs such as $a(t)$, $\phi(t)$, $f(t)$ and other instantaneous expressions theoretically could perform fine time-varying features of IMF signals. This behavior has more significant performances to describe time-frequency changes than to apply HT computing to just a general band-pass filtered signal.

The necessary conditions required by IMF signals can be summarized as these: (a) Over the entire time series the number of extrma (maxima plus minima) and the number of zero-crossings differ by, at most, one, i.e., an essentially oscillatory process. (b) At any point the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. With these conditions, an IMF signal is not being a narrow-band signal in the strict sense. It is actually a modulated function which could often have a finite bandwidth, such as an amplitude-modulated (AM) one [14], a frequency-modulated (FM) one or an AM-FM one [20].

The basic EMD algorithm adopted to extract IMF signals essentially consists of a three-step processing named sifting [10]. The standard EMD process can be described as the following three parts (or steps):

First, set a given real signal $x(t)$ as the input signal of sifting process.

Second, apply sifting process to get the first level IMF. How to do this? There are several details. (*a*) Identify the successive local minima and the local maxima of $x(t)$, interpolate the local maxima and the local minima with a cubic spline so as to form an upper and a lower envelope in the whole signal span. (*b*) Compute the instantaneous mean values of the two envelopes, denoted by $m(t)$. (*c*) Determine whether the difference between $x(t)$ and $m(t)$ meets IMF signals' conditions.

$$x(t) - m(t) = c_{11}(t) \quad (4)$$

When $c_{11}(t)$ meets the conditions , $c_{11}(t)$ is just the first level IMF $c_1(t)$ decomposed from original signal, set $c_{11}(t)$= $c_1(t)$. Otherwise, make $c_{11}(t)$ as original signal and repeat the sifting steps above until find the first IMF $c_1(t)$. In order to cease the sifting process, a stop criterion is required. That is the standard deviation (S.D.) calculated from two consecutive sifting results should be limited to a threshold value as in (5):

$$S.D. = \sum_{t=0}^{t=T}\frac{\left|c_{1(k-1)}(t) - c_{1k}(t)\right|^2}{c_{1(k-1)}^2(t)} < threshold \quad (5)$$

where T is the length of signal in time domain, and k is the order number of sifting . The threshold value is usually set between 0.2 and 0.3.

And third, apply sifting process to get other levels' IMF $c_n(t)$, n=2,3,4,…N. N is the number of total IMF levels. This is another cycle process. It begins with setting the difference signal $r_1(t)$ in (6) between $x(t)$ and $c_1(t)$ as the original signal once again.

$$x(t) - c_1(t) = r_1(t) \qquad (6)$$

Repeat steps in second part to extract the second level IMF $c_2(t)$. Carry on extracting others with these rules until meet other stopped criteria, one of which is that the number of extrema of the difference signal in certain level is less than 2. Finally EMD is achieved and yields results in (7).

$$x(t) = \sum_{n=1}^{N} c_n(t) + r_N(t) \qquad (7)$$

where N is usually a finite integer, $r_N(t)$ is final residue of the signal.

Thus the original signal could be decomposed into N empirical modes and one residue. These empirical intrinsic mode functions are ready for the using of HT and for the outputs of instantaneous information. Because IMF signals are some kind of modulated ones, EMD might be thought as a filter bank. Be different from some traditional filters, EMD process is adaptive to original signal in time domain.

3. FIC's Extraction

Generally speaking, basic idea of FIC is derived from two points.

Firstly, according to the principles of speech production, there are the formants in vowels and voiced consonants. These formants come from function of vocal tract, while formants of nasals come from nasal cavity. Different shapes of vocal tract have some connection with different phonemes, but different persons pronounce different outcomes even as speaking the same phoneme. The formant features are basic responses of vocal tract, nevertheless, there is being existed the fact that they are not only relevant to speech contents, but also to special speaker traits [6]. The main durations of formants determine contents of speech basically, while the fluctuation of fine formant frequencies marks which person is speaking. Therefore, it is necessary to pay attention to fine formants' fluctuation in SI studies. FIC's extraction what is in the article is just one type of these works.

Secondly, speech signal has being studied with AM-FM model. That is speech signal could be considered as the oral and nose cavity modulated by the glottal airflow from the view of AM-FM model theory. With the thought, speech signal is the superposition of a series modulated AM-FM signals, with special AM and FM signals included. The outstanding performances are the resonance phenomena around formant frequencies. According to studies of speech signals based on AM-FM model, the formants are thoughtfully the modulation effects of vocal tract responding to glottal flow, which represents as convolution of the two. So the original speech signal is a kind of multi-component signals. It is necessary to decompose the original speech signal into some single component signals or some modulated ones. For example, when the carrier frequency is one of the formant frequencies, the frequency's variances around formant frequencies and spectrum envelope around formant frequency are the results of the frequency modulation and the amplitude modulation. What is welcome in SI studies is that all of these modulation results are just one type of reflections about speakers' individual differences.

This article takes all the modulation around formant frequency and basic speech theory into consideration. Try to extract instantaneous information around formant frequencies with the help of HT or HHT computing. Then some statistical methods could be adopted here to form suitable parameters, which are able to be used as the characteristic parameters for SI. All of these parameters could be collectively referred to as FIC parameters.

The main framework of FIC's extraction is described in Fig. 1. The procedure consists of three blocks to form FIC parameters. They are formants' position estimation, instantaneous information acquisition and statistical parameters' extraction.

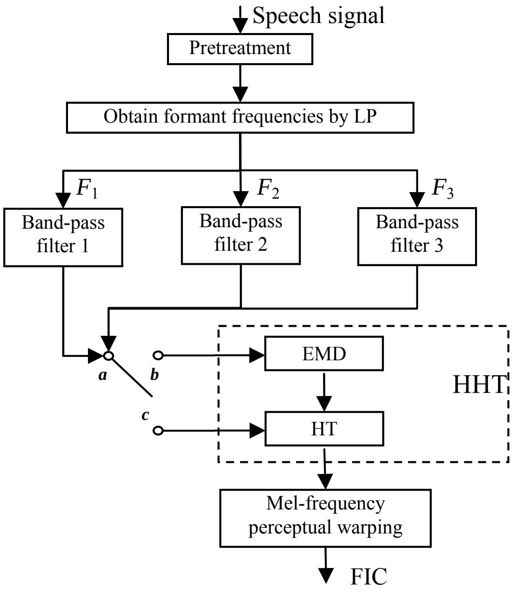


Figure 1. Flow chart of FIC's extraction based on HT or HHT

Pretreatment is the first step of specific, removing silence and dividing speech signal into short frames modified by Hamming window. By LP method, the second is to find first three gross formant frequencies (i.e. F_1, F_2, and F_3). If this step is unsuccessful for a certain frame, discard this frame. Vowels and voiced consonants occupy a major number of frames, so there are a few the discarded frames which are unvoiced.

After three band-pass filters, the flow chart reaches the node *a* in Fig.1. If the node *a* links up to the node *b*, the next step is to carry out HHT process. It is necessary to indicate that just the first IMF component was used for HT in our experiments. If the node *a* links up to the node *c*, the next step is to carry out HT process directly. Finally, instantaneous outputs through HT could be comprehensively considered to extract some statistical FIC parameters. Additional instruction is that mel frequency perceptual warping is needed after HT as experimental results showed.

In particular, it should be noticed that each filter's center frequency is one of formant frequencies, and its cutoff frequencies are corresponding to 3dB bandwidth around filter center frequency.

The results of HT or HHT are instantaneous descriptions, such as instantaneous amplitude and instantaneous frequency. Mel-frequency perceptual warping is used to transform instantaneous frequency to mel instantaneous frequency (hereinafter referred to as IF). In this regard, expressions $a_{iH}(t)$, $a_{iHH}(t)$ and $f_{iH}(t)$, $f_{iHH}(t)$, $i = 1,2,3$ indicate respectively instantaneous amplitude and instantaneous frequency after three band-pass filtered signals with HT or HHT computing. Expression $a_i(t)$ normally refers to any one of $a_{iH}(t)$ and $a_{iHH}(t)$, while expression $f_i(t)$ normally refers to any one of $f_{iH}(t)$ and $f_{iHH}(t)$.

FIC parameters considered in this article are principally mean and variance values of IF, that is the formant frequency and bandwidth measurements. So these instantaneous parameters are statistical values referring to formants. There are two primary methods to obtain FIC parameters [7] [21]. One way is directly computation as in (8) and (9) to get FIC parameters F_i and B_i, and another indirectly computation to get weighted F_{wi} and B_{wi} by instantaneous amplitude $a(t)$ as in (10) and (11).

$$F_i = \frac{1}{T}\int_{t_0}^{t_0+T} f_i(t)dt \tag{8}$$

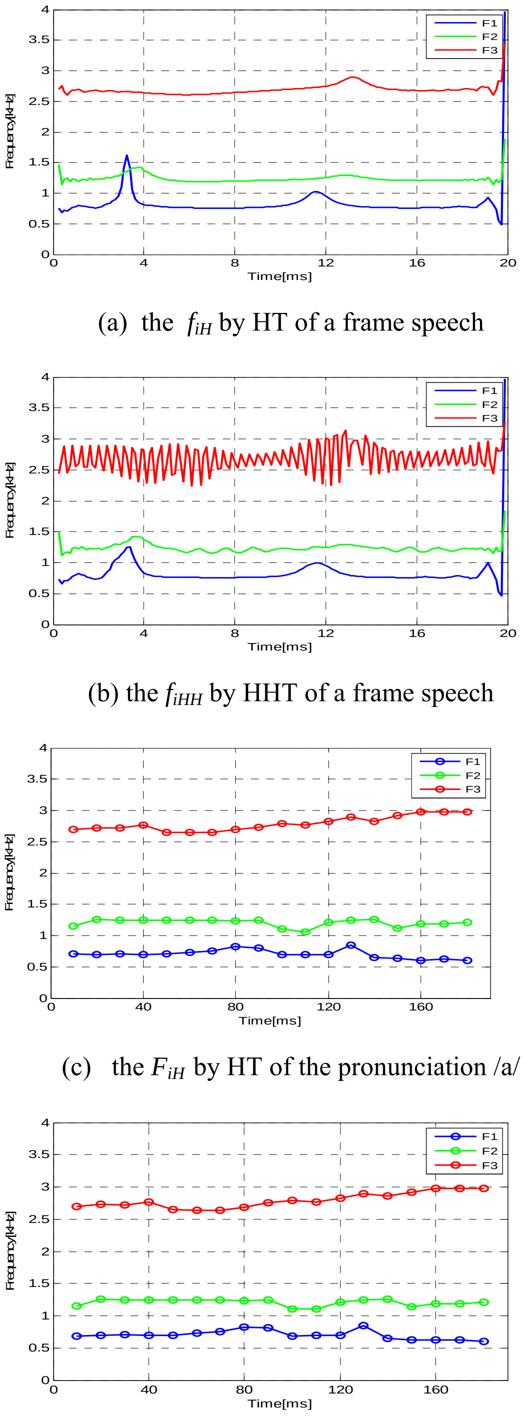
$$B_i = \frac{1}{T}\int_{t_0}^{t_0+T} \left[f_i(t) - F_i\right]^2 dt \tag{9}$$

$$F_{wi} = \frac{\int_{t_0}^{t_0+T} f_i(t)\left[a_i(t)\right]^2 dt}{\int_{t_0}^{t_0+T} \left[a_i(t)\right]^2 dt} \tag{10}$$

$$B_{wi} = \frac{\int_{t_0}^{t_0+T} \left\{\left[\frac{1}{2\pi}\cdot\frac{da_i(t)}{dt}\right]^2 + \left[f_i(t) - F_{wi}\right]^2\left[a_i(t)\right]^2\right\}dt}{\int_{t_0}^{t_0+T} \left[a_i(t)\right]^2 dt} \tag{11}$$

where t_0 and T are the start and duration of the analysis frame, respectively.

There are two groups of FIC parameters by HT or HHT in Fig.1. Expressions F_{iH}, B_{iH}, F_{wH}, B_{wH} below indicate respectively any one of band-pass filtered signals' results corresponding to the formula (8)-(11) are based on only HT computing, while F_{iHH} , B_{iHH} , F_{wHH} , B_{wHH} below indicate respectively any one of band-pass filtered signals' results from the formula (8)-(11) are based on HHT computing. That is, the node a in Fig.1 respectively links up to the node c to carry out HT process, and to the node b to carry out HHT process. For the convenience of description, results of the formula (8) and (10) will be described as mean FIC parameters calculated from either HT or HHT; results of the formula (9) and (11) will be described as bandwidth FIC parameters calculated from either HT or HHT in the same way.



(a) the f_{iH} by HT of a frame speech

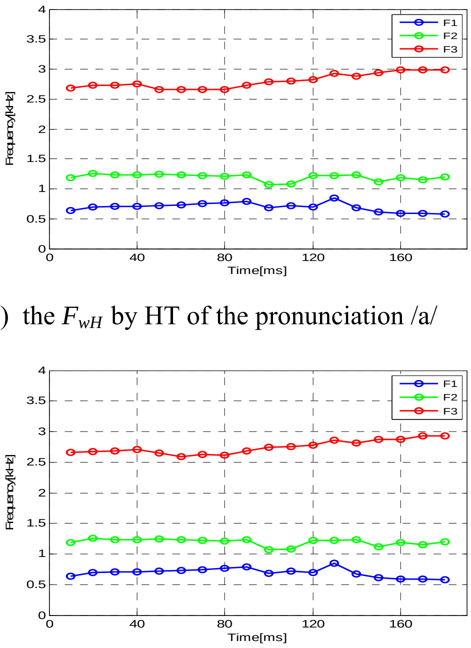
(b) the f_{iHH} by HHT of a frame speech

(c) the F_{iH} by HT of the pronunciation /a/

(d) the F_{iHH} by HHT of the pronunciation /a/

Figure 2. part of the FIC parameters by formulas(8)

There is part of results by the HT and showed in Fig.2 (a) (c) and Fig.3 (a). There is part of results by the HHT and showed in Fig.2 (b) (d) and Fig.3 (b). In the FIC's extraction processing, the instantaneous frequency f_{iHH} by the HHT has the dense frequency modulating in the third formant, but the f_{iH} by the HT has not frequency modulating property, seen in Fig.2 (a) and (b). The F_{iHH} gotten by HHT has more smoothness than the F_{iH} gotten by HT in the first and the second formants, seen in Fig.2 (c) and (d). However the F_{WH} got by HT and the F_{WHH} got by HHT have almost properties, seen in Fig.3 (a) and (b). Compared with the F_{iH} and the F_{iHH}, there is obvious difference seen in the Fig.2(c), (d) and the Fig.3 (a), (b).



(a) the F_{wH} by HT of the pronunciation /a/

(b) the F_{wHH} by HHT of the pronunciation /a/

Figure 3. part of the FIC parameters by formulas(10)

4. Experimental Evaluation for SI

4.1 Database Description

The PKU-SRSC corpus is used, which speech is 16bit quantized and 8 kHz sampled. The speech recording is through a microphone (AKG D 222 EB), speakers are students from Peking University of China, aging around 20 year-old. The language is mandarin.

There are 48 speakers' (24 male and 24 female) utterances in our experiments. Every person's training data were sentences about 50 seconds, everyone's testing data were 10 seconds with different-content.

4.2 Experiments

The GMM model was chosen to do SI experiments, with 64 Gaussian components. All the FIC parameters are evaluated in a closed-set text-independent SI task. MFCC's performance was used as a reference benchmark to the other experimental results. Here, MFCC's extraction [1] in experiments could be summarized as follows: pretreatment, weigh square magnitude using 24 Mel-bank filters, do the Discrete Cosine Transform (DCT), obtain 16 cepstral coefficients [22] after cepstral mean removal, and form 48 cepstral coefficients together with the first-order and second-order differences between frames. In each case, the number of male speakers for experiments was the same as the number of female speakers. Both training and testing experimental platforms for all parameters are the same.

To observe different behaviors of mean FIC and bandwidth FIC parameters in SI, they were concatenated with MFCC parameters respectively. Their results are showed in both Fig.4 and Fig.5. Besides, Fig.6 and Tab.1 give different parameters' average accuracies. They implicate that IF's variance can be ignored due to their relatively poor accuracies. Fig.5 shows the results of using MFCC only, MFCC combined with formant frequencies directly coming from LP, MFCC with IF's mean and the weighted respectively by HT or HHT computing, so as to clearly observe behaviors of IF's mean values obtained from different statistical ways.

4.2.1 FIC Parameters by HT

Fig.4 summarizes the accuracies of the SI classifier while varying the number of speakers. We compare seven kinds of parameters totally here. They are the typical MFCC and other six types of FIC parameters concatenated with MFCC. All of FIC parameters in this figure are from HT computing.

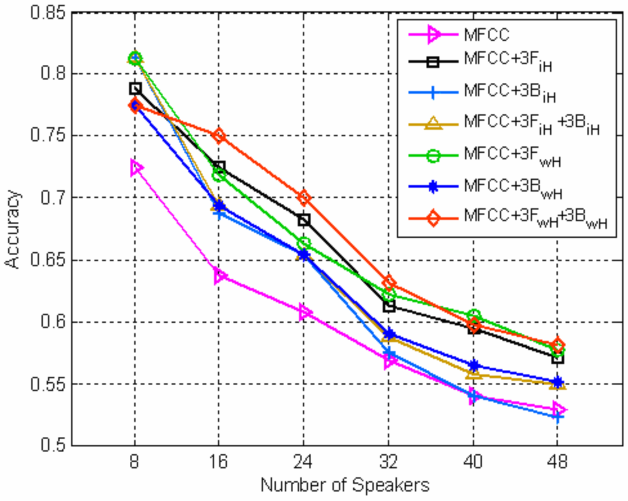


Figure 4. Results of MFCC combined with IF's mean 、 variance and the weighted ones from HT.

To some extent, Fig.4 displays a basic influence that different kinds of phase parameters (FIC) really achieved auxiliary function to MFCC in SI experiments. The better identification performance and greater stability while increasing the number of speakers are from mean FIC parameters: the black line marked with "□" are mean FIC parameters by formula(8), the green line marked with "○" are weighted mean FIC parameters by formula(10), and the red line marked with"◇" are weighted mean FIC and weighted bandwidth FIC parameters together.

There is a little function from two groups of features which are bandwidth FIC parameters by the formula (9), the light blue line marked with "+", and weighted bandwidth FIC parameters by formula (11), the blue line marked with "*".

Being relative to bandwidth FIC parameters, mean FIC parameters own better auxiliary function. And from another angle, all the combined parameters of MFCC and FIC show more or less, a performance improvement over using MFCC alone.

4.2.2 Different Mean FIC Parameters

In order to observe auxiliary roles' discrepancies of different mean FIC parameters in detail, experiments in Fig. 3 were made. The compared parameters include five types of mean FIC parameters obtained from HT or HHT or three formant frequencies directly gotten from LP estimation.

What should be attended is that we take both HT and HHT computing into account to obtain mean FIC parameters in Fig.5, while only HT is thought alone in Fig.4. This is because mean FIC parameters performs better than bandwidth ones from HT computing. HHT is the developed algorithm over HT. It digs more properties of single component signal and is a type of more fined professional tools. With a thought of instantaneous information being more significant, contrast experiments in Fig.5 were taken on.

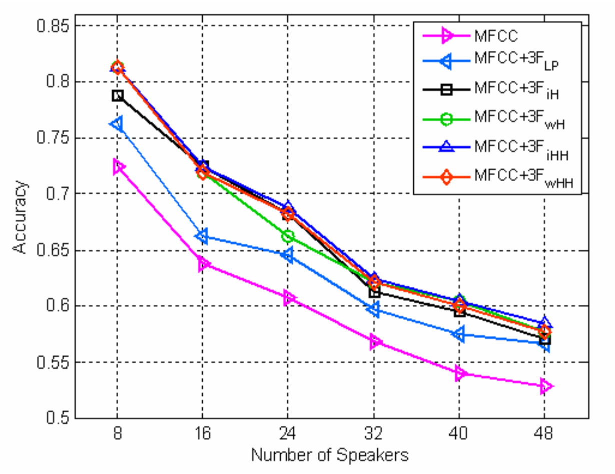


Figure 5. Results of MFCC only, and combined with IF's mean parameters from HT or HHT.

In Fig.5, it is clear to see one conclusion that is mean FIC parameters from HHT computing could reach higher recognition accuracy than FIC parameters just from HT computing in the mass. This result should not be divorced from the extra step of HHT, which is EMD step.

The results of all the FIC parameters in Fig.5 exhibit the better and steadier performance than MFCC alone with different number of testing speakers. The best type is mean FIC parameters from HHT, the red line marked with "◇", and weighted mean FIC parameters from HHT, the blue line marked with "△". The better are FIC parameters from HT marked with the green and the black line. However, the formants only decided with LPC has little compensation function, marked with light blue line "◁".

The instantaneous frequency weighted by instantaneous amplitude as the formula (10) and (11) is little function for SI system.

4.2.3 Overall Evaluation

The observations from Fig.4 and Fig.5 are concrete identification accuracy with different parameters and different number of speakers used in experiments. For any one type of parameters, we can set the average accuracy among different number of testing speakers' experiments as its overall evaluation value. Without considering the factor of testing speakers' amount, the average accuracy is a feasible reference value. These values are summarized in Fig.6.

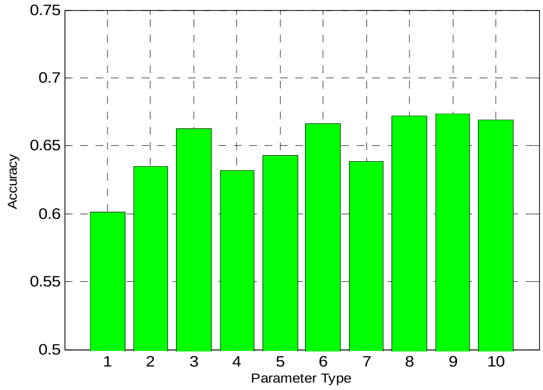


Figure 6. Average Accuracy of Different Type Parameters

Table 1. Different Type Parameters

number	Parameter Type
1	MFCC
2	MFCC + $3F_{LP}$
3	MFCC + $3F_{iH}$
4	MFCC + $3B_{iH}$
5	MFCC + $3F_{iH}$ + $3B_{iH}$
6	MFCC + $3F_{wH}$
7	MFCC + $3B_{wH}$
8	MFCC + $3F_{wH}$ + $3B_{wH}$
9	MFCC + $3F_{iHH}$
10	MFCC + $3F_{wHH}$

Tab. 1 gives the different parameters in Fig.6 as the SI system features. In SI experiments, FIC parameters play a compensatory role for MFCC. The compensatory effects change with different parameters. The relative improved average accuracy rate could be up to 11.96%. Compared with bandwidth FIC parameters, compensatory roles from mean FIC parameters are more obvious.

4.3 Analysis of Experiment Results

The overall analysis of the above experiments is here. Theoretical analysis in Section II and III has deduced the fact that FIC parameters are the function compensatory parameters to MFCC in SI studies. Next we will analyze the final conclusion from several different visual angles. Basic arguments are as follows.

- The FIC parameters are helpful for SI. Different kinds of the FIC parameters from the HT or HHT compensate for MFCC in SI system to a certain extent. When combined F_{wHH} with MFCC as the total features of the SI, the relative improved correct accuracy ratio was up to 11.96%. The fact that recognition accuracies improved indicates that phases of the different frequencies have independent information compared with the amplitudes, especially for the formant frequencies. There are the personality features of the speaker in the formant frequencies captured by FIC parameters. In other words, phase information is not just a part that could be neglected totally in speech processing and SI studies.

- Instantaneous frequencies at the formants reflect not only the oral cavity shape of pronunciation, but also synthesis of individual oral cavity. The first formant is direct proportion of the derivative of individual glottal area, and the third formant is closely interrelated with the personal, is not relevant with the phoneme. So F_{uH} and F_{uHH} parameters combined the MFCC get better performance than only the MFCC parameters, which are seen in Fig.4, Fig.5 and Fig.6.
- Instantaneous amplitude respectively computed from HT or HHT displays the same function for SI system. While it weighed to IF in formulas (10) and (11), it still brings little variances for IF in formulas (8) and (9). That is compared between the Fig.2(c) and Fig.3 (a), Fig.2 (d) and Fig.3 (b). So F_{wH} and F_{wHH} parameters get almost same performance as the F_{uH} and F_{uHH} parameters, which are seen in Fig.6.
- The explanation for this attributes to whether EMD process's existence. After band-pass filtered signals carry out EMD process, the outputs of the process are several IMF components. Contrasted to the input signal of the process, characteristics of IMF stand much closer to single component signal and have the ability to make results of HT being more meaningful. In other words, IMF signal is a type of actual expressions which are the modulation results of vocal trace and glottal wave around formant frequencies to a certain extent. So the F_{iHH} and F_{wHH} parameters gotten by the HHT have more elaborate FIC than the F_{iH} and F_{wH} parameters gotten only by the HT, shown in Fig.2 (a) and (b). The results of using HHT are better than the HT shown in Fig.5.
- Among all kinds of FIC parameters obtained from HT computing, mean FIC parameters have better performance than bandwidth FIC parameters shown in Fig.4. However, there is not stabilization in bandwidth FIC parameters for SI. The mean FIC parameters being the center frequency take better effects than the bandwidth FIC parameters.

Here, FIC parameters totally came from mel instantaneous frequency rather than instantaneous frequency. That's because the latter's performances were not as good as the former's, when without perceptual frequency wrapping.

5. Conclusions

In this work, formants' fine structures have been specially considered to the SI with the view of spectral phase information and instantaneous measurement. Using LP, band-pass filters, and HT or HHT, FIC parameters as phase features were adopted. Together with MFCC, a type of conventional spectral magnitude parameters, serial combined parameters were applied in the SI experiments. Combined information of spectral magnitude and phase, these parameters show us better performances than only MFCC. In order to look into different FIC parameters' comparable significance to reflect speaker identity, the weighted and unweighted FIC parameters are both discussed.

As a result of the AM-FM model, information at formats includes more speaker personal information. Formants' fine structures display compensating function for SI system. The HHT and HT are more efficient methods drawing phase information than the traditional LP. There is much more work still required to investigate in this article. The FIC's further expanding for all frequencies would be used to extract personal identity for SI system and to interpret deeper properties of the vocal tract modulated by glottis during human pronunciation.

Acknowledgment

This work has accomplished with the help of PKU-SRSC corpus. We would like to thank the anonymous reviews for their pertinent suggestions and criticisms that help improve this article.

Project supported by the Natural Science Foundation of China (Grant No. 60903186) and Shanghai Municipal Education Commission' Key Disciplines (Grant No. J50104)

References

[1] D. A. Reynolds, and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions Speech Audio Processing*, 3(1), pp. 72-83, Jan. 1995.

[2] C. H. Knapp and G.Carter, "The Generalized Correlation Method for Estimation of Time Delay", *IEEE Transactions Audio, Speech, and Language Processing*, ASSP-24(4), pp. 190-202, Jan. 2007.

[3] L. D. Alsteris, and K. K. Paliwal, "Iterative reconstruction of speech from short-time Fourier transform phase and magnitude spectra", *Computer Speech and Language*, 21(1), pp. 174-186, Jan. 2007.

[4] L. D. Alsteris, and K. K. Paliwal, "Short-time Phase Spectrum in Speech Processing: A Review and Some Experimental Results", *Digital Signal Processing*, 17(3), pp.578–616, 2007.

[5] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the Modified Group Delay Feature in Speech Recognition", *IEEE Transactions Audio, Speech, and Language Processing*, 15(1), pp. 190-202, Jan. 2007.

[6] M. D. Plumpe, T. F. Quatieri, and D.A.Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification", *IEEE Transactions Speech and Audio Processing,* 7(5), pp.569-586, Sept.1999.

[7] B. Boualem, "Estimating and Interpreting the Instantaneous Frequency of a Signal-Part 1: Fundamentals", *IEEE Proceeding*, 80(4), pp. 520-538, Apr. 1992.

[8] D. Dimitriadis and P. Maragos, "Continuous Energy Demodulation Methods and Application to Speech Analysis", *Speech communication*, 48(7), pp. 819-837, Jul. 2006.

[9] G. Marco, and C. Fred, "Speaker Identification Using Instantaneous Frequencies", *IEEE Transactions Audio,*

Speech, and Language Processing, 16(8), pp.1097-1111, Aug.2008.

[10] N. E. Huang, Z. Shen, S. R. Long, M. L. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung and H. H. Liu, "The Empirical Mode Decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A*, 454, pp. 903-995, 1998.

[11] J. Xun and S. Yan, "A revised Hilbert-Huang transformation based on the neural networks and its application in vibration signal analysis of a deployable structure", *Mechanical Systems and Signal Processing*, 22(7), pp. 1705-1723, Oct. 2008.

[12] R. Jiang and H. Yan, "Studies of spectral properties of short genes using the wavelet subspace Hilbert-Huang transform (WSHHT)", *Physica A: Statistical Mechanics and its Applications*, 387(16-17), pp.4223-4227, Jul. 2008.

[13] Q. Gao, C.D. Duan, H. Fan, et al., "Rotating Machine Fault Diagnosis Using Empirical Mode Decomposition", *Mechanical Systems and Signal Processing*, 22(5), pp.1072-1081, 2008.

[14] D. S. Laila, A. R. Messina and B. C. Pal, "A refined Hilbert-Huang transform with applications to interarea oscillation monitoring", *IEEE Transactions on Power Systems*, 24(2), pp. 610-620, 2009.

[15] K. Khaldi, A. Boudraa, A. Bouchikhi, and M. T. Alouane1, "Speech enhancement via EMD", *Eurasip Journal on Advances in Signal Processing*, 2008, 2008.

[16] A. Rao and R. Kumaresan, "On Decomposing Speech into Modulated Components," *IEEE Transactions Speech Audio Processing*, 8(3), pp. 240-254, May 2000.

[17] I. C. Bruce, N. V. Karkhanis, E. D. Young, and M. B. Sachs, "Robust Formant Tracking in Noise," In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (I*CASSP*)*, l, pp. 281-284, 2002.

[18] K. Mustafa and I. C. Bruce, "Robust Formant Tracking for Continuous Speech with Speaker Variability", *IEEE Transactions on Audio, Speech and Language Processing*, 14(2), pp. 435-444, Mar. 2006.

[19] H. Huang and X. Chen, "Speech Formant Frequency Estimation Based on Hilbert Huang Transform", *Journal of Zhejiang University (Engineering Science)*, 40(11), pp.1920-1930, 2006.

[20] G. Rilling, and P. Flandrin, "One or Two Frequencies? The Empirical Mode Decomposition Answers", *IEEE Transactions on Signal Processing*, 56(1), pp. 85-95, 2008.

[21] A. Potamianos, and P. Maragos, "Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Demodulation", *Journal of Acoustical Society of America*, 99(6), pp. 3795-3806, Jun. 1996.

[22] B. Zhen, X. Wu, Z. Liu, and H. Chi, "On the Importance of Components of the MFCC in Speech and Speaker Recognition", *Acta Scientiarum Naturalium Universitatis Pekinensis*, 37(3), pp. 371-378, May 2001.

Author Biographies

Limin Hou was born in Shanxi province of China in 1962. She received B.Eng. in electronics and telecommunication from Xi'an Jiaotong University, China, in 1982, M.Sc. in electronics and information engineering from the Lanzhou University, China, in 1996, and Ph.D. degree in communication and information system from Shanghai University,China, in 2005. She is currently an associate professor in the School of Communication and Information Engineering, Shanghai University, China. Her current research interests include speech processing and speaker recognition.

Juanmin Xie was born in Gansu province of China in 1986. She received the B.E. degree in economics and information engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2007, and will receive the M.E. degree in communication and information system from the Shanghai University, Shanghai, China, in 2010. She is currently studying as a master degree candidate in Shanghai University. Her research interests include speaker identification, speech recognition and other pattern recognition techniques.