On the Estimation of Fundamental Frequency From Nonstationary Noisy Speech Signals Based on the Hilbert–Huang Transform

L. Zão, *Member, IEEE*, and R. Coelho, *Senior Member, IEEE*

***Abstract*—This letter introduces a method based on the Hilbert–Huang transform (HHT) to estimate the fundamental frequency of nonstationary noisy speech signals. For this purpose, the target signals are analyzed by means of the ensemble empirical mode decomposition and the Hilbert transform. The main contribution of the proposed solution, namely HHT-Amp, relies on the extraction of pitch information from the instantaneous amplitude of the first decomposition modes. The HHT-Amp and four competitive algorithms are evaluated considering speech signals corrupted by five acoustic noises with different nonstationarity degrees. The HHT-Amp achieves the lowest gross error rate and mean absolute error for the most severe noisy conditions. This demonstrates that the proposed approach outperforms the baseline methods in estimating the fundamental frequency of noisy speech.**

***Index Terms*—Fundamental frequency estimation, Hilbert–Huang transform (HHT), nonstationary acoustic noises.**

I. Introduction

THE estimation of the speech fundamental frequency is a major issue for many speech processing applications, such as speech coding, speech synthesis, voice disorders detection, and speaker recognition. In a voiced speech segment, the fundamental frequency (F_0) consists on the rate of vibration of the vocal folds, which corresponds to the inverse of the pitch period (T_0). In the literature, different approaches have been proposed for F_0 detection. The YIN [1] and other methods based on the autocorrelation [2] and crosscorrelation [3] have been presented in the time domain, whereas the subharmonic-to-harmonic ratio (SHR) [4] and sawtooth waveform inspired pitch estimator (SWIPE) [5] are examples of spectral approaches. However, the accurate estimation of F_0 in noisy conditions, particularly in low signal-to-noise ratio (SNR), is still a challenging topic [6].

In the last decade, the adaptive processing of empirical mode decomposition (EMD) [7] solution and variants [8], [9] have been extensively applied for speech signal analysis [10]–[14]. This decomposition entails a series of intrinsic mode functions (IMF) which are completely based on the local properties of the analyzed signal. Additionally, EMD is generally combined with the Hilbert transform to analyze a wide range of nonlinear and nonstationary signals. The Hilbert–Huang transform (HHT) leads to the instantaneous frequencies and amplitudes of the IMFs as functions of time.

This letter proposes a F_0 estimation approach for speech signals corrupted by nonstationary acoustic noises based on the HHT. Different from other EMD-based techniques [10], [12], the F_0 values are not obtained from the instantaneous frequencies of the IMFs. Instead, the F_0 information is extracted from the instantaneous amplitude functions. The ensemble EMD (EEMD) [8] is here adopted to decompose the voiced segments of the speech signals. It is demonstrated that the F_0 values may be estimated even when the signals are severely corrupted by different acoustic noises.

Several experiments are conducted to prove the effectiveness of the proposed HHT-Amp solution in estimating the speech F_0. For this purpose, the speech signals collected from the Centre for Speech Technology Research (CSTR) database [15] are corrupted by noises collected from five real acoustic sources considering three SNR values: 0 dB, 5 dB, and 10 dB. The index of nonstationarity (INS) [16] analysis of the noisy signals is also included in this work. The autocorrelation [2], YIN [1], SWIPE [5], and the EMD-based F_0 estimator introduced in [12] are used as competitive methods. The results demonstrate that the proposed approach achieves the lowest values in terms of gross error (GE) rate and mean absolute error (MAE) for most of the noisy conditions.

II. Proposed HHT-Amp F_0 Estimation Method

Considering a speech signal $x(t)$, the proposed HHT-Amp solution is performed in three main steps. First, the HHT is applied to $x(t)$ in order to obtain a series of instantaneous amplitude functions. The second step consists on the computation of the autocorrelation function (ACF) from the amplitude signals. The *lags* where the ACF peaks occur are used to define a set of speech pitch period candidates, one per decomposition mode. Finally, a decision criteria is adopted to select which candidate better represents the pitch period $\hat{T}_0$. The estimated fundamental frequency is given by $\hat{F}_0 = 1/\hat{T}_0$.

Manuscript received October 18, 2017; revised November 22, 2017; accepted December 7, 2017. Date of publication December 11, 2017; date of current version December 29, 2017. R. Coelho was supported in part by the National Council for Scientific and Technological Development under Grant 307866/2015-7. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong He. *(Corresponding author: R. Coelho.)*

The authors are with the Laboratory of Acoustic Signal Processing, Military Institute of Engineering, Rio de Janeiro 22290-270, Brazil (e-mail: coelho@ime.eb.br).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2782267

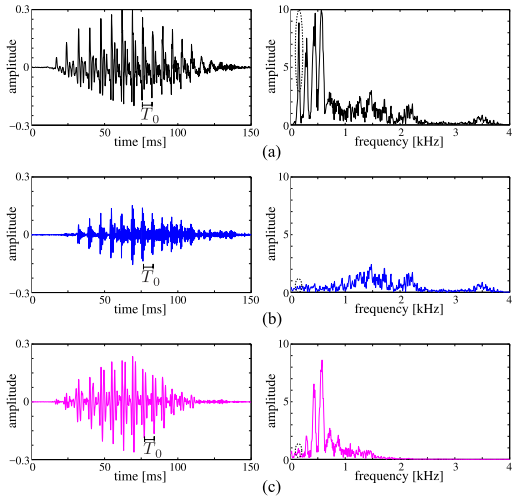

Fig. 1. Left column: (a) samples of a clean speech segment with 150 ms duration, (b) IMF 1, and (c) IMF 2 obtained with EEMD. Right column: the corresponding frequency responses obtained after the Fourier transform.

A. Application of the HHT (Step 1)

In this letter, the EEMD is selected to decompose the target signal $x(t) = \sum_{k=1}^{K} \text{IMF}_k(t) + r(t)$, where $r(t)$ is the residual. The EEMD algorithm states that the first IMFs are mainly composed of the fastest oscillations, i.e., highest frequencies, of the analyzed signal. However, they also present some nonnegligible low-frequency content of the input target signal [17]. This fact is illustrated in Fig. 1, where the left column depicts the samples of a clean speech segment with duration of 150 ms, and also the two first IMFs obtained with the EEMD. The sampling rate is 8 kHz. The speech segment and the two IMFs are Fourier transformed and the amplitude results are plotted in the right column of Fig. 1. Note from Fig. 1(a) that, for the time region around 80 ms, the pitch period can be manually labeled as $T_0 \approx 6.7$ ms. The fundamental frequency $F_0 = 1/T_0 \approx 149$ Hz approximately corresponds to the location of the first peak in the Fourier amplitude. As in Fig. 1(a), the T_0 value is detectable as the time interval between consecutive local minima of the two first IMFs, as highlighted in Fig. 1(b) and (c). It means that the speech pitch period can also be estimated from the periodicity of the decomposition modes.

After the decomposition, the Hilbert transform is then used to obtain a set of analytic signals $Z_k(t), k = 1, \ldots, K$, as

$$Z_k(t) = \text{IMF}_k(t) + j\, H\{\text{IMF}_k(t)\}, \tag{1}$$

where $H\{\text{IMF}_k(t)\}$ refers to the Hilbert transform of $\text{IMF}_k(t)$. The instantaneous amplitude $a_k(t)$ and the instantaneous frequency $\omega_k(t)$ of $\text{IMF}_k(t)$ are directly derived from the polar representation $Z_k(t) = a_k(t) \exp\left\{j \int \omega_k(t) dt\right\}$.

The left column of Fig. 2 depicts the instantaneous amplitude functions of the two IMFs presented in Fig. 1(b) and (c). Note that $a_k(t)$ is a slowly varying signal when compared to $\text{IMF}_k(t)$, since it corresponds to the amplitude modulation (AM) part of the IMF. It must be also observed from Fig. 2 that the $a_k(t)$ signals contain quasi-periodic segments in time interval $20 \leq t \leq 120$, in ms, with the same fundamental period T_0 as in the original speech segment. It means that the instantaneous amplitudes enable the estimation of the fundamental frequency of $x(t)$. This is the opposite behavior when compared to the Fourier analysis, illustrated in the right column of Fig. 1, where the F_0 value cannot be easily computed from the IMFs frequency responses.

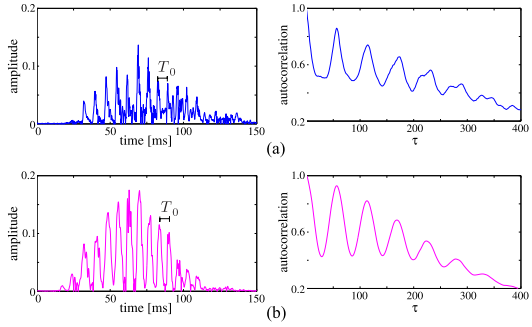

Fig. 2. Left column: instantaneous amplitude of (a) IMF 1 and (b) IMF 2, obtained from a clean speech signal. Right column: ACF computed from the instantaneous amplitude.

The extraction of speech F_0 from the instantaneous amplitude functions is one of the main contributions of this letter. In the literature, EMD-based F_0 estimators [10], [12] generally obtain the fundamental frequency from the instantaneous frequency functions $\omega_k(t)$. In this proposal, however, the F_0 is estimated directly from the amplitude signals $a_k(t)$. The main motivation is the quasi-periodicity obtained from voiced segments (refer to Fig. 2) of the analyzed signal.

B. Computation of the ACF (Step 2)

In the second step of the proposed HHT-Amp method, the autocorrelation is used to estimate the pitch period from the instantaneous amplitude functions. This is motivated by the ACF plots depicted in the right column of Fig. 2. In this example, the ACF curves are obtained considering all the samples from the amplitude signals $a_k(t), k = 1, 2$ presented in the left column. The ACF curves show the first peak at *lag* $\tau = 54$, which corresponds to the estimated pitch period $\hat{T}_0 = 54/8000 = 6.8$ ms, and the fundamental frequency $\hat{F}_0 = 148$ Hz. This demonstrates that the ACF of $a_k(t)$ is able to provide information about the fundamental frequency of the target speech signal $x(t)$.

Considering the problem of estimating the fundamental frequency of $x(t)$ at each time instant in the set $S = \{t_1, t_2, \ldots, t_Q\}$, consider

$$r_{t_q,k}(\tau) = \sum_{t=t_q-W/2}^{t_q+W/2-\tau} a_k(t)\, a_k(t+\tau) \tag{2}$$

the ACF computed from a frame of $a_k(t)$ with W samples centered at time instant $t_q, 1 \leq q \leq Q$. For each mode index k, let τ_0 correspond to the lowest value of τ where an ACF peak is located, restricted to the interval $\tau_{\min} \leq \tau_0 \leq \tau_{\max}$. The choices of $\tau_{\min}$ and $\tau_{\max}$ define the range $[F_{\min}, F_{\max}]$ of possible values of $\hat{F}_0$. If τ_0 exists, the pitch candidate $P_{t_q}(k)$ is then given as τ_0/f_s, where f_s refers to the sampling rate. Thus, up to K pitch candidates are found for each time instant.

Fig. 3 shows the instantaneous amplitude functions and the corresponding ACF from a noisy speech segment with duration

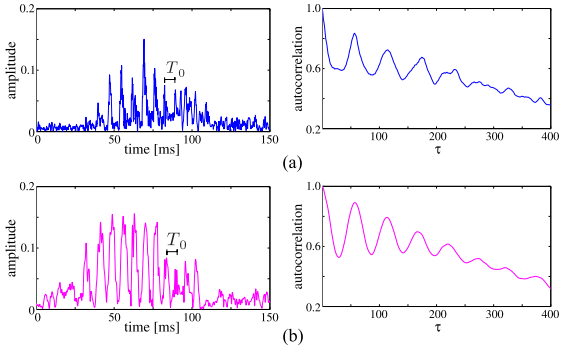

Fig. 3. Left column: instantaneous amplitude of (a) IMF 1 and (b) IMF 2, obtained from a speech signal corrupted with Babble noise. Right column: ACF computed from the instantaneous amplitude.

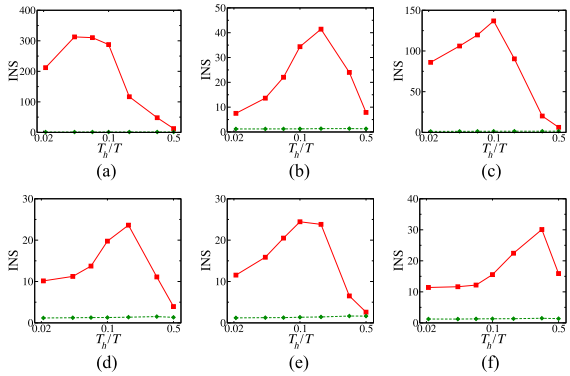

Fig. 4. Index of nonstationarity of (a) clean speech and speech corrupted by (b) Babble, (c) Car, (d) Helicopter, (e) Traffic, and (f) Train acoustic noises.

of 150 ms. In this example, the signal depicted in Fig. 1 is corrupted by the Babble noise [18] with SNR of 0 dB. Note from the left column that the amplitude functions also contain quasi-periodic segments with fundamental period close to those found in the original speech segment. It may be observed that, as for the clean speech, the ACF curves are able to capture the F_0 information, since the first ACF peaks also occur at $\tau = 54$. This fact indicates that the estimation of F_0 based on the autocorrelation of instantaneous amplitude is robust in terms of acoustic noise corruption.

C. Selection of the Pitch Candidates (Step 3)

Let $S_{\text{V}} = \{t_j, t_{j+1}, \ldots, t_{j+J}\}$ be the subset of S that contains all the time instants that are included in the same voiced speech segment. For each $t_q \in S_{\text{V}}$, this third step aims at selecting one of the pitch candidates $P_{t_q}(k), k = 1, \ldots, K$, as the fundamental pitch period $\hat{T}_0(t_q)$. For this purpose, the estimated pitch period is initially set as the pitch candidate obtained from IMF 1, i.e.,

$$\hat{T}_0(t_q) \leftarrow P_{t_q}(1), \quad q = j, \ldots, j+J\,. \tag{3}$$

The value of $\hat{T}_0(t_q)$ is then compared to the mean value $m(t_q)$ computed at the nearest neighbors time instants as

$$m(t_q) = E\left[\hat{T}_0(t_l)\,,\ |l-q| < V\right], \quad q = j, \ldots, j+J\,. \tag{4}$$

If $\hat{T}_0(t_q)$ deviates from $m(t_q)$ in more than a predefined threshold R, i.e., $|\hat{T}_0(t_q) - m(t_q)|/m(t_q) > R$, the $\hat{T}_0(t_q)$ value is updated by the pitch candidate obtained from the next IMF, $\hat{T}_0(t_q) \leftarrow P_{t_q}(2)$. The parameter V defines the maximum difference between neighboring time instants. R is the threshold of an acceptable deviation between the pitch candidate $\hat{T}_0(t_q)$ and the mean $m(t_q)$. For instance, considering that a $\hat{T}_0$ value is estimated every 4 ms, and setting $V = 10$ and $R = 0.2$ means that the pitch period is expected to vary less than 20% within a 40 ms interval. These values were selected from $R \in [0.05, 0.3]$ and $V \in [5, 20]$ as to achieve the lowest GE rates in preliminary experiments. For this purpose, the speech signals described in Section III are corrupted by five acoustic noises considering SNR of 10 dB.

After this procedure is conducted for each $t_q \in S_{\text{V}}$, the mean values are recalculated and the selection process is repeated until $|\hat{T}_0(t_q) - m(t_q)|/m(t_q) \leq R$ for every time instant t_q. If there is no candidate left to update $\hat{T}_0(t_q)$, the stopping criteria is guaranteed by setting $\hat{T}_0(t_q) \leftarrow m(t_q)$.

III. Experiments and Results

The proposed method is evaluated in F_0 detection experiments conducted with the CSTR database [15]. It is composed of 100 utterances sampled at 8 kHz spoken by a male (50) and a female (50) speakers. The reference F_0 values are available based on the recordings of laryngograph data. The experiments are conducted considering five acoustic noises collected from real sources: Babble and Car, from NOISEX [18], and Helicopter, Traffic, and Train from FreeSfx[1] database.

Different F_0 estimators are examined as baseline for the proposed HHT-Amp method. The YIN[2] and SWIPE[3] algorithms were obtained from the websites provided by the authors. The ACF was implemented using the Praat [19] software. For the EMD-based method [12], the Praat is adopted to obtain F_0 values with the cepstrum method. These are then used to select the instantaneous frequencies that compose the final F_0 estimates. The F_0 estimation methods are compared in terms of the GE rate and MAE. The GE is computed as the percentage of voiced frames for which the estimated value of $\hat{F}_0$ deviate by more than 20% from the reference F_0. The MAE corresponds to the the absolute error $|\hat{F}_0 - F_0|$ averaged over all the available reference values.

The index of nonstationarity [16] is here adopted to examine the nonstationarity degrees of the noisy speech signals. Fig. 4(a) depicts the INS values (continuous line) of a clean speech utterance with total time duration $T = 1.5$ s. The INS obtained from the same signal corrupted with the five noises, considering SNR of 0 dB, are depicted in Fig. 4(b)–(f). The values are calculated considering different observation scales T_h/T, where T_h is the length adopted in the short-time spectral analysis. The values of T_h, in ms, are: 32, 64, 100, 150, 250, 500, and 750. The dashed green lines refer to the stationarity threshold $\gamma \approx 1$. The

[1] Available in http://www.freesfx.co.uk/.
[2] Available in http://audition-backend.ens.fr/adc/.
[3] Available in http://www.cise.ufl.edu/ acamacho/publications/swipep.m.

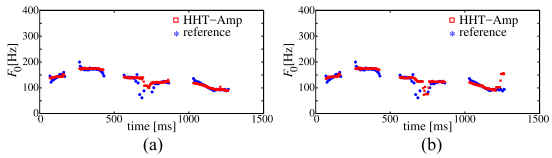

Fig. 5. F_0 values estimated with the HHT-Amp method: (a) clean speech and (b) speech corrupted with Babble noise.

TABLE I
GE RATE (%) OBTAINED IN SPEECH F_0 ESTIMATION WITH THE PROPOSED HHT-AMP AND THE BASELINE METHODS

Noise	SNR	ACF	YIN	SWIPE	EMD [12]	HHT-Amp
Babble $\overline{\text{INS}} = 16.1$	0 dB	36.4	33.3	34.1	45.6	**25.2**
	5 dB	16.6	16.8	**14.6**	34.3	17.1
	10 dB	7.6	8.5	**6.4**	24.3	14.1
Traffic $\overline{\text{INS}} = 10.9$	0 dB	49.9	59.8	46.1	36.6	**12.8**
	5 dB	31.0	36.3	27.1	26.5	**10.9**
	10 dB	17.1	20.1	13.8	18.8	**9.8**
Train $\overline{\text{INS}} = 8.5$	0 dB	35.0	44.3	43.7	34.6	**20.9**
	5 dB	20.1	27.2	26.2	26.8	**14.7**
	10 dB	**10.3**	14.9	12.8	20.1	11.7
Helicopter $\overline{\text{INS}} = 1.2$	0 dB	51.1	58.4	58.7	43.0	**23.0**
	5 dB	25.4	30.9	25.6	29.6	**15.4**
	10 dB	11.0	13.6	**9.9**	20.6	12.2
Car $\overline{\text{INS}} = 1.1$	0 dB	14.9	18.5	18.4	17.1	**9.9**
	5 dB	**7.6**	9.5	10.1	12.7	9.2
	10 dB	**4.9**	6.2	5.9	10.4	9.0
Average		22.6	26.6	23.6	26.7	**14.4**

INS analysis shows that the clean and noisy speech signals are nonstationary, since the INS values are greater than γ for all the time scales. Depending on the acoustic noise sources, the maximum INS values vary from 24 to 140 with Helicopter and Car, respectively. This reinforces the adoption of HHT to analyze noisy speech signals, since it is mainly defined to process nonstationary signals.

In the HHT-Amp experiments, the EEMD is applied considering 100 different realizations of white noise. The ratio between the white noise and the signal variances is 0.01. For the ACF computation, the IMFs are upsampled to 32 kHz, and the IMFs frames are composed with $W = 1024$ samples. The pitch candidates are collected from the first $K = 3$ IMFs with ACF peaks restricted to the interval $64 \leq \tau_0 \leq 640$. This corresponds to $\hat{F}_0$ values in the range $[50, 500]$, in Hz.

Fig. 5 illustrates the speech F_0 estimation with the HHT-Amp method. The values depicted in Fig. 5(a) are obtained from the same clean speech utterance analyzed in Fig. 4(a). The first voiced speech segment corresponds to the samples depicted in Fig. 1(a). Note that for most of the time instants the estimated F_0 is close to the reference values. This achievement also holds for the values depicted in Fig. 5(b), which concerns the same speech utterance corrupted with the nonstationary Babble noise [18] and SNR of 0 dB. The F_0 estimation results indicate that the proposed HHT-Amp method is robust to acoustic noise corruption considering low SNR values.

Table I presents the GE rates obtained in the experiments with the proposed solution and the baseline methods. The results are shown according to the average INS ($\overline{\text{INS}}$) of the acoustic noises. It may be observed that the HHT-Amp solution outperforms the competitive approaches for most of the noisy conditions (9 from a total of 15). The proposed technique is particularly interesting for the most severe noisy situations, e.g., SNR of 0 dB and different nonstationarity degrees. For this scenario, the HHT-Amp achieves the lowest GE results. For example, considering the Traffic noise with SNR of 0 dB and $\overline{\text{INS}}$ of 10.9, the GE rate is decreased from 36.6% with the EMD [12] to 12.8% with HHT-Amp, i.e., a reduction of 23.8 percentage points (p.p.). The GE rates also achieves reduction of 15.6 p.p. and 9.0 p.p. for SNR of 5 dB and 10 dB, respectively. In average, the proposed solution reaches a GE rate of 14.4%, which is 8.2 p.p. and 12.3 p.p. lower than the ACF and EMD methods, respectively.

TABLE II
MAE (HZ) OBTAINED WITH THE PROPOSED HHT-AMP AND THE BASELINE METHODS

Noise	SNR	ACF	YIN	SWIPE	EMD [12]	HHT-Amp
Babble $\overline{\text{INS}} = 16.1$	0 dB	34.0	40.1	28.5	41.4	**26.8**
	5 dB	16.1	21.4	**14.2**	21.5	18.2
	10 dB	**8.7**	12.2	9.2	12.3	14.0
Traffic $\overline{\text{INS}} = 10.9$	0 dB	52.0	71.9	38.4	55.9	**15.1**
	5 dB	32.9	43.4	22.8	36.6	**12.2**
	10 dB	19.6	24.7	13.0	22.7	**9.6**
Train $\overline{\text{INS}} = 8.5$	0 dB	32.2	57.4	21.1	39.5	**19.1**
	5 dB	18.0	33.5	**12.4**	23.4	13.6
	10 dB	10.6	19.7	**8.7**	14.1	10.8
Helicopter $\overline{\text{INS}} = 1.2$	0 dB	47.0	73.2	23.9	53.9	**18.0**
	5 dB	24.8	39.2	**12.1**	30.5	13.9
	10 dB	11.6	19.0	**8.4**	15.4	11.8
Car $\overline{\text{INS}} = 1.1$	0 dB	33.0	28.4	9.8	31.8	**9.0**
	5 dB	17.2	16.3	**8.5**	18.6	8.6
	10 dB	11.4	12.5	**7.5**	13.3	8.2
Average		24.6	34.2	15.9	28.7	**13.9**

The proposed HHT-Amp method is also evaluated in terms of MAE, as shown in Table II. Note that the HHT-Amp yields to the lowest error for most of the noisy situations. Once again, it attains the best results for the most severe conditions, i.e., SNR of 0 dB. In average, the HHT-Amp leads to an absolute error of 13.9 Hz, which is 2.0 Hz lower than the SWIPE and 14.8 Hz lower than the EMD.

IV. CONCLUSION

This letter introduced a method based on the HHT to estimate the fundamental frequency of speech signals. The F_0 information is extracted from the instantaneous amplitude of each IMF. The proposed HHT-Amp estimator is evaluated in experiments considering speech signals corrupted by real acoustic noises with different nonstationarity degrees. The results show that, in average, the HHT-Amp outperforms four competitive F_0 techniques considering speech signals corrupted by acoustic noises. The average GE rate is reduced from 22.6% with ACF to 14.4% with the HHT-Amp. It leads to MAE 2.0 Hz lower than the spectral SWIPE method. Furthermore, the proposed solution achieves the best results for speech signals corrupted by acoustic noises considering SNR of 0 dB.

REFERENCES

[1] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[2] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 25, pp. 24–33, Feb. 1977.

[3] S. A. Samad, A. Hussain, and L. K. Fah, "Pitch detection of speech signals using the cross-correlation technique," in *Proc. TENCON*, 2000, vol. 1, pp. 283–286.

[4] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," in *Proc. Int. Conf. Spoken Lang. Process.*, Oct. 2000, vol. 4, pp. 676–679.

[5] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 124, no. 3, pp. 1638–1652, Sep. 2008.

[6] G. Aneeja and B. Yegnanarayana, "Extraction of fundamental frequency from degraded speech using temporal envelopes at high SNR frequencies," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 829–838, Apr. 2017.

[7] N. Huang *et al.*, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A, Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[8] Z. Wu and N. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 1, pp. 1–41, 2009.

[9] M. Torres, M. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 4144–4147.

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

[11] T. Hasan and M. Hasan, "Suppression of residual noise from speech signals using empirical mode decomposition," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 2–5, Jan. 2009.

[12] H. Hong, Z. Zhao, X. Wang, and Z. Tao, "Detection of dynamic structures of speech fundamental frequency in tonal languages," *IEEE Signal Process. Lett.*, vol. 17, no. 10, pp. 843–846, Oct. 2010.

[13] R. Coelho and L. Zão, "Empirical mode decomposition theory applied to speech enhancement," in *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering and Processing*, R. Coelho, V. Nascimento, R. Queiroz, J. Romano, and C. Cavalcante, Eds. Boca Raton, FL, USA: CRC Press, 2015.

[14] L. Zão, R. Coelho, and P. Flandrin, "Speech enhancement with EMD and hurst-based mode selection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 5, pp. 899–911, May 2014.

[15] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *Proc. EUROSPEECH'93*, Sep. 1993, pp. 1003–1006.

[16] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3459–3470, Jul. 2010.

[17] P. Flandrin, G. Rilling, and P. Gonçalvès, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 112–114, Feb. 2004.

[18] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.