Dynamic formant extraction of Wa language based on adaptive variational mode decomposition

Meijun Fu, Huazhen Dong and Wenlin Pan*

School of Mathematics & Computer Science, Yunnan Minzu University, Kunming, Yunnan, China, 650500.
Email: 779685578@qq.com, 1248571027@qq.com, panwenlin@sina.cn

Abstract. Wa language is one of Chinese minority languages spoken by the Wa nationality who lives in Yunnan Province, China. Until now, it has not been studied from the perspective of Engineering Phonetics. In this paper, for the above reason, by the adaptive variational mode decomposition (AVMD) we have investigated the dynamic formant characteristics of Wa language. Firstly, more precisely, use the synthetic dimension to split Wa language isolated words into voiceless and voiced segment, initials and finals. Secondly, use Linear Prediction Coding to estimate the first three formant frequencies and their bandwidths roughly. Thirdly, select the appropriate equilibrium constraint parameter and the number of decomposed layers so that Adaptive Variational Mode Decomposition (AVMD) can decompose the signal into some intrinsic mode functions (IMFs) without pattern aliasing. Fourthly, use the estimated formant frequencies and bandwidths to determine precisely the required IMFs. Fifthly, use the Hilbert transform to calculate the instantaneous frequency of the above determinate IMFs. Further, we implement the weight average operation on instantaneous frequencies to obtain the first three formant frequencies for each frame. Finally, comparing the first three formant frequencies obtained by the adaptive variance modal decomposition and by Praat software respectively, so we have drawn the conclusion that the relative correct rate of the former to the latter can reach 86% averagely in terms of the selected isolated words, which has shown that our method is effective on Wa language.

1. Introduction

The Wa nationality [1] is an important member of the Chinese nations. They live mainly in the southwest border of Yunnan Province, China and northern Myanmar. And the native language they use to communicate with each other is called Wa language. Wa language belongs to Va-De'ang branch, Mon-Khmer languages, and Austro-Asiatic family. Wa language can be divided into three kinds of dialects, which are "Barao dialect", "Lawa dialect" and "Wa dialect" respectively. Currently, Wa language is based on the Barao dialect and the voice of YanShuai is standard pronunciation. Until now, Wa language has not been studied by people from the perspective of Engineering Phonetics. For the above reason, we have decided to try some new things. So we try to investigate the dynamic formant frequencies on isolated words of Wa language. Here, the speech corpus used in this paper are all recorded in YanShuai.

It is well known that the glottal excitation signal at a quasi-periodic impulse into the sound channel will cause resonance characteristic, resulting in a group of resonant frequencies which are called formant frequencies or formants [2]. The formant includes two important parameters: the formant frequency and its width of bandwidth (bandwidth), which are important parameters that distinguish different voiced sounds, of course, including vowels. The formant information is lied in the envelope of the speech spectrum. Therefore, the key to the extraction of formant parameters is to estimate the
natural speech spectrum envelope, and thinking that the frequency corresponding to the maximum value of the spectral envelope is just the resonance peak. The resonant frequency can be arranged from low to high, respectively, the first formant \((F1)\), the second formant \((F2)\), the third formant \((F3)\), and the like. In general, there are about five resonant peaks that can be told in a voiced, of which the first three have a critical effect on distinguishing different voiced sounds. While the voiceless short-term spectrum has no these two characteristics, it is very similar to a random noise spectrum.

Most of the traditional speech signal processing methods are based on deterministic linear system theory [3]. And Teager et al. [4] [5] in 1980 found that in speech signals there exists a nonlinear, random and chaotic mechanism. The chaotic dynamical system converges to a certain attractor, which can be fractal modelling in the phase space [6] [7] [8]. So the chaotic fractal theory can be used to study the speech signal. The study has found that in the voice signal, consonant (without periodicity) [2] chaos is much greater than the vowel (with periodicity) [2]. Multi-fractal [9] is an infinite set of singular measures defined by multiple scale exponents on fractal, and it can be fully presented probability measure distribution ratio and unevenness of speech signal in the overall fractal structure, it therefore, can be expressed as a sum of fractal subsets with different dimensions, which improves signal geometrical characteristic and local scale. Based on the above analysis, Multi-fractal can be used to split the voiceless and voiced segment, initials and finals.

Based on the nonlinearity of speech signals [4] [5], Maragos et al. [10] [11] [12] have proposed a nonlinear speech model—-amplitude-modulated-frequency-modulated model (AM-FM model), which can be used to describe the amplitude and frequency of real-time speech formant [10]. At present, based on the AM-FM model, the main adaptive decomposition methods are: Empirical mode decomposition (EMD) [13], ensemble empirical mode decomposition (EEMD) [14], local mean decomposition (LMD) [15] and variational mode decomposition (VMD) [16]. When the signal is decomposed by EMD, there will be mode mixing and end effect [17] [18] [19] in the intrinsic mode functions (IMFs). In order to solve mode mixing, Huang proposed intermittent criterion [17] and ensemble empirical mode decomposition (EEMD) [14]. Due to that the EEMD has to add Gaussian white noise continuously, thereby increasing the amount of calculation. LMD, only to a certain extent weakens mode mixing and end effect. VMD [16], as a new signal processing method, it is assumed that the signal is composed of some IMFs which possess finite bandwidths with different center frequencies. In order to minimize the sum of the estimated bandwidth of each IMF, through the conversion to solve the variational problem, the IMFs are demodulated to the corresponding base band, and finally extract the IMFs and the corresponding center frequencies.

This paper is organized as follows. In Section 2, introduce the synthetic dimension that can better split Wa-language isolated words into voiceless and voiced segment, initials and finals. In Section 3, introduce the variational mode decomposition (VMD) that can eliminate the mode mixing, which helps us to decompose the formant frequency into the responding intrinsic mode function (IMF). In Section 4, introduce the adaptive variational mode decomposition (AVMD) on the basis of VMD. Section 5 illustrates the formant extraction algorithm based on AVMD; Section 6 includes experiment and results.

2. Synthetic Dimension

2.1. Definition of Generalized Fractal Dimension

Let \(\delta\) be the side length of a box and \(N(\delta)\) the least box number covering the set \(\Gamma\), where \(\Gamma = \{x(i)\}_{i=1}^{N}\) be a time series. And let \(P_{i}\) be the probability of falling into the \(i\) th box. Then, for any \(q \in \mathbb{R}\), define Renyi’s information entropy as follows.
\[
K_q(\delta) = \begin{cases} \\
\ln \left( \sum_{i=1}^{N(\delta)} (P_i)^q \right)^{1/(1-q)}, q \neq 1; \\
- \sum_{i=1}^{N(\delta)} P_i \ln P_i, & q = 1.
\end{cases}
\]

(1)

So we can define the following Renyi generalized dimension:

\[
D_q = - \lim_{\delta \to 0} \left( K_q(\delta)/\ln \delta \right)
\]

(2)

Especially, if \( q = 0, 1, 2 \) respectively, \( D_0, D_1, D_2 \) is called Box, Information and Correlation dimensions respectively. Concretely, they are:

1) Box dimension

\[
D_0 = - \lim_{\delta \to 0} \left( \ln N(\delta)/\ln \delta \right)
\]

(3)

2) Information dimension

\[
D_1 = - \lim_{\delta \to 0} \left( \ln \left( \sum_{i=1}^{N(\delta)} P_i \ln P_i \right)/\ln \delta \right)
\]

(4)

3) Correlation dimension

\[
D_2 = \lim_{\delta \to 0} \left( \ln \left( \sum_{i=1}^{N(\delta)} P_i^2 \right)/\ln \delta \right)
\]

(5)

2.2. Calculation of Generalized Dimension

Let \( \Delta t \) be the sampling interval, and \( \delta_j = 2^j \Delta t \) the side length of box, where \( j = 1, 2, \cdots, J \). The steps of calculation on generalized dimension are the following:

1) According to the different value \( \delta_j \), compute the total number \( N(\delta_j) = N_j \) of boxes covering a frame of speech signal and the total number \( d_{ji} \) which fall into the \( ji \) th box, where \( i = 1, 2, \cdots, N_j, j = 1, 2, \cdots, J \). As a result, \( p_{ji} = d_{ji}/L \) represents the probability of a point falling into \( ji \)-th box. Here, \( L \) is the length of a frame of speech signal, i.e., the total number of points in the set \( \Gamma \).

2) Substitute \( p_{ji} \) for \( P_i \), then we are able to obtain \( K_q(\delta_j), j = 1, 2, \cdots, J \). Construct the following Equation (6), such that it attains the minimum.

\[
\min = \sum_{j=1}^{J} \left( y_j + D_q x_j + b \right)^2
\]

(6)

Here \( x_j = \ln \delta_j, y_j = K_q(\delta_j), j = 1, 2, \cdots, J \).

3) We get easily the computational formula of \( D_q \) as follows.

\[
D_q = \left( \sum_{j=1}^{J} x_j y_j - \sum_{j=1}^{J} x_j \sum_{j=1}^{J} y_j \right) \left/ \left( \sum_{j=1}^{J} x_j^2 - \left( \sum_{j=1}^{J} x_j \right)^2 \right) \right.
\]

(7)

4) For any definite value \( q \), by the above formula we can obtain different character dimension \( D_q \).
2.3. The Definition and the Calculation of Synthetic Dimension
Box dimension $D_0$ describes the geometric scale of a set, information dimension $D_1$ gives the point distribution in a set, and correlation dimension $D_2$ reflects the relational degree amongst points in a set. Therefore, we can define synthetic dimension as follows.

For $0 \leq \alpha, \beta, \gamma \leq 1; \alpha + \beta + \gamma = 1$, define

$$S_D = \alpha D_0 + \beta D_1 + \gamma D_2$$

(8)

The above synthetic dimension possesses are better ability to identify voiceless and voiced segment, initials and finals than single fractal dimension $D_0, D_1$ and $D_2$. We finally obtained a set of optimal solutions by a large number of experiments, $\alpha = 0.8$, $\beta = 0.1$ and $\gamma = 0.1$.

3. Variational Mode Decomposition [16]

3.1. The Definition of IMF [16]
We define an intrinsic mode function (IMF) as a frequency-modulated-amplitude-modulated signal (FM-AM signal), written as:

$$u_k(t) = A_k(t) \cos(\varphi_k(t))$$

(9)

Here, $A_k(t)$ and $\varphi_k(t)$ represent respectively the instantaneous amplitude and phase of $u_k(t)$, and satisfying

$$A_k(t) \geq 0; \quad \omega_k(t) = d\varphi_k(t)/dt \geq 0$$

(10)

Further, $A_k(t)$ and $\omega_k(t)$ vary much slower than the phase $\varphi_k(t)$, that is to say, in the long interval $[t - \delta, t + \delta]$ (where $\delta = 2\pi/\omega_k(t)$), $u_k(t)$ can be seen as a pure harmonic signal with amplitude $A_k(t)$ and instantaneous frequency $\omega_k(t)$.

3.2. The Construction of VMD [16]
The variational mode decomposition (VMD) algorithm can be described as follows: in terms of constraints, the sum of the intrinsic mode functions (IMFs) is equal to the input signal $x(t)$, seeking $k$ IMFs $u_k(t)$ such that the sum of the estimated bandwidth of each IMF $u_k(t)$ is minimized. Then the corresponding constraint variational mode function expression is as follows.

$$\min_{\{u_k, \omega_k\}} \left\{ \sum_k \left[ \| \delta_x \left[ \left(\delta(t) + j/(\pi t)\right) * u_k(t) \right] e^{-j\omega_k t} \|_2^2 \right] \right\}, \text{ s.t. } \sum_k u_k = x(t)$$

(11)

where, $\{u_k\} = \{u_1, \ldots, u_k\}$ represents the $k$ IMF components; $\{\omega_k\} = \{\omega_1, \ldots, \omega_k\}$ represents the center frequency of each IMF component $u_k(t)$ correspondingly.

3.3. The Solution of VMD [16]
In order to obtain the optimal solution of the constrained IMFs, by the VMD transform, we convert the constraint variational problem to the unconstrained variational problem by introducing the penalty factor $\alpha$ and the Lagrange operator $\lambda(t)$, in which $\alpha$ can be guaranteed the reconstruction precision of the signal and $\lambda(t)$ can be strengthened constraints. The extended Lagrange expression is as follows.
\[ L \{ \{ u_k \}, \{ \omega_k \}, \lambda \} = a \sum_k \left \{ \| \sum_j \left [ d_j (t) + j/(\pi t) \right ] u_k (t) \right \|_2^2 + \left \| x(t) - \sum_k u_k (t) \right \|_2^2 + \left \langle \lambda (t), x(t) - \sum_k u_k (t) \right \rangle \right \} \]

(12)

where, \( a \) is the penalty factor (also called the equilibrium constraint parameter) and \( \lambda \) is the Lagrange multiplier. In the VMD, Alternate Direction Method of Multipliers (ADMM) is used to solve the above variational problem, and by alternately updating the parameters \( u_k^{n+1} \), \( \omega_k^{n+1} \), and \( \lambda^{n+1} \) to seek to extend the "saddle point" of the Lagrange expression. The VMD calculation step can be taken as follows.

1) Initialize the parameters \( \{ u_k^1 \}, \{ \omega_k^1 \}, \) and \( \{ \lambda_k^1 \} \), and set \( n = 0 \).
2) Set \( n = n + 1 \), and execute the entire loop.
3) Update the parameters \( u_k \) and \( \omega_k \), according to Formulas (13) and (14).

\[ \hat{u}_k^{n+1} = \arg \min_{u_k} \int_0^\infty (\omega - \omega_k)^2 | \hat{u}_k (\omega) |^2 d\omega \]

(13)

\[ \omega_k^{n+1} = \int_0^\infty \omega | \hat{u}_k (\omega) |^2 d\omega / \int_0^\infty | \hat{u}_k (\omega) |^2 d\omega \]

(14)

4) Set \( k = k + 1 \), repeat Step 3 until \( k = K \).
5) According to Formula \( \lambda^{n+1} = \lambda^n + \tau \left ( f - \sum u_k^{n+1} \right ) \), update \( \lambda \), where \( \tau \) is a noise margin parameter and here take \( \tau = 0 \).
6) Give any accuracy \( \varepsilon > 0 \), if the iteration stop condition that
\[ \sum \| u_k^{n+1} - u_k^n \|_2^2 / \sum \| u_k^n \|_2^2 < \varepsilon \]

(15)

is satisfied, then stop the entire cycle, output the results and get \( K \) narrowband IMF components \( u_k \). Otherwise, repeat Step2~Step5.

3.4. Simulation Example

Let the simulation signal \( x(t) = x_1 (t) + x_2 (t) + x_3 (t) \), where \( x_1 (t) = \cos (4 \pi t) \), \( x_2 (t) = 1/4 \cos (48 \pi t) \) and \( x_3 (t) = 1/16 \cos (576 \pi t) \). \( x(t) \) is decomposed by VMD, 3 IMFs with \( u1 \), \( u2 \) and \( u3 \), which respectively correspond to the input signals \( x1 \), \( x2 \) and \( x3 \). The simulation input signals and IMFs are shown in Figure 1.
Comparing $u_1$, $u_2$, $u_3$ and $u (u = u_1 + u_2 + u_3)$ with $x_1$, $x_2$, $x_3$ and $x$ respectively, we can find that they are not much difference in whether the amplitude or the frequency, which are shown in Figure 2.

Figure 1. Simulation Signal and VMD Decomposition.

Figure 2. The mode contrast figure of the simulation signal and signals decomposed by VMD.
It can also be seen from the spectrum distribution of input signal $x$ (See Figure 3) and 3 IMFs $u_1$, $u_2$, $u_3$ (See Figure 4), which can be well matched with the original signal in the frequency domain, and their center frequencies are shown in Figure 5.

**Figure 3.** The frequency spectrum distribution of the input signal.

**Figure 4.** The Mode Signal Spectrum by VMD.
4. Adaptive Variational Mode Decomposition

Based on VMD, we decompose the simulation signal into different layers ($K = 2, 3, 4, 5$). As a result, we have found that the decomposition accuracy is the highest if $K = 3$, that the signal is under-decomposed if $K < 3$, and that the signal is over-decomposed if $K > 3$. Under different decomposition layers $K$, the center frequencies of the above simulation signal are shown in Figure 6 respectively.

**Figure 5.** The Change of Center Frequency.

**Figure 6.** The change of center frequencies under different decomposition layers $K$

In addition, the equilibrium constraint parameter $a$ also has an influence on the signal decomposition result. The smaller the value of $a$, the larger the bandwidth of each IMF, and the result is prone to center frequency overlap and mode aliasing phenomenon; the larger the value of $a$, the smaller the bandwidth of each IMF, and the result is that the center frequency overlap and the aliasing
phenomenon disappears. Based on the above analysis, we proposed an Adaptive Variational Mode Decomposition (AVMD) method. Firstly, concretely speaking, preset value $a$ according to the signal sampling frequency. Secondly, decompose the signal by VMD, and calculate the correlation coefficients $\rho$ between each IMF and the original signal and set the threshold. Thirdly, if the minimum value of $\rho$ is less than the threshold, stop the decomposition. Otherwise, increase the number $K$ of decomposition layer, and continue to break down until the stop condition is met. The flow chart of AVMD is shown in Figure 7.

![Flowchart of AVMD](image)

**Figure 7.** The Flow chart of AVMD

5. Formant Extraction Algorithm Based on AVMD

In this paper, we propose a formant detection algorithm based on the above AVMD. The algorithm uses the AVMD to decompose the formants into different IMFs, and then uses the Linear Predictive Coding (LPC) to extract the first three formants and its center frequencies of each frame. The flow chart of formant detection algorithm based on AVMD is shown in Figure 8.

![Flowchart of Formant Extraction Algorithm Based on AVMD](image)

**Figure 8.** The Flow chart of Formant Extraction Algorithm Based on AVMD.

6. Experiment and Results

In order to verify the validity to extract the formant by AVMD, we select 100 isolated words from Wa language, which were recorded in YanShuai, as mentioned before. Now we illustrate the main steps of our experiment as follow.
1) Using Synthetic Dimension to split the 100 selected Wa language isolated words into voiceless and voiced segment, initials and finals. The division result of the Wa language word "romma", meaning "Heaven" in Chinese, is shown in Figure 9.

![Figure 9. The Segmentation Result by Synthetic Dimension.](image)

2) For every isolated word, using LPC to estimate the first three formant frequencies and their bandwidths roughly.

3) Selecting the appropriate equilibrium constraint parameter $a$ (generally taking the half of sampling frequency as the value $a$ ) and the threshold of the correlation coefficient $\rho$ such that AVMD can decompose the signal into some IMFs without pattern aliasing.

4) For every isolated word, using the estimated formant frequencies and bandwidths to determine precisely the required IMFs.

5) For every isolated word, using the Hilbert transform to calculate the instantaneous frequency of the determinate IMFs. Weighted average operations on instantaneous frequencies are carried on to obtain the first three formant frequencies for each frame. The result of the Wa language word "romma", meaning "Heaven" in Chinese, is shown in Figure 10.

![Figure 10. The First Three Formant Frequencies by AVMD.](image)
6) Comparing the first three formant frequencies depending on the AVMD method and extracted by Praat software respectively, we can draw the conclusion that the relative correct rate of the former to the latter can reach 86% averagely in terms of the 100 selected isolated words, which shows that our AVMD method is effective on Wa language. The first three formant frequencies extracted by Praat software, belonging to the Wa language word "romma", meaning "Heaven" in Chinese, are shown in Figure 11.

7) Figure 11. The First Three Formant Frequencies by PRAAT software.

7. Acknowledgments
This work is financially supported by Yunnan Provincial Department of Education Science Research Fund Project (2016YJS078), Yunnan Minzu University Overseas Master’s Program (3019901).

8. References
[1] Zhou Zhizhi, Yan Qixiang, Chen Guoqing. The wa dialect study [M]. Yunnan: Nationalities Publishing House, 2004.
[2] Kong Jiangping. Experimental phonetics basic course [M]. Peking University Press, 2015.
[3] Ritsema C J. Acoustic theory of speech production [M]. Mouton, 1970.
[4] Teager H. Some observations on oral air flow during phonation [J]. Acoustics Speech \& Signal Processing IEEE Transactions on, 1980, 28(5):599-601.
[5] Teager H M, Teager S M. Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract [M]. Speech Production and Speech Modelling. Springer Netherlands, 1990:241-261.
[6] Karonen J O, Ostergaard L, Vainio P, et al. Practical numerical algorithms for chaotic systems [M]. Practical numerical algorithms for chaotic systems. Springer Verlag, 1989:125-8.
[7] Thompson C, Mulpur A, Mehta V, et al. Transition to chaos in acoustically driven flows [J]. Journal of the Acoustical Society of America, 1991, 90(4):2097-2108.
[8] Peitgen H O, Jürgens H, Saupe D. Chaos and fractals: new frontiers of science [J]. Mathematical Gazette, 2004, 79(484):241–255.
[9] Grassberger P. Generalized dimensions of strange attractors [J]. Physics Letters A, 1983, 97(6):227-230.
[10] Maragos P, Quatieri T F, Kaiser J F. Speech nonlinearities, modulations, and energy operators[C]. International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1991:421-424 vol.1.
[11] Maragos P, Kaiser J F, Quatieri T F. On amplitude and frequency demodulation using energy operators [M]. IEEE Press, 1993.
[12] Maragos P, Kaiser J F, Quatieri T F. Energy separation in signal modulations with application to speech analysis [J]. IEEE Transactions on Signal Processing, 1993, 41(10):3024-3051.
[13] N. E. Huang, Zheng Shen, Steven R. Long … et. al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time seri [M]. The Royal Society, 1998.
[14] Zhaohua W U, Huang N. E. ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD [J]. Advances in Adaptive Data Analysis, 2011, 01(01):0900004.-

[15] Smith J S. The local mean decomposition and its application to EEG perception data [J]. Journal of the Royal Society Interface, 2005, 2(5):443.

[16] Dragomiretskiy K, Zosso D. Variational Mode Decomposition [J]. IEEE Transactions on Signal Processing, 2014, 62(3):531-544.

[17] Huang N E, Shen Z, Long S R. A NEW VIEW OF NONLINEAR WATER WAVES: The Hilbert Spectrum1 [J]. Annual Review of Fluid Mechanics, 2003, 31(1):417-457.

[18] Datig M, Schlurmann T. Performance and limitations of the Hilbert–Huang transformation (HHT) with an application to irregular water waves [J]. Ocean Engineering, 2004, 31(14–15):1783-1834.

[19] Rilling G, Flandrin P. One or Two Frequencies? The Empirical Mode Decomposition Answers [J]. IEEE Transactions on Signal Processing, 2008, 56(1):85-95.