sensors

MDPI

Article

Time-Varying Vocal Folds Vibration Detection Using a 24 GHz Portable Auditory Radar

Hong Hong [1,†], Heng Zhao [1,†], Zhengyu Peng [2,†], Hui Li [1], Chen Gu [1], Changzhi Li [2,*] and Xiaohua Zhu [1]

1 School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; hongnju@njust.edu.cn (H.H.); soniczhao@live.com (H.Z.); lihui_njust@126.com (H.L.); gc_njust@163.com (C.G.); zxh_njust@126.com (X.Z.)

2 Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, TX 79409, USA; zhengyu.peng@ttu.edu

* Correspondence: changzhi.li@ttu.edu; Tel.: +1-806-834-8682; Fax: +1-806-742-1245

† These authors contributed equally to this work.

Academic Editor: Vittorio M. N. Passaro

Received: 12 June 2016; Accepted: 25 July 2016; Published: 28 July 2016

Abstract: Time-varying vocal folds vibration information is of crucial importance in speech processing, and the traditional devices to acquire speech signals are easily smeared by the high background noise and voice interference. In this paper, we present a non-acoustic way to capture the human vocal folds vibration using a 24-GHz portable auditory radar. Since the vocal folds vibration only reaches several millimeters, the high operating frequency and the 4×4 array antennas are applied to achieve the high sensitivity. The Variational Mode Decomposition (VMD) based algorithm is proposed to decompose the radar-detected auditory signal into a sequence of intrinsic modes firstly, and then, extract the time-varying vocal folds vibration frequency from the corresponding mode. Feasibility demonstration, evaluation, and comparison are conducted with tonal and non-tonal languages, and the low relative errors show a high consistency between the radar-detected auditory time-varying vocal folds vibration and acoustic fundamental frequency, except that the auditory radar significantly improves the frequency-resolving power.

Keywords: non-acoustic; vocal folds; fundamental frequency; VMD; auditory radar

1. Introduction

Speech, human's most used means of communication, has been the object of intense study for more than 150 years. Much significant progress has been made in speech signal processing, such as speech synthesis, speech recognition, speech enhancement, speech coding, speaker identification, etc. [1]. As we know, the microphone is the most common device to record the speech signals. However, the recorded speech signals are easily smeared by the high background noise and voice interference, which will considerably degrade the quality of the recorded signals. Since the speech signal and noise always have the same frequency band, it becomes very difficult to separate speech signals from high background noise, which gains more and more attention [2,3].

In the past two decades, studies using non-acoustic sensors have shown that the glottal excitation and vocal folds articulator movements can be measured in real-time as an acoustic speech signal is produced [4]. Relevant non-acoustic sensors could be classified into two categories: the physical instruments and microwave devices. In physiology, instruments including the electroglottography (EGG) [5,6], throat microphones [7], and bone-conduction microphones [8,9] have been proposed to detect the motion of human vocal folds. From the microwave devices, the general electromagnetic motion sensor (GEMS) attracts considerable concern [6,8,10], which can be used to measure tissue

movements during voiced speech and speech involving vocal folds vibration. With GEMS, an antenna is typically strapped on the throat at the laryngeal notch or other facial locations. However, most of the physical instruments and GEMS need to be placed on the skin or close to the mouth, which makes the user discomfort and leads to the skin irritation.

In recent years, biomedical radar technology has attracted great interest in various fields, such as medical monitoring, military applications, etc. [11–17]. In medical monitoring, this technique could not only free the subject from being planted with directly contacted sensors, but also widen the monitoring period and avoid the measurement bias because of psychological stress. In military applications, such a technique can be used to find hidden enemies behind walls, or rapidly evaluate the status of victims on the battlefield.

More recently, biomedical radar technology has been extended to detect the speech signal information. In [18], a novel millimeter microwave radar was proposed to detect speech signals. The acquired speech quality is comparable to the microphone signal. Incorporating with this radar, this group used a higher-order statistics algorithm to enhance the speech quality [19]. Moreover, they presented a 94-GHz radar in [20] and an algorithm to improve the detection quality [21]. The results show that the noise is greatly suppressed. However, these works focus on the suppression of background noise. In [22], a 925-MHz speech radar system was proposed. From the system, the speech induced vocal vibration can be detected reliably. However, the experiments just reveal the similarity of radar-detected and microphone-detected signals. In our previous work, we showed the feasibility of detecting human speech via the biomedical radar and demonstrated its advantages of noise rejection and directional discrimination [23]. However, the inherent relationship between radar-detected and microphone-detected signals needs to be further explored before it can be used widely.

Therefore, all these limitations necessitate a reliable radar system that can capture the tiny human vocal folds vibration, and an accurate signal processing algorithm that can realize high frequency resolving power to demonstrate the inherent time-varying characteristics of radar-detected signal. In this paper, a 24-GHz portable auditory radar system is designed and fabricated. The high operating frequency and the 4×4 array antennas are applied to achieve the high sensitivity. The Variational Mode Decomposition (VMD) based algorithm is proposed to decompose the radar-detected auditory signal into a sequence of intrinsic modes firstly, and then, extract the time-varying vocal folds vibration frequency from the vocal folds vibration bearing mode. The VMD algorithm is entirely non-recursive and shows attractive performance with respect to existing decomposition models [24,25]. Its model provides a solution to the decomposition problem that is theoretically well founded. The basic detection theory and the radar hardware are presented in Section 2. The VMD based algorithm is described in Section 3. Then, in Section 4, two sets of experiments on non-tonal language (English) and tonal language (Chinese) are presented. Finally, a conclusion is drawn in Section 5.

2. Auditory Radar Theory

2.1. Basic Detection Theory

The auditory radar is functioning based on the phase estimation of the signals reflected by the vibrating vocal folds. The block diagram of the auditory radar system is illustrated in Figure 1.

The auditory radar typically transmits a radio-frequency continuous wave (CW) signal as follows:

$$T(t) = \cos[2\pi f t + \Phi(t)] \tag{1}$$

where f is the carrier frequency and $\Phi(t)$ is the phase noise. If the human subject is located d_0 away from the radar with the vocal folds vibration $x(t)$, the total transmitted distance becomes $2d(t) = 2d_0 + 2x(t)$. Thus, the reflected signal captured by the radar sensor at the moment t is actually the signal transmitted at the moment $t - 2d(t)/c$, where c is the signal's propagation speed (i.e., speed

of light in free space). As a result, the reflected signal captured by the radar sensor at the moment t can be written as [26]:

$$R(t) = T(t - \frac{2d(t)}{c}) = \cos[2\pi f(t - \frac{2d(t)}{c}) + \Phi(t - \frac{2d(t)}{c})] \quad (2)$$

Substituting $d(t)$ with $d_0 + x(t)$, the received signal can be further written as:

$$R(t) = \cos[2\pi ft - \frac{4\pi d_0}{\lambda} - \frac{4\pi x(t)}{\lambda} + \Phi(t - \frac{2d_0}{c} - \frac{2x(t)}{c})] \quad (3)$$

where $\lambda = c/f$ is the wavelength. Because the vocal folds vibration $x(t)$ is much smaller than nominal detection distance of d_0, the change of the phase noise Φ is negligible. Therefore, the received signal can be finally approximated as:

$$R(t) \approx \cos[2\pi ft - \frac{4\pi d_0}{\lambda} - \frac{4\pi x(t)}{\lambda} + \Phi(t - \frac{2d_0}{c})] \quad (4)$$

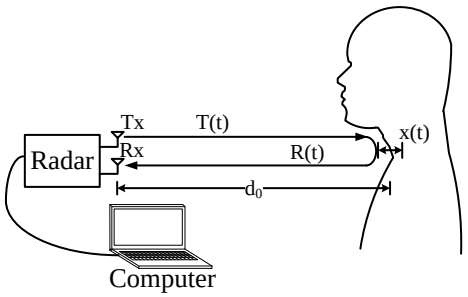


Figure 1. The basic mechanism of the auditory radar.

The received radio frequency (RF) signal is down-converted to a baseband directly by mixing with the local oscillator (LO) signal $T(t)$. In order to avoid the optimal/null point problem, the quadrature architecture is adopted in the radar [26]. As a result, the baseband quadrature signals can be written as [26]:

$$B_I(t) = \cos[\theta + \frac{4\pi x(t)}{\lambda} + \Delta\Phi(t)] \quad (5)$$

$$B_Q(t) = \sin[\theta + \frac{4\pi x(t)}{\lambda} + \Delta\Phi(t)] \quad (6)$$

where $\theta = 4\pi d_0/\lambda + \theta_0$ is the constant phase shift depending on the nominal distance to the target d_0, $\Delta\Phi(t) = \Phi(t) - \Phi(t - 2d_0/c)$ is the total residual phase noise.

As we can see, the vocal folds vibration is involved in the phase of baseband signals. In practice, when a human speaks, the vibration displacement $x(t)$ is non-sinusoidal [22]. It is well-known that this non-sinusoidal waveform contains the fundamental frequency of the speech signal, which is variable when the words or the tone changes. To extract the phase information, the complex signal demodulation (CSD) method is used to combine the quadrature channel outputs as [27]:

$$\begin{aligned} S(t) &= B_I(t) + j \cdot B_Q(t) \\ &= \cos[\theta + \frac{4\pi x(t)}{\lambda} + \Delta\Phi(t)] + j \cdot \sin[\theta + \frac{4\pi x(t)}{\lambda} + \Delta\Phi(t)] \\ &= \exp\{j[\frac{4\pi x(t)}{\lambda} + \Delta\Phi(t)]\} \end{aligned} \quad (7)$$

As demonstrated in [27], the CSD is immune from the direct current (DC) offset but can be affected by noise due to random body movement. The high operating frequency makes the radar very sensitive to random body movement.

2.2. The 24 GHz Portable Auditory Radar

Figures 2 and 3 present the block diagram and photographs of the portable auditory radar, respectively. The transmit power is 8 dBm and the DC power consumption is 1.1 W. The carrier frequency used in this work is 24 GHz, which has a μm-scale motion detection sensitivity [28]. The RF signal is divided into two channels, one is transmitted through the transmitting antenna and the other serves as the local oscillator (LO) signal in the receiver chain.

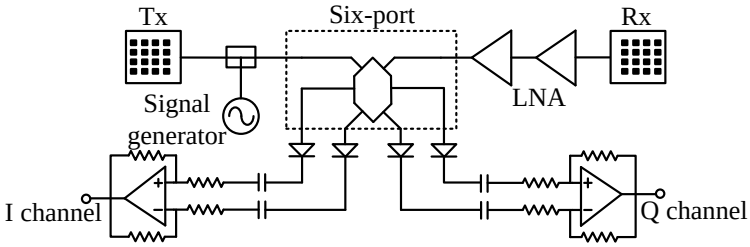
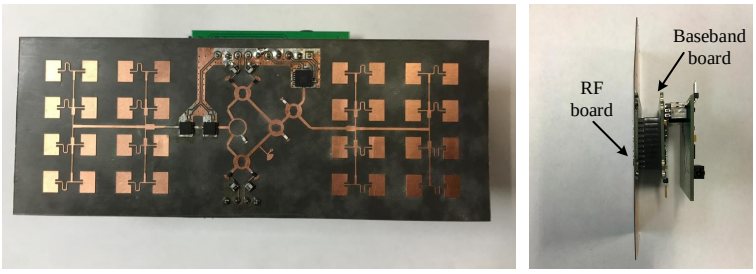


Figure 2. The block diagram of the 24-GHz auditory radar.



(a) **(b)**

Figure 3. The photographs of the 24-GHz auditory radar from (**a**) front side and (**b**) right hand side.

To enhance the directivity, a pair of 4 $\times$ 4 antenna arrays are designed, offering an antenna directivity of 19.8 dBi. As shown in Figure 3a, the antenna arrays are fabricated and integrated on a Rogers RT/duroid 5880 flexible microwave substrate (Chandler, AZ, USA) along with the RF front-end, which reduces the total device size to 11.9 cm $\times$ 4.4 cm.

In the receiver chain, the received signal is first amplified by two-stage low noise amplifiers (LNAs). Compared with the existing integrated mixer chips at 24 GHz, the six-port structure is simpler and cheaper. The outputs of the six-port downconverter are differential quadrature signals, which are amplified by two differential amplifiers to generate the baseband *I*/*Q* signals. The received RF gain and baseband gain are 34 dB and 26 dB, respectively. The baseband signals are fed to a 3.5 mm audio jack, which can be easily connected to the audio interface of a laptop or a smart phone for real-time signal processing.

3. Algorithm and Its Implementation

In this section, we start by merging the basic theory of Variational Mode Decomposition (VMD) into the time-varying vocal folds vibration detection framework [24]. Then, the details of the implementation are described. Finally, an example is presented for illustration.

3.1. VMD

The goal of the VMD is to decompose a real valued input signal f into K discrete number of modes u_k ($k = 1, 2, 3, \ldots, K$), that have specific sparsity properties, so that the original signal can be reconstructed as:

$$f(t) = \sum_k u_k \tag{8}$$

Each mode u_k is assumed to be mostly compact around a center angular frequency ω_k, which is determined along with the decomposition. The VMD algorithm to assess the bandwidth of a one dimension signal is as follows: (1) for each mode u_k, compute the associated analytic signal through Hilbert transform in order to obtain a unilateral frequency spectrum; (2) for each mode u_k, shift the mode's frequency spectrum to baseband by means of mixing with an exponential tuned to the respective estimated center frequency; and (3) estimate the bandwidth through Gaussian smoothness of the demodulated signal, i.e., the squared L^2-norm of the gradient. Then, the constrained variational problem is given as follows:

$$\min_{u_k, \omega_k} \{ \sum_k \| \partial_t [(\delta(t) + \frac{j}{\pi t}) * u_k(t)] e^{-j\omega_k t} \|_2^2 \} \tag{9}$$

where $\{u_k\} = \{u_1, \ldots, u_K\}$ and $\{\omega_k\} = \{\omega_1, \ldots, \omega_K\}$ are shorthand notations for the set of all modes and their center frequency, δ is the Dirac distribution, t is time script, k is the number of modes, and $*$ represents convolution.

In the VMD framework, the original real valued input signal f is decomposed into a set of k modes u_k each having a bandwidth in Fourier domain and compacted around a center angular frequency ω_k. The solution to the original minimization problem is the saddle point of the following augmented Lagrangian (L) expression:

$$L(u_k, \omega_k, \lambda) = \alpha \sum_k \| \partial_t [(\delta(t) + \frac{j}{\pi t}) * u_k(t)] \|_2^2 + \| f - \sum u_k \|_2^2 + \langle \lambda, f - \sum u_k \rangle \tag{10}$$

where λ is the Lagrange multiplier and $\| \bullet \|_p$ denotes the usual vector ℓ_p norm, where $p = 2$. The solution to Equation (8) are found in a sequence of k iterative sub-optimizations. Finally, the solutions for u and ω are found in Fourier domain and are given by:

$$u_n^{n+1} = (f - \sum_{i \neq k} u_i + \frac{\lambda}{2}) \frac{1}{1 + 2\alpha(\omega - \omega_k)^2} \tag{11}$$

$$\omega_n^{n+1} = \frac{\int_0^\infty \omega |u_k(\omega)|^2 d\omega}{\int_0^\infty |u_k(\omega)|^2 d\omega} \tag{12}$$

where α is known as the balancing parameter of the data-fidelity constraint, and n is the number of iterations. More details could be found in [24]. We therefore are provided with an opportunity to analyze the received signal $R(t)$ from few individual modes that are remarkably simpler than the original $R(t)$.

3.2. Implementation

The flowchart of the proposed signal processing algorithm is shown in Figure 4. The details of the implementation of the algorithm are described as follows.

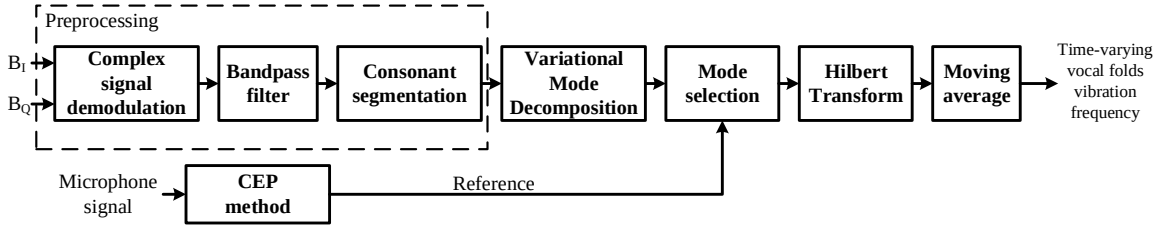


Figure 4. The flowchart of the signal processing algorithm.

3.2.1. Preprocessing

The complex signal demodulation method is used to combine the quadrature channel outputs B_I and B_Q. After complex signal demodulation, the demodulated signal $S(t)$ is filtered by a bandpass filter (50–1500 Hz) to reduce the magnitude of both the DC component and the random body movement. Then, the filtered signal $s(t)$ is preprocessed by consonant segmentation so as to remove those segments that are identified as being silent or unvoiced. Here, we adopt the short-term energy techniques to realize the segmentation [1]. In the meantime, applying the cepstrum (CEP) method to the microphone-detected signal yields a set of discrete values of the vocal folds vibration frequency [29–31]. Let f_v, $(v = 1, 2, 3, \ldots, V)$ denote those of the non-vanishing values in the data set. Then, the mean value $\bar{f}_v$ is calculated as:

$$\bar{f}_v = V^{-1} \sum_v f_v \tag{13}$$

which is represented as the reference to locate the mode that contain the vocal folds vibration information.

3.2.2. Decomposition and Mode Selection

The preprocessed radar signal $r(t)$ is decomposed with the VMD as described above, yielding a set of modes, u_k $(k = 1, 2, 3, \ldots, K)$. In this study, the moderate bandwidth constraint α is set to 2000, the noise-tolerance τ is taken as 0, the number of modes k is arbitrarily set to 2, the center frequencies ω_k are uniformly initialized, the tolerance of the convergence criteria tol is taken as $1e-7$ without imposing any DC component.

Then, there arises a crucial issue on how to single out, among the modes, the one or the ones that contain the time-varying vocal fold vibration information. We tackle this issue by catching the time-varying trend of vocal folds vibration and using it as a reference to track the variation of vocal folds vibration. This reference is evaluated by means of the CEP method, which was mentioned before. Finally, one of the decomposed modes whose average frequency is the closest to the reference $\bar{f}_v$ is picked up for the vocal folds vibration bearing mode, which can be described as:

$$\underset{k}{\operatorname{argmin}} \, |\bar{f}_v - \max[\mathcal{F}(u_k)]| \tag{14}$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform.

3.2.3. Calculation of Time-Varying Vocal Folds Vibration Frequency

The time-varying frequency, $\hat{f}(t)$, is calculated via Hilbert Transform [32]. To eliminate the possible random fluctuation induced by noise and computational errors, these time-varying frequency curves are smoothed by a simple averaging after windowing:

$$f_i(t) = \frac{1}{T} \int_{-T/2}^{+T/2} \hat{f}(\tau) W(\tau - t) \mathrm{d}\tau \tag{15}$$

where $W(t)$ is a rectangular window function of width T and height 1, with T set to be three times of the dominant oscillatory period of the selected mode.

3.3. Illustration

An illustration of the operation is shown in Figure 5. Here, we take the decomposition of English character "A" as an example. Figure 5a shows the demodulated phase information after consonant segmentation of the English character "A", and Figure 5b,c display the two modes decomposed by means of VMD. Figure 5b shows that only the first mode contains the time-varying vocal folds vibration information, while Figure 5c is the noise. Then, the time-varying frequency is captured successfully as shown in Figure 5d.

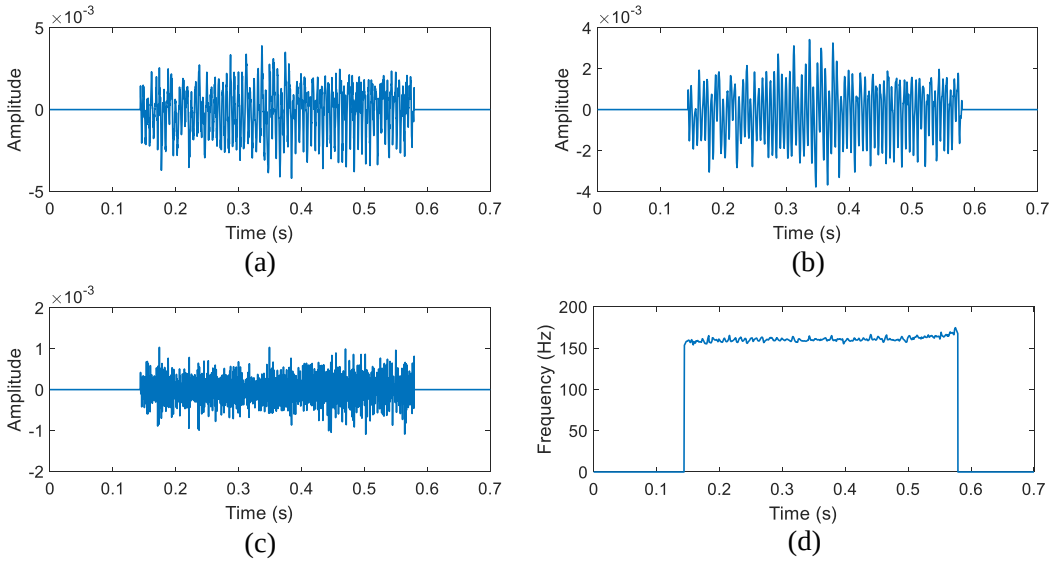


Figure 5. The results from the Variational Mode Decomposition (VMD)-based algorithm of a single character "A". (**a**) The phase information after segmentation; (**b**) The first mode decomposed by the VMD; (**c**) The second mode decomposed from the VMD; and (**d**) the time-varying frequency of the signal in (**a**).

4. Experiments

In this section, two sets of experiments are carried out. In the experiment, two healthy subjects (Subject A: male, 26-year-old; Subject B: male, 27-year-old) were asked to each sit on a chair and read the required characters, words or phrases. The auditory radar was placed 40 cm away from the human subject with its array antennas facing the subject's throat. For comparison, a smart phone was placed 40 cm away to record the acoustic signal simultaneously.

4.1. Non-Tonal Language

English, which is known as a typical non-tonal language, was tested firstly. In the first set of experiments, the subject read some English characters, words and a sentence in standard speed and intonation. The experiments were conducted in a quiet room to reduce the acoustic interference. The radar-detected time-domain waveform after filtering and segmentation is shown in Figure 5a, along with the decomposed modes in Figure 5b,c. As we analyzed before, the vibration-related periodic oscillation only exists in the first mode, while the second mode is noise. The hand-labeled method is used to extract the microphone-detected fundamental frequency for comparison, which is known as one of the most accurate methods in speech signal processing [29]. First, we manually find out each peak or valley in the time-domain speech signal. Then, the discrete points are sorted as a series t_i ($i = 1, 2, 3, \ldots$,N). As a result, the discrete fundamental frequency can be found as:

$$f(t = \frac{t_i + t_{i+1}}{2}) = \frac{1}{\triangle t} = \frac{1}{t_{i+1} - t_i} \tag{16}$$

The time-varying result of the spoken character "A" is shown in Figure 6a. Also plotted in the figures for comparison are those hand-labeled fundamental frequency values of microphone-detected speech signals. From this figure, we can see the radar-detected vocal folds vibration frequency is located around 170 Hz, and the trend of time-varying envelop fits the hand-labeled acoustic fundamental frequency values well, except that the auditory radar greatly improves the frequency-resolving power. Figure 6b presents the comparative result of a word "hello". Similarly, the radar-detected frequency closely matches the microphone-detected one. Moreover, a rising and a falling can be observed in the time-varying frequency. It indicates the frequency variations of the two different vowels in this word.

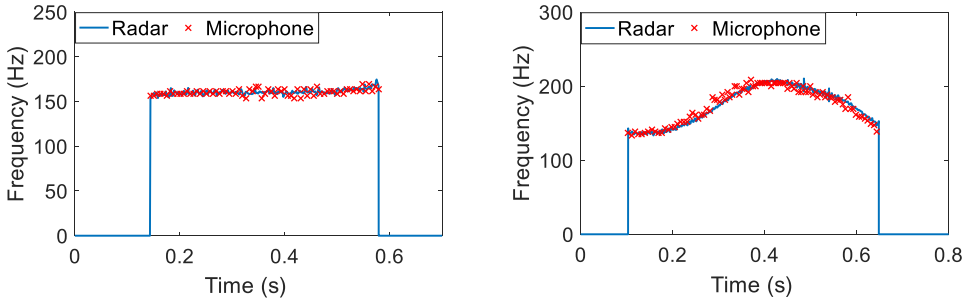


Figure 6. The auditory radar-detected time-varying vocal folds vibration frequency of one English character and one English word: (**a**) "A" and (**b**) "hello". The "×" symbols represent the hand-labeled acoustic fundamental frequency values.

To further demonstrate the effectiveness of the auditory radar, intensive tests were performed for seven characters and two words. Here, we define the deviation degree of the hand-labeled fundamental frequency values as relative error:

$$e = \frac{1}{N} \sum_{1}^{N} \frac{\{f_r(t = t_n) - f_v[n]\}}{f_v[n]} * 100\% \tag{17}$$

where $f_r(t)$ means the time-varying vocal folds vibration frequency at the moment t_n, $n = 1, 2, 3.....N$. Table 1 summarizes the relative errors of seven characters and two words. From this table, we find that the relative errors of the character and word are below 10%. It is well-known that there are more consonants in a word than a single character. When the subject pronounces the consonants, the vocal folds vibration becomes disordered and non-vibratory, which may be smeared or difficult to capture by the radar. The low relative errors show a high consistency between the radar-detected vibration and acoustic fundamental frequency. In addition, the durations of these English characters and words are given in the table to illustrate the difference between characters and words.

Table 1. The relative errors of the tested English characters and words. ("A","B","C","D","F","hello" and "boy" are from Subject A. "E" and "O" are from Subject B.)

Character/Word	A	B	C	D	E	F	O	Hello	Boy
Duration/s	0.44	1.07	0.41	0.66	0.41	0.08	0.29	0.55	0.81
Relative error	3.23%	4.45%	1.47%	4.09%	6.22%	3.89%	2.49%	5.46%	9.88%

Finally, we show a more general example of detecting the time-varying vocal folds vibration of an English sentence "I am a boy". The results are illustrated in Figure 7. After the segmentation, four

separated segments can be found, which are related to the four words in the sentence. We can observe the intonation variation in each word. In addition, the time-varying vocal folds vibration captured by the presented method agrees very well with the hand-labeled acoustic fundamental frequency track, whether for English characters "I" and "a" or two words "am" and "boy".

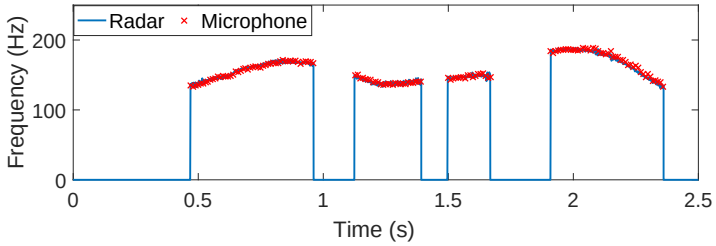


Figure 7. The auditory radar-detected time-varying vocal folds vibration frequency of the continuous speech signal, "I am a boy". The "×" symbols represent the hand-labeled acoustic fundamental frequency values.

4.2. Tonal Language

Reliable detection of the time-varying vocal folds vibration frequency is of crucial importance in speech processing of tonal languages as Chinese. In the second set of experiments, we tested some Mandarin Chinese monosyllable, multisyllabic word and continuous speech in standard speed and intonation. The experiments were also conducted in a quiet room to reduce the noise interference. We utilized the 24-GHz portable auditory radar to capture the vocal folds vibration of monosyllable and disyllable, and then applied the proposed method to signal processing to examine whether it could extract the time-varying vocal folds vibration information.

Figure 8 presents the time-varying vocal folds vibration frequency detected using the new scheme for one Chinese monosyllable and one Chinese disyllable: (a) /tiān/ in Chinese, meaning "Sky" in English and (b) /píng guǒ/ in Chinese, meaning "Apple" in English. Those hand-labeled fundamental frequency values of microphone-detected speech signals are plotted in the figures for comparison. For /tiān/, the time-varying vocal folds vibration frequency fits the hand-labeled acoustic fundamental frequency values very well as shown in Figure 8a, except that the auditory radar greatly improves the frequency-resolving power. For Chinese disyllable /píng guǒ/, the hand-labeled fundamental frequency values are distorted due to the limitation of significant digits; however, the auditory radar gives an encouraging result that correctly delineates the temporal variation of this disyllable. Note that the distortion in Figure 8b is due to the low sound intensity.

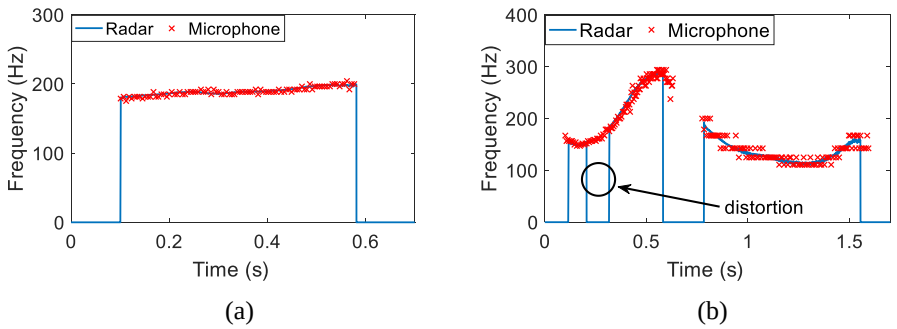


Figure 8. The auditory radar-detected time-varying vocal folds vibration frequency of one Chinese monosyllable and one Chinese disyllable: (**a**) /tiān/ and (**b**)/píng guǒ/. The "×" symbols represent the hand-labeled acoustic fundamental frequency values.

To further demonstrate the effectiveness of the auditory radar, intensive tests were performed for seven monosyllables, one disyllable and one trisyllable. Table 2 summarizes the relative errors of these syllables. From this table, it can be observed that the relative errors of the monosyllables and multisyllabic words are below 10%. The low relative errors show a high consistency between the radar-detected time-varying vocal folds vibration and acoustic fundamental frequency. In addition, the durations of these syllables are given in the table to illustrate the difference between monosyllable and multisyllabic words.

Table 2. The relative errors of Chinese monosyllables and multisyllabic words. (/nǐ/, /xíng/, /wǒ/, /wǒ/, /hǎo/, /tiān/, /píng guǒ/ and /zǎo shàng hǎo/ are from Subject A. /hāo/ and /yǒu/ are from Subject B.)

Character/Word	/nǐ/	/xíng/	/wǒ/	/hǎo/	/hāo/	/yǒu/	/tiān/	/píng guǒ/	/zǎo shàng hǎo/
Duration/s	0.43	0.448	0.188	0.415	0.21	0.17	0.48	1.44	1.66
Relative error	1.62%	2.36%	1.96%	8.31%	2.29%	6.06%	1.03%	3.20%	7.07%

Finally, we present a more general example of detecting the time-varying vocal folds vibration of a continuous Chinese speech signal, which is the sentence '/qǐng zhù yì ān quán/' in Chinese, meaning 'Attention please'. As shown in Figure 9, the present method results in the dynamic structures of time-varying vocal folds vibration, which agrees very well with the hand-labeled acoustic fundamental frequency track, whether for single word /qǐng/ or two phrases /zhù yì/ and /ān quán/.

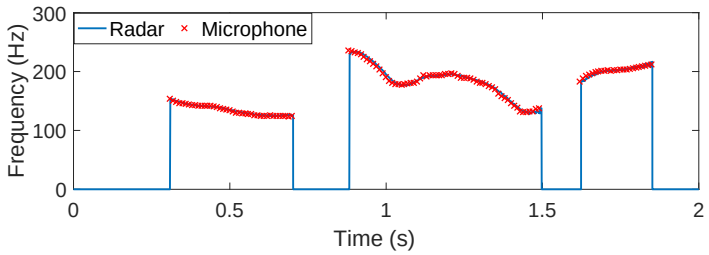


Figure 9. The auditory radar-detected time-varying vocal folds vibration frequency of the continuous speech signal, '/qǐng zhù yì ān quán/'. The "×" symbols represent the hand-labeled acoustic fundamental frequency values.

5. Conclusions

This paper has presented a non-acoustic way to capture the human vocal folds vibration using a 24-GHz portable auditory radar. The auditory radar is highly integrated with a small size and low power consumption. The VMD based algorithm is proposed to decompose the radar-detected auditory signal into a sequence of intrinsic modes. Therefore, the time-varying vocal folds vibration frequency is extracted from the corresponding mode. The low relative errors show a high consistency between the radar-detected auditory time-varying vocal folds vibration and acoustic fundamental frequency, which enable potential applications in robust speech recognition, recovery and surveillance.

Acknowledgments: This work was supported by the National Science Foundation (NSF) under Grant ECCS-1254838, the National Natural Science Foundation of China under Grant 61301022, the National Key Technology Support Program 2015BAI02B04, the Special Foundation of China Postdoctoral Science under Grant 2013T6054, and by the the Natural Science Foundation of Jiangsu Province under Grant BK20140801.

Author Contributions: Hong Hong, Heng Zhao and Changzhi Li conceived and designed the experiments; Zhengyu Peng developed the hardware system; Hui Li and Chen Gu performed the experiments; Hong Hong, Heng Zhao and Hui Li analyzed the data; Hong Hong, Heng Zhao, Changzhi Li and Xiaohua Zhu participated in the analysis of the results; and Hong Hong wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Quatieri, T.F. *Discrete-Time Speech Signal Processing Principles and Practice*; Prentice Hall: Upper Saddle River, NJ, USA, 2001.
2. Varela, O.; San-Segundo, R.; Hernandez, L.A. Robust speech detection for noisy environments. *IEEE Aerosp. Electron. Syst. Mag.* **2011**, *26*, 16–23.
3. Jain, P.; Pachori, R.B. Event-Based Method for Instantaneous Fundamental Frequency Estimation from Voiced Speech Based on Eigenvalue Decomposition of the Hankel Matrix. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1467–1482.
4. Barnes, T.; Burnett, G.; Gable, T.; Holzrichter, J.F.; Ng, L. Direct and indirect measures of speech articulator motions using low power EM sensors. In Proceedings of the XlVth International Congress of Phonetic Sciences, San Francisco, CA, USA, 1–7 Augest 1999.
5. Brady, K.; Quatieri, T.F.; Campbell, J.P.; Campbell, W.M.; Brandstein, M.; Weinstein, C.J. Multisensor MELPe using parameter substitution. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004 (ICASSP '04), Motreal, QC, Canada, 17–21 May 2004; Volume 1, pp. 77–80.
6. Holzrichter, J.F.; Ng, L.C.; Burke, G.J.; Champagne, N.J.; Kallman, J.S.; Sharpe, R.M.; Kobler, J.B.; Hillman, R.E.; Rosowski, J.J. Measurements of glottal structure dynamics. *J. Acoust. Soc. Am.* **2005**, *117*, 1373–1385.
7. Erzin, E. Improving Throat Microphone Speech Recognition by Joint Analysis of Throat and Acoustic Microphone Recordings. *IEEE Trans. Audio Speech Lang. Proc.* **2009**, *17*, 1316–1324.
8. Burnett, G.C.; Holzrichter, J.F.; Ng, L.C.; Gable, T.J. The use of glottal electromagnetic micropower sensors (GEMS) in determining a voiced excitation function. *J. Acoust. Soc. Am.* **1999**, *106*, 2183–2184.
9. Campbell, W.M.; Quatieri, T.F.; Weinstein, C.J. Multimodal speaker authentication using nonacoustic sensors. In Proceedings of the in Workshop Multimodal User Authentication, Santa Barbara, CA, USA, 11–12 December 2003; pp. 215–222.
10. Holzrichter, J.F.; Burnett, G.C.; Ng, L.C.; Lea, W.A. Speech articulator measurements using low power EM-wave sensors. *J. Acoust. Soc. Am.* **1998**, *103*, 622–625.
11. Chen, K.M.; Huang, Y.; Zhang, J.; Norman, A. Microwave life-detection systems for searching human subjects under earthquake rubble or behind barrier. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 105–114.
12. Li, C.; Cummings, J.; Lam, J.; Graves, E.; Wu, W. Radar remote monitoring of vital signs. *IEEE Microw. Mag.* **2009**, *10*, 47–56.
13. Mikhelson, I.V.; Lee, P.; Bakhtiari, S.; Elmer, T.W.; Katsaggelos, A.K.; Sahakian, A.V. Noncontact Millimeter-Wave Real-Time Detection and Tracking of Heart Rate on an Ambulatory Subject. *IEEE Trans. Inf. Technol. Biomed.* **2012**, *16*, 927–934.
14. Kim, H.J.; Kim, K.H.; Hong, Y.S.; Choi, J.J. Measurement of human heartbeat and respiration signals using phase detection radar. *Rev. Sci. Instrum.* **2007**, *78*, 104703–104703-3.
15. Zhao, H.; Hong, H.; Sun, L.; Xi, F.; Li, C.; Zhu, X. Accurate DC offset calibration of Doppler radar via non-convex optimisation. *Electron. Lett.* **2015**, *51*, 1282–1284.
16. Sun, L.; Hong, H.; Li, Y.; Gu, C.; Xi, F.; Li, C.; Zhu, X. Noncontact Vital Sign Detection based on Stepwise Atomic Norm Minimization. *IEEE Signal Process. Lett.* **2015**, *22*, 2479–2483.
17. Sun, L.; Li, Y.; Hong, H.; Xi, F.; Cai, W.; Zhu, X. Super-resolution spectral estimation in short-time non-contact vital sign measurement. *Rev. Sci. Instrum.* **2015**, *86*, 105–133.
18. Jiao, M.; Lu, G.; Jing, X.; Li, S.; Li, Y.; Wang, J. A novel radar sensor for the non-contact detection of speech signals. *Sensors* **2010**, *10*, 4622–4633.
19. Tian, Y.; Li, S.; Lv, H.; Wang, J.; Jing, X. Smart radar sensor for speech detection and enhancement. *Sens. Actuators A Phys.* **2013**, *191*, 99–104.
20. Li, S.; Tian, Y.; Lu, G.; Zhang, Y.; Lv, H.; Yu, X.; Xue, H.; Zhang, H.; Wang, J.; Jing, X. A 94-GHz Millimeter-Wave Sensor for Speech Signal Acquisition. *Sensors* **2013**, *13*, 14248.
21. Chen, F.; Li, S.; Li, C.; Liu, M.; Li, Z.; Xue, H.; Jing, X.; Wang, J. A Novel Method for Speech Acquisition and Enhancement by 94 GHz Millimeter-Wave Sensor. *Sensors* **2015**, *15*, doi: 0.3390/s16010050.
22. Lin, C.S.; Chang, S.F.; Chang, C.C.; Lin, C.C. Microwave Human Vocal Vibration Signal Detection Based on Doppler Radar Technology. *IEEE Trans. Microw. Theory Tech.* **2010**, *58*, 2299–2306.

23. Zhao, H.; Peng, Z.; Hong, H.; Zhu, X.; Li, C. A Portable 24-GHz Auditory Radar for Non-Contact Speech Sensing with Background Noise Rejection and Directional Discrimination. In Proceedings of the 2016 IEEE MTT-S International Microwave Symposium, San Francisco, CA, USA, 22–27 May 2016.
24. Dragomiretskiy, K.; Zosso, D. Variational Mode Decomposition. *IEEE Trans. Signal Proc.* **2014**, *62*, 531–544.
25. Wang, Y.; Markert, R.; Xiang, J.; Zheng, W. Research on variational mode decomposition and its application in detecting rub-impact fault of the rotor system. *Mech. Syst. Signal Process.* **2015**, *60*, 243–251.
26. Droitcour, A.D.; Boric-Lubecke, O.; Lubecke, V.M.; Lin, J.; Kovacs, G.T.A. Range correlation and I/Q performance benefits in single-chip silicon Doppler radars for noncontact cardiopulmonary monitoring. *IEEE Trans. Microw. Theory Tech.* **2004**, *52*, 838–848.
27. Li, C.; Lubecke, V.M.; Boric-Lubecke, O.; Lin, J. A Review on Recent Advances in Doppler Radar Sensors for Noncontact Healthcare Monitoring. *IEEE Trans. Microw. Theory Tech.* **2013**, *61*, 2046–2060.
28. Gu, C.; Inoue, T.; Li, C. Analysis and Experiment on the Modulation Sensitivity of Doppler Radar Vibration Measurement. *IEEE Microw. Wirel. Compon. Lett.* **2013**, *23*, 566–568.
29. Hong, H.; Zhao, Z.; Wang, X.; Tao, Z. Detection of Dynamic Structures of Speech Fundamental Frequency in Tonal Languages. *IEEE Signal Proc. Lett.* **2010**, *17*, 843–846.
30. Childers, D.G.; Skinner, D.P.; Kemerait, R.C. The cepstrum: A guide to processing. *IEEE Proc.* **1977**, *65*, 1428–1443.
31. Noll, A.M. Cepstrum Pitch Determination. *J. Acoust. Soc. Am.* **1967**, *41*, 293.
32. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. A* **1998**, *454*, 903–995.