A NEW METHOD FOR DETERMINATION OF INSTANTANEOUS PITCH FREQUENCY FROM SPEECH SIGNALS

Abhay Upadhyay, Ram Bilas Pachori

Discipline of Electrical Engineering
Indian Institute of Technology Indore
Indore, India 452017
{phd1301102002, pachori}@iiti.ac.in

ABSTRACT

This paper presents a new method for instantaneous pitch frequency determination from speech signals. The proposed method is based on the variational mode decomposition (VMD) and the Hilbert transform. The VMD is applied in iterative way with specific input parameters in order to determine the fundamental frequency component from the speech signals. The fundamental frequency component has been used for detection of voiced and non-voiced regions from speech signals. The instantaneous pitch frequency is computed using Hilbert transform of the fundamental frequency component corresponding to voiced regions of speech signals. The experimental results are shown on speech signals taken from Keele pitch extraction reference database. The experimental results obtained from the proposed method are compared with the other existing methods for determining pitch frequency from speech signals.

***Index Terms*—** Speech signal analysis, Variational mode decomposition, Voiced and non-voiced detection, Pitch frequency determination

1. INTRODUCTION

Instantaneous pitch frequency is one of the essential attribute in the area of speech signal processing. The speech signals can be divided mainly into two categories namely, voiced speech signals and non-voiced speech signals [1]. The non-voiced speech signals combine unvoiced speech signals and silent regions. The voiced speech signal can be considered as a output of the vocal-tract system with nearly periodic excitation source (impulse train) as an input [1]. The instantaneous pitch frequency can be considered as the rate of vibration of vocal folds during the generation of voiced speech signals by vocal-tract system [2]. The nature of pitch frequency is time-varying [3]. It may depend on the many factors like as, gender, emotion, language, and speaker's age. The determination of pitch frequency from speech signals has many applications in speech signal processing [4].

Several approaches for determination of pitch frequency from speech signals have been reported in the literature. Many block based methods have been proposed for pitch frequency estimation such as: short-time average magnitude difference function (AMDF) based method [5], autocorrelation based method [6], cepstrum based method [7], simplified inverse filter tracking (SIFT) based method [8], modulation model based method [9], weighted autocorrelation based method [10], a harmonic sinusoidal autocorrelation (HSAC) model based method [11], and subharmonic summation (SHS) based method [12]. In the block-based methods, voiced speech signal is segmented into parts and these parts are considered as stationary signals. Due to this way of processing, the block-based methods can not determine the variation of pitch frequency within the segment of voiced speech signal [13]. The event-based methods for determination of pitch frequency [4, 14–17] are based on glottal closure instants (GCIs) [18–21] in the voiced regions of the speech signals. In [22], the GCIs are identified using discrete energy separation algorithm (DESA) applied on the band-limited voiced speech signals in the low-frequency region. These obtained GCIs are used for determining the pitch frequency. The instantaneous methods for pitch frequency detection from speech signals include many methods like as, B-spline expansion based method [13], the Hilbert-Huang transform (HHT) based method [23], band-pass filters based method [24], ensemble empirical mode decomposition (EEMD) based method [25], and empirical wavelet transform based method [26].

We propose a new method to determine the instantaneous pitch frequency based on the variational mode decomposition (VMD) and the Hilbert transform in this paper. The voiced and non-voiced regions from speech signals have been determined using VMD based method [27]. The fundamental frequency component corresponding to voiced regions has been used for determining the instantaneous pitch frequency using the Hilbert transform. In [27], the fundamental frequency component extracted from speech signals has been applied for detection of voiced and non-voiced regions. In this work,

978-1-4673-9169-6/15/$31.00 ©2015 IEEE

the same fundamental frequency component extracted from speech signals has been explored for determination of pitch frequency in instantaneous way means at each sample point.

The organization of this paper is as follows: Section 2 provides a brief overview of VMD. In Section 3, the proposed methodology is explained for instantaneous pitch frequency detection. The description of existing pitch frequency estimation methods used for comparison is provided in Section 4. Section 5 presents the experimental results. The paper has been concluded in Section 6.

2. VARIATIONAL MODE DECOMPOSITION

The variational mode decomposition (VMD) [28] represents the input signal $p(t)$ into different modes $m_k(t)$ which satisfy the new definition of intrinsic mode functions (IMFs) proposed in [29]. The VMD is an adaptive non-recursive signal decomposition method, in which each mode is compact around the corresponding center frequency ω_k. Where, k represents the mode number. The formulation of variational optimization problem in VMD method requires following steps [28]: (1) Obtain one-sided frequency spectrum using the Hilbert transform of mode. (2) Shift each mode's frequency spectrum according to estimated center frequency using modulation property. (3) Estimate the bandwidth of each mode using the gradient of squared L^2 - norm of the demodulated signal. The constrained variational problem based on the above mentioned steps can be defined as follows [28]:

$$\min_{\{m_k\},\{\omega_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * m_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}$$
$$\text{such that} \sum_k m_k(t) = p(t) \tag{1}$$

$\{m_k\} := \{m_1, m_2, ..., m_K\}$ and $\{\omega_k\} := \{\omega_1, \omega_1, ..., \omega_K\}$ represent the modes and their corresponding center frequencies, respectively. Where, K represents the total number of modes. The constrained variational problem is converted to unconstrained optimization problem and it is solved using alternate direction method of multipliers (ADMM) [30]. The final estimation of modes in frequency-domain and their corresponding updated center frequencies can be given as follows [28]:

$$\hat{m}_k^{n+1}(\omega) = \frac{\hat{p}(\omega) - \sum_{i \neq k} \hat{m}_i(\omega) + \dfrac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \tag{2}$$

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{m}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{m}_k(\omega)|^2 d\omega} \tag{3}$$

Where, $\hat{\lambda}$ and α are the estimated Lagrangian multiplier and the balancing parameters of the data-fidelity constraint, respectively. The presence of Wiener filter expression in (2) [28] makes VMD method suitable for robust analysis of noisy signals.

3. PROPOSED METHODOLOGY FOR DETERMINATION OF INSTANTANEOUS PITCH FREQUENCY

The flow chart in Fig. 1 presents an overview of the proposed methodology for instantaneous pitch frequency estimation. Firstly, the voiced and non-voiced regions from speech signal have been extracted using VMD based method which requires extraction of fundamental frequency component based on selection of specific parameters [27].

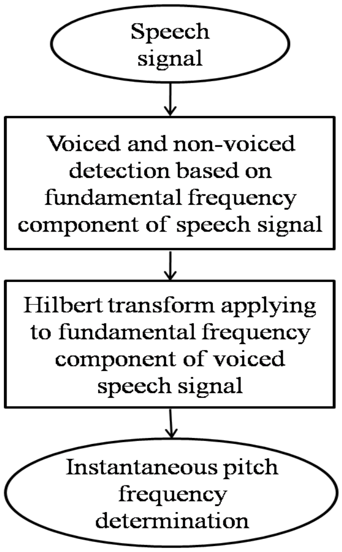
Speech signal
Voiced and non-voiced detection based on fundamental frequency component of speech signal
Hilbert transform applying to fundamental frequency component of voiced speech signal
Instantaneous pitch frequency determination

Fig. 1. The flow chart of the proposed method for estimation of instantaneous pitch frequency from speech signals.

After extracting the voiced regions of speech signals, the same fundamental frequency component corresponding to voiced speech signal denoted by $d(t)$ has been used for computing the instantaneous pitch frequency. The analytic signal of $d(t)$ has been computed by Hilbert transform. It is used for determination of instantaneous pitch frequency. The analytic signal of $d(t)$ can be expressed as follows [31]:

$$z(t) = d(t) + jd_H(t) = A(t)e^{j\phi(t)} \tag{4}$$

In (4), $d_H(t)$ is the Hilbert transform of signal $d(t)$. The amplitude envelope and instantaneous phase of analytic signal $z(t)$ denoted as $A(t)$ and $\phi(t)$, respectively and can be computed as follows [31]:

$$A(t) = \sqrt{d^2(t) + d_H^2(t)} \tag{5}$$

$$\phi(t) = \arctan\left[\frac{d_H(t)}{d(t)}\right] \quad (6)$$

The instantaneous pitch frequency of the analytic signal $z(t)$ is determined as follows:

$$\omega(t) = \frac{d\phi(t)}{dt} \quad (7)$$

The smoothing operation is required on the estimated instantaneous pitch frequencies in order to remove the discontinuities and random fluctuations. The smoothing operation on instantaneous pitch frequency is performed using a moving average filter used in [26].

4. DESCRIPTION OF EXISTING PITCH FREQUENCY ESTIMATION METHODS USED FOR COMPARISON

In order to compare the performance of the proposed method with exiting methods, we have used three existing methods namely, Pratt's autocorrelation (AC) based method [32], cross-correlation (CC) based method [33], and subharmonic summation (SHS) based method [12]. The brief information about these methods of determining pitch frequency from speech signals is as follows:

1). Pratt's autocorrelation (AC) based method [32]: In the Pratt's AC based method, the autocorrelation of the windowed voiced speech signal is divided by the autocorrelation of the window function to attenuate the windowing artifacts. A sinc interpolation is employed near to the local maxima corresponding to the pitch frequency to overcome the limitation due to the sampling rate.

2). Cross-correlation (CC) based method [33]: In the CC based method, a cross-correlation function which operates on two different data windows has been used to remove the rolling off effect in the autocorrelation values at higher lags. The autocorrelation function values at higher lags are significant for male speech signals having low pitch frequencies.

3). Subharmonic summation (SHS) based method [12]: In the SHS based method, a spectral compression model is used for the pitch frequency determination. Each spectral component introduces a series of subharmonics in the central pitch processor and the resulting sum spectrum of these subharmonic components is maximum for the pitch frequency.

The software implementations of Pratt's AC based method, CC based method, and SHS based method are available in [34].

5. EXPERIMENTAL RESULTS

The experimental study has been performed on Keele pitch extraction reference database [35] in the presence of white noise at various signal to noise ratios (SNRs) in order to evaluate the performance of the proposed method for determining the instantaneous pitch frequency. The white noise data has been obtained from NOISEX-92 database [36].

The proposed method and existing pitch frequency estimation methods which have been used for comparison are studied on speech signals available in the Keele pitch extraction reference database [35]. The Keele pitch extraction reference database contains the speech signals recorded from five male and five female speakers in the English language. The speech signals in the Keele pitch extraction reference database are sampled at rate of 20 kHz with resolution of 16 bits. This speech database also provides a reference pitch frequency value at intervals of 10 ms, which have been referred to as the ground truth. This additional information is procured from a simultaneously recorded laryngograph trace. The laryngograph signal which is available in the database has been divided into overlapped analysis frames of size 25.6 ms at a frame rate of 100 Hz and the reference pitch frequency values were computed using the autocorrelation function over each frame of the laryngograph signal. Positive reference pitch frequency values are indicated for voiced frames, negative reference pitch frequency values are provided for uncertain frames, and zero reference pitch frequency values are provided for unvoiced frames.

In order to evaluate the performance of the proposed method for pitch frequency estimation in the presence of noise, the white noise environments are obtained from the NOISEX-92 database [36]. The noise environments are resampled to 20 kHz before adding them to the speech signals. The experiment is performed at various SNRs, 20 dB, 10 dB, 5 dB, and 0 dB.

The Gross error (GE) as a performance evaluation measure [16] has been used to determine the efficacy of the proposed pitch frequency estimation method in comparison of the existing methods which have been used for comparison. The GE is defined as a percentage of voiced frames of 10 ms duration in which the estimated pitch frequency value deviates from the reference pitch frequency value by more than 20 %.

In Fig. 2, the experimental results obtained from the proposed method on a clean speech signal segment are shown together with other existing pitch frequency estimation methods namely, Pratt's AC based method, CC based method, and SHS based method. Fig. 2(a) shows the clean speech signal segment. The fundamental frequency component of this clean speech signal is shown in Fig. 2(b). The estimated instantaneous pitch frequency using proposed method has been shown in Fig. 2(c). The estimation of pitch frequencies using Pratt's AC based method, CC based method, and SHS based method have been shown in Fig. 2(d), Fig. 2(e), and Fig. 2(f), respectively. It is clear from the Figs. 2(c), 2(d), 2(e), and 2(f) that the estimated pitch frequency using proposed method closely matches with the reference pitch frequency as compared to other existing methods.

In the proposed method, the value of performance measure GE for clean speech signal segment shown in Fig. 2(a) is 1.2354 %. For the same signal, the value of GE for Pratt's AC based method, CC based method, and SHS based method are 4.6310 %, 4.1260 %, and 9.1720 %, respectively.

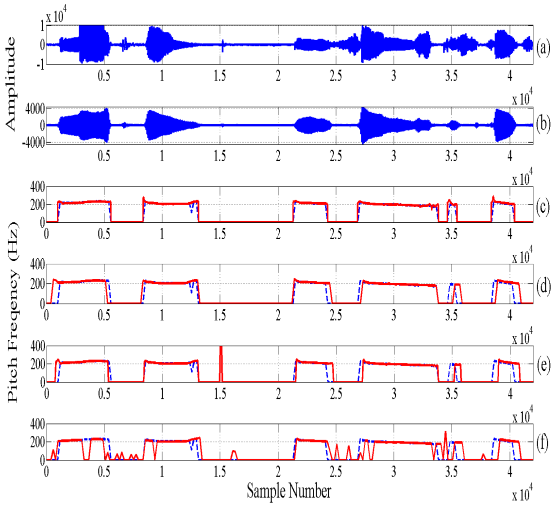

Fig. 2. (a) Clean speech signal segment (b) Fundamental frequency component of clean speech signal. Instantaneous pitch frequency contour obtained by (c) Proposed method (d) Praat's AC based method (e) CC based method (f) SHS based method. (The reference pitch frequency contour is shown by the dashed line. The (c)-(f) show the value of pitch frequency in Hz.)

Fig. 3 shows the results of estimated instantaneous pitch frequency of the speech signal segment under white noise environment at 10 dB SNR. The speech signal segment under white noise environment at 10 dB SNR and its fundamental frequency component are shown in Figs. 3(a) and 3(b), respectively. The instantaneous pitch frequency determined using proposed method is shown in Fig. 3(c). Figs. 3(d), 3(e), and 3(f) depict the estimated pitch frequencies using of Pratt's AC based method, CC based method, and SHS based method, respectively. It is clear from the Fig. 3 that the proposed method has also provided better estimation of instantaneous pitch frequency as compared with other existing methods.

Figs. 4(a) and 4(b) depict the speech signal segment under white noise environment at 0 dB SNR and its fundamental frequency component, respectively. The estimated instantaneous pitch frequency based on proposed method has been shown in Fig. 4(c). The estimated pitch frequencies based on Pratt's AC based method, CC based method, and SHS based method have been shown in Fig. 4(d), Fig. 4(e), and Fig. 4(f), respectively. In Fig. 4, the estimation of the instantaneous pitch frequency of the proposed method is also better than the existing methods at 0 dB SNR.

The proposed method has been compared with existing

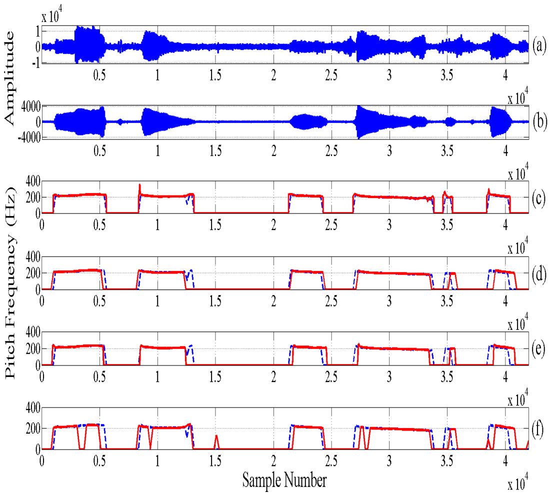

Fig. 3. (a) Speech signal segment under white noise environment at 10 dB SNR (b) Fundamental frequency component of speech signal segment under white noise environment at 10 dB SNR. Instantaneous pitch frequency contour obtained by (c) Proposed method (d) Praat's AC based method (e) CC based method (f) SHS based method. (The reference pitch frequency contour is shown by the dashed line. The (c)-(f) show the value of pitch frequency in Hz.)

pitch frequency estimation methods on two male and two female speech signals under white noise environments at SNRs, 20 dB, 10 dB, 5 dB, and 0 dB in terms of GE in percentage which is shown in Table 1. It can be observed from Table 1, the performance of proposed method is better than other existing methods under white noise environment with above mentioned SNRs. Even at 0 dB SNR, the performance of proposed method in terms of GE is better than the other existing methods.

Table 1. Gross error (GE) in percentage for different pitch frequency estimation methods at various SNRs under white noise environment.

SNR	GE in percentage			
	Proposed method	Praat's AC based method [32]	CC based method [33]	SHS based method [12]
20 dB	1.3596	5.9730	7.1260	10.5260
10 dB	2.9856	7.8210	9.2530	15.0310
5 dB	3.6580	9.4920	13.8940	18.9370
0 dB	4.1710	16.9480	21.4550	29.4940

6. CONCLUSION

A new instantaneous pitch frequency estimation method based on VMD has been proposed in this paper. The proposed

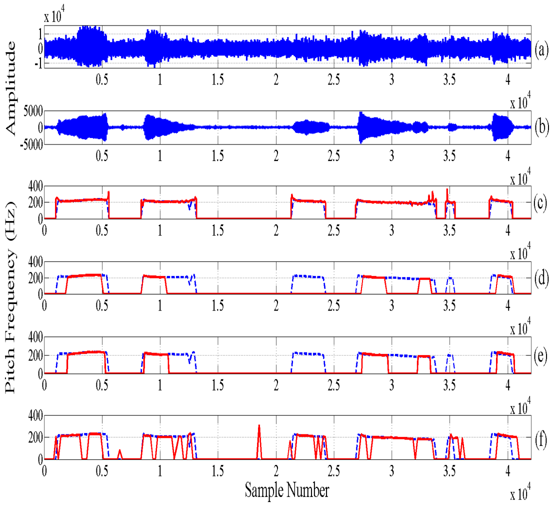

Fig. 4. (a) Speech signal segment under white noise environment at 0 dB SNR (b) Fundamental frequency component of speech signal segment under white noise environment at 0 dB SNR. Instantaneous pitch frequency contour obtained by (c) Proposed method (d) Praat's AC based method (e) CC based method (f) SHS based method. (The reference pitch frequency contour is shown by the dashed line. The (c)-(f) show the value of pitch frequency in Hz.)

method also provided better estimation of instantaneous pitch frequency in the presence of additive white Gaussian noise (AWGN). This method has been used to extract the fundamental frequency component using VMD method from speech signals, which has been used for determining instantaneous voiced and non-voiced regions. The same fundamental frequency component has been explored for determination of instantaneous pitch frequency corresponding to voiced regions of the speech signals. Moreover, the Wiener filter structure is also embedded in the VMD method, due to this feature, the proposed method for determination of instantaneous pitch frequency from speech signals is suitable for noisy environment also. The proposed method has provided better estimation of instantaneous pitch frequency even in the presence of noise as compared with other existing methods.

It would be of interest to study the proposed method in different types of noisy environments like babbel and vehicular noise with different SNRs. The proposed methodology for determination of instantaneous pitch frequency can also be studied on different standard speech signal databases.

7. REFERENCES

[1] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals*, New Delhi, India: Wiley-India, 2011.

[2] D. O'shaughnessy, *Speech Communication: Human and Machine*, Piscataway, NJ: IEEE Press, 2000.

[3] M. Markaki and Y. Stylianou, "Dimensionality reduction of modulation frequency features for speech discrimination," in *Proceedings of Interspeech, Brisbane, Australia*, 2008, pp. 646–649.

[4] P. Jain and R.B. Pachori, "Event-based method for instantaneous fundamental frequency estimation from voiced speech based on eigenvalue decomposition of the Hankel matrix," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1467–1482, Oct. 2014.

[5] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 5, pp. 353–362, Oct. 1974.

[6] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 1, pp. 24–33, Feb. 1977.

[7] A.M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," *The Journal of the Acoustical Society of America*, vol. 36, no. 2, pp. 296–302, Feb. 1964.

[8] J. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 5, pp. 367–377, Dec. 1972.

[9] K. Gopalan, "Pitch estimation using a modulation model of speech," in *5th International Conference on Signal Processing*. IEEE, Aug. 2000, pp. 786–791.

[10] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 727–730, Oct. 2001.

[11] C. Shahnaz, W.P. Zhu, and M.O. Ahmad, "Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 322–335, Jan. 2012.

[12] D.J. Hermes, "Measurement of pitch by subharmonic summation," *The Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, Jan. 1988.

[13] B. Resch, M. Nilsson, A. Ekman, and W.B. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 813–822, Mar. 2007.

[14] Y.M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 1805–1815, Dec. 1989.

[15] G. Seshadri and B. Yegnanarayana, "Performance of an event-based instantaneous fundamental frequency estimator for distant speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1853–1864, Sep. 2011.

[16] B. Yegnanarayana and K.S.R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, May 2009.

[17] S. Kadambe and G.F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 917–924, Mar. 1992.

[18] P.A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[19] K.S.R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[20] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, March 2012.

[21] P. Jain and R.B. Pachori, "GCI identification from voiced speech using the eigen value decomposition of Hankel matrix," in *8th International Symposium on Image and Signal Processing and Analysis*. IEEE, Sep. 2013, pp. 371–376.

[22] P.S. Rathore and R.B. Pachori, "Instantaneous fundamental frequency estimation of speech signals using DESA in low-frequency region," in *2013 International Conference on Signal Processing and Communication*. IEEE, pp. 470–473.

[23] H. Huang and J. Pan, "Speech pitch determination based on Hilbert-Huang transform," *Signal Processing*, vol. 86, no. 4, pp. 792–803, Apr. 2006.

[24] L. Qiu, H. Yang, and S.N. Koh, "Fundamental frequency determination based on instantaneous frequency estimation," *Signal Processing*, vol. 44, no. 2, pp. 233–241, Jan. 1995.

[25] G. Schlotthauer, M.E. Torres, and H.L. Rufiner, "A new algorithm for instantaneous F_0 speech extraction based on ensemble empirical mode decomposition," in *17th EURASIP Signal Processing Conference, Glasgow, Scotland, UK*, 2009, pp. 2347–2351.

[26] Y. Li, B. Xue, H. Hong, and X. Zhu, "Instantaneous pitch estimation based on empirical wavelet transform," in *19th International Conference on Digital Signal Processing*. IEEE, 2014, pp. 250–253.

[27] A. Upadhyay and R.B. Pachori, "Instantaneous voiced/non-voiced detection in speech signals based on variational mode decomposition," *Journal of the Franklin Institute*, vol. 352, no. 7, pp. 2679–2707, 2015.

[28] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, Feb. 2014.

[29] I. Daubechies, J. Lu, and H.T. Wu, "Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 243–261, 2011.

[30] D. P. Bertsekas, "Constrained Optimization and Lagrange Multiplier Methods," *Computer Science and Applied Mathematics, Boston: Academic Press*, 1982.

[31] L. Cohen, *Time-Frequency Analysis*, vol. 1, Englewood Cliffs, NJ: Prentice Hall, 1995.

[32] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of The Institute of Phonetic Sciences*. Amsterdam, 1993, vol. 17, pp. 97–110.

[33] R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*, Boca Raton, FL: CRC Press, 2000.

[34] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]," Version 5.3.51, June 2013, `http://www.praat.org/`.

[35] P. Fabrice, F.M. Georg, and A.A. William, "A pitch extraction reference database," in *4th European Conference on Speech Communication and Technology, Madrid, Spain, Sep. 18-21, 837-840*, 1995.

[36] "Noisex-92, [online]," Available: `http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html`.