Noisy Speech Based Temporal Decomposition to Improve Fundamental Frequency Estimation

Anderson Queiroz, *Student Member, IEEE*, and Rosângela Coelho, *Senior Member, IEEE*

***Abstract*—This article introduces a novel method to separate voiced frames of noisy speech signals into low-frequency or high-frequency. This separation improves the accuracy of fundamental frequency (F0) estimators. In this proposal, the target signal is analyzed by means of the ensemble empirical mode decomposition. Next, the pitch information is extracted from the first decomposition modes. This feature indicates the frequency region where the speech F0 should be located, thus separating the frames into low-frequency or high-frequency. The frames separation is then applied to correct pitch candidates extracted from a F0 detection method, improving the estimation accuracy. The proposed method and a baseline separation approach are evaluated considering four different F0 estimation algorithms. Experiments are conducted with the CSTR and TIMIT databases, and six noises with various signal-to-noise ratios. The Gross Error (GE) and Mean Absolute Error (MAE) metrics are adopted to evaluate the solutions in terms of F0 estimation errors. Results show that the proposed method outperforms the baseline, in terms of low/high frequency separation accuracy. Moreover, the novel solution is able to better improve F0 detection accuracy under different noisy conditions.**

***Index Terms*—F0 estimation, low/high frequency separation, noisy speech, TF decomposition.**

I. Introduction

URBAN noisy acoustic scenarios may affect speech temporal and spectral attributes. A robust estimation of the fundamental frequency (F0) feature is a requirement for a diversity applications, such as speech coding [1], speech synthesis [2], speaker and speech [3], [4] recognition. In a voiced speech segment, the F0 consists on the vibration rate of the vocal folds, which corresponds to the inverse of the pitch period. The investigation of harmonic noisy components of speech signals has also gained significant attraction for strategies that attain intelligibility gain [5]–[7]. These harmonic components and also the formants play a significant role for speech intelligibility in noise [8]–[10]. Therefore, methods and systems which estimate fundamental frequency accurately need to be explored, particularly, at low signal-to-noise ratios (SNRs), where noise components may cause errors in estimation algorithms.

Several approaches have been proposed in the literature for F0 detection in the clean speech. Classical time domain methods are generally based on the Auto-Correlation Function (ACF) [11]–[13]. YIN [14] is an alternative solution that uses the local minima of the Normalized Mean Difference Function (NMDF) with some post-processing procedures to avoid estimation errors caused by signal amplitude changes. Nevertheless, SHR [15] and SWIPE [16] are examples of spectral techniques. The SHR method introduces the concept of sub-harmonic to harmonic ratio, and F0 detection is performed by looking for values which maximize this ratio. SWIPE estimates the F0 as the frequency of the sawtooth waveform whose spectrum best matches the spectrum of the input signal. Despite the variety of proposed methods, an accurate estimation of F0 in severe noisy conditions is still a challenging task in speech signal processing.

In recent years, solutions have been proposed for F0 estimation in noisy speech [17]–[19]. In [17], the Pitch Estimation Filter with Amplitude Compression (PEFAC) is introduced, which applies a prefiltering to reduce the noise effects with interesting accuracy results. Machine learning based approaches [20]–[23] investigate classifiers that include feed-forward, recurrent and convolutional neural networks. Furthermore, other F0 detection methods [24], [25] are based on the Hilbert-Huang Transform (HHT) [26]. Specifically, the Empirical Mode Decomposition (EMD) [26] or its variations, which realizes a time-frequency (TF) decomposition to analyze the noisy speech signal for various tasks [27]–[29]. In the HHT-Amp technique [25] the F0 estimates are achieved from the instantaneous amplitude functions of the target signal. Although these F0 estimation methods are examined for noisy environments, some errors can still occur, e.g., F0 subharmonics detection, octave errors, and indistinct periodic background noise from the voiced speech [13].

In a recent work [30], a strategy is introduced to improve pitch detection attained by conventional estimation algorithms considering noisy scenarios. To this end, a Deep Convolutional Neural Network (DCNN) is trained in order to classify the voiced speech frames into low or high frequency. According to the classification, new F0 candidates are computed considering typical probable types of errors for a certain F0 value estimated by a detection method. Taking into account a procedure which involves two spectral attributes of speech frames and a cost function, the enhanced pitch is selected from the new candidates.

This work introduces the Frequency Separation for Fundamental Frequency Enhancement (FSFFE) method to separate

Manuscript received 9 December 2021; revised 7 April 2022 and 26 May 2022; accepted 3 July 2022. Date of publication 13 July 2022; date of current version 5 August 2022. This work was supported in part by the National Council for Scientific and Technological Development under Grant 308155/2019, in part by Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro under Grant 203075/2016, and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil under Grant Code 001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Xie.
(Corresponding author: Rosângela 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; anderson.queiroz@ime.eb.br).
Digital Object Identifier 10.1109/TASLP.2022.3190670

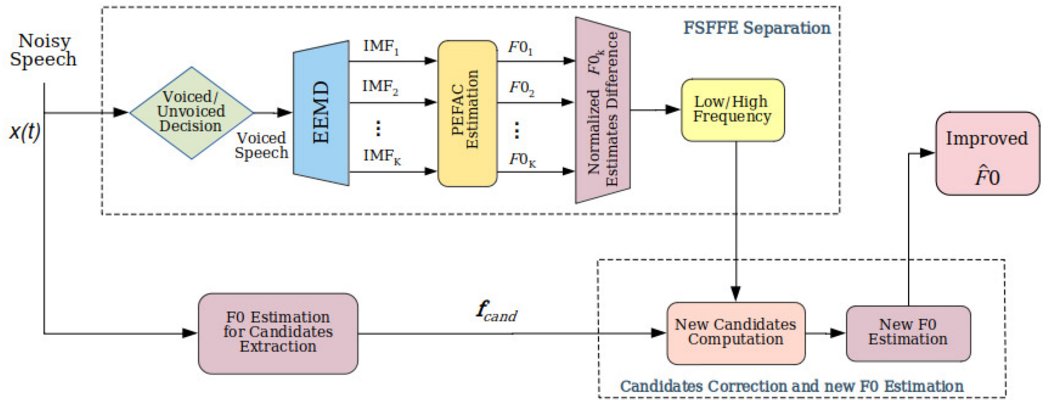

Fig. 1. Block diagram of the proposed FSFFE method for low/high frequency separation and improvement of F0 estimation accuracy.

harmonic noisy speech into low-frequency or high-frequency frames, in order to improve the F0 estimation accuracy. The proposed method applies the time-frequency Ensemble EMD (EEMD) [31] to decompose the noisy signal into a series of Intrinsic Mode Functions (IMFs). The first IMFs present the fastest oscillations referring to speech, which allows to attenuate the low-frequency noisy masking components. In this first step, PEFAC [17] is considered to estimate F0 from decomposition modes, expressing the low/high frequency tendency of speech frames. Then, a normalized distance that reflects the variation property of F0 is computed, selecting only the two IMFs with less variation in comparison to each other. Finally, the mean F0 of selected IMFs is compared to a threshold to indicate if the frame is considered low- or high-frequency. According the proposed separation, frequency candidates attained from a F0 detection method are corrected, leading to a set of enhanced candidates, and consequently improving the accuracy of F0 estimation.

Several experiments are conducted to examine the effectiveness and accuracy of the FSFFE method. For this purpose, speech utterances collected from two databases (CSTR [32] and TIMIT [33]) are corrupted by six real acoustic noises, considering five SNR values: -15 dB, -10 dB, -5 dB, 0 dB and 5 dB. The FSFFE method and the baseline approach [30] (denoted in this work by DCNN-Based Pitch Separation - DCNN-BPS) are examined in terms of improving the accuracy of fundamental frequency estimation, considering three F0 detection techniques: SHR [15], SWIPE [16] and HHT-Amp [25]. The Gross Error rate (GE) [34] and Mean Absolute Error (MAE) [35] are considered to evaluate the proposed and baseline method. Experiments demonstrate that F0 approaches enhanced by FSFFE method achieve the lowest error values. Furthermore, FSFFE+HHT-Amp shows the best overall scores when compared to competitive techniques.

The main contributions of this work are:

- Introduction of the FSFFE method for voiced speech frames separation into low/high frequency.
- Definition of a criterion to correct the candidates achieved from F0 estimation techniques.
- Accuracy improvement of F0 estimation approaches with error reduction from noisy speech signals with low SNR values.

The remaining of this article is organized as follows. Section II introduces the FSFFE method for voiced speech low/high frequency separation and correction of F0 candidates. It also provides descriptions of the F0 estimation algorithms adopted in the experiments. An explanation of the baseline DCNN-BPS for error correction of pitch estimates is included in Section III. Section IV presents the evaluation experiments and results. Finally, Section V concludes this work.

II. The Proposed Method

The FSFFE method to improve the fundamental frequency estimation accuracy includes four main steps: time-frequency noisy speech signal decomposition (EEMD), F0 estimation of Intrinsic Mode Functions (IMFs), normalized distance computation, that reflects the F0 variation property for frame separation in low/high frequency, and finally, candidates correction from the F0 estimators. Fig. 1 illustrates the block diagram of the proposed method.

A. TF Decomposition of Noisy Speech

The first step is devoted to the Ensemble Empirical Mode Decomposition (EEMD). The general idea of the EMD [26] is to analyze a signal $x(t)$ between two consecutive extrema (minima or maxima), and define a local high-frequency part, also called detail $d(t)$, and a local trend $a(t)$, such that $x(t) = d(t) + a(t)$. An oscillatory IMF is derived from the detail function $d(t)$. The high- versus low-frequency separation procedure is iteratively repeated over the residual $a(t)$, leading to a new detail and a new residual. Since the decomposition can only be applied if the last computed residual $a(t)$ is composed of at least two extrema, the input signal can be decomposed into a finite number of IMFs. Thus, the decomposition leads to a series of IMFs and a residual, such that

$$x(t) = \sum_{k=1}^{K} \mathrm{IMF}_k(t) + r(t) \tag{1}$$

where $\mathrm{IMF}_k(t)$ is the k-th mode of $x(t)$ and $r(t)$ is the last residual. In contrast to other signal decomposition methods, a set of basis functions is not demanded for the EMD. Besides,

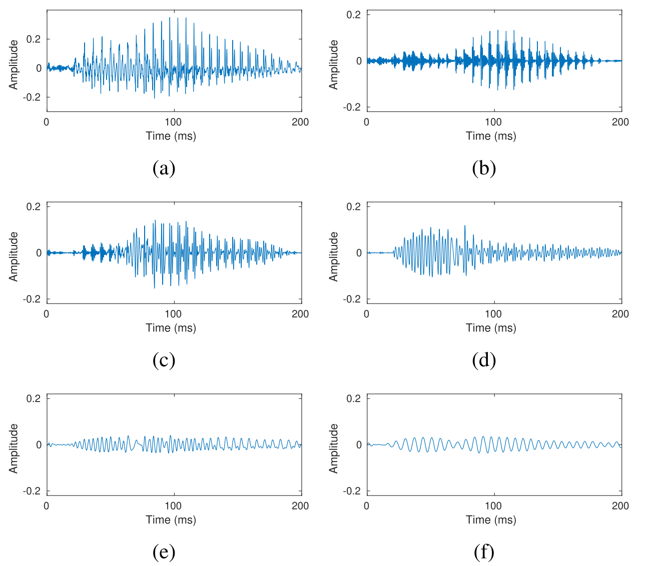

Fig. 2. (a) Samples of a clean speech segment with 200 ms duration, (b) IMF_1, (c) IMF_2, (d) IMF_3, (e) IMF_4 and (f) IMF_5 obtained with EEMD.

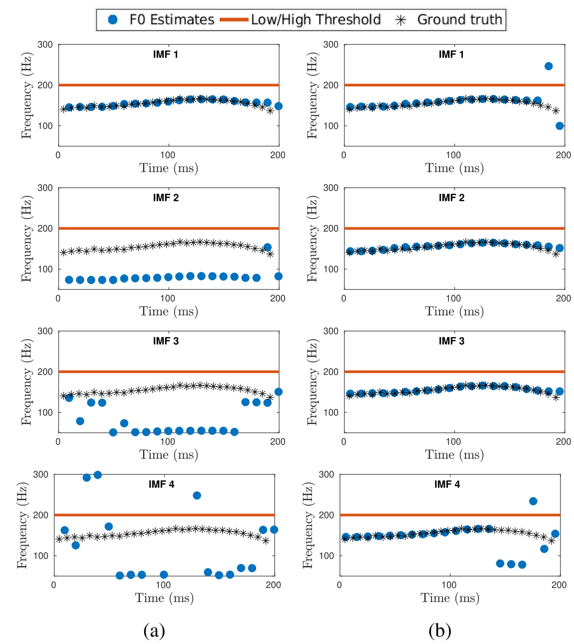

Fig. 3. F0 estimation with (a) HHT-Amp and (b) PEFAC, for the first four IMFs of a speech segment corrupted with Babble noise and SNR $= 0$ dB.

this strategy leads to fully data-driven decomposition modes and does not require the stationarity of the target signal.

The EEMD was introduced in [36] to overcome the mode mixing problem that generally occurs in the original EMD [26]. The key idea is to average IMFs obtained after corrupting the original signal using several realizations of White Gaussian Noise (WGN). Thus, EEMD algorithm can be described as:

1) Generate $x^n(t) = x(t) + w^n(t)$ where $w^n(t)$, $n = 1, \ldots, N$, are different realizations of WGN;
2) Apply EMD to decompose $x^n(t)$, $n = 1, \ldots, N$, into a series of components $\text{IMF}_k^n(t)$, $k = 1, \ldots, K$;
3) Assign the k-th mode of $x(t)$ as

$$\text{IMF}_k(t) = \frac{1}{N} \sum_{n=1}^{N} \text{IMF}_k^n(t); \tag{2}$$

4) Finally, $x(t) = \sum_{k=1}^{K} \text{IMF}_k(t) + r(t)$, where $r(t)$ is the residual.

Fig. 2 illustrates the first five IMFs obtained with EEMD from a sample speech segment of 200 ms collected from the CSTR [32] database. Note that the first IMFs are composed of faster oscillations than the last ones. Therefore, the first modes contain the high-frequency components of the signal. Furthermore, it is possible to observe that the cutoff frequency between IMFs is time-varying and signal dependent. Although the lack of mathematical formulation, it is interesting to point out that EEMD is a very powerful tool for analyzing non-stationary real signals and has been successfully applied in several research areas [29], [37].

B. F0 Estimation of IMFs

In this step, the F0 of frames of each IMF is estimated by the selected PEFAC [17] algorithm. This solution detects the F0 by convolving the power spectrum of the frame in the log-frequency domain, using a filter that sums the energy of the pitch harmonics. This filter highlights the harmonic components, improving the noise-robustness of the algorithm. Thus, let $\hat{F}0_{k,q}$ denote the F0 value estimated from frame q of $\text{IMF}_k(t)$, the $\hat{F}0_q$ vector is composed as

$$\hat{F}0_q = \begin{bmatrix} \hat{F}0_{1,q} \\ \hat{F}0_{2,q} \\ \vdots \\ \hat{F}0_{K,q} \end{bmatrix}, \tag{3}$$

to express the tendency that the frame is placed in a low or high frequency region. Although the first IMFs are composed of fastest oscillations, these components also present some nonnegligible low-frequency content of the speech signal [25], [38]. Therefore, in this article it is considered only the first four IMFs ($K = 4$) in order to avoid the acoustic noise masking effect, whose energy is mostly concentrated at low frequencies [27], [38], [39].

Fig. 3 shows the F0 estimated with HHT-Amp [25] and PEFAC in the first four IMFs of a speech segment collected from the CSTR database [32] and corrupted by Babble [40] noise with SNR $= 0$ dB. Default parameters of PEFAC are considered, whose voiced segment is split into overlapping frames with 90 ms length and 10 ms frame shift. Note from Fig. 3(a) that HHT-Amp presents F0 estimates with greater accuracy particularly for IMF_1. However for other IMFs, the estimated F0 values differ from the ground truth. As can be noted from Fig. 3(b), PEFAC estimates are more accurate than HHT-Amp in all the four IMFs. In most frames the frequencies have similar values for all the IMFs, particularly when the decomposition attenuates the noisy component of speech. A comparison with other F0 estimation algorithms was conducted, such as SHR [15] and SWIPE [16] approaches. The PEFAC technique provided a greater convergence with the ground truth in the four IMFs, resulting in a more accurate detection of the frequency region in which the true F0 is placed.

C. Low/High Frequency Separation

A normalized distance is computed between IMFs for the successive frames, in order to detect and overcome the differences in the estimated F0. Let k and k' denote IMF indexes, the distance $\delta^q_{\hat{F}0}$ is described as

$$\delta^q_{\hat{F}0}(k,k') = \left|\frac{\hat{F}0_{k,q} - \hat{F}0_{k,'q}}{\hat{F}0_{k,q} + \hat{F}0_{k,'q}}\right|. \quad (4)$$

This distance is determined for different values of k and k', resulting in the following matrix

$$\delta^q_{\hat{F}0} = \begin{bmatrix} 0 & \delta^q_{\hat{F}0}(1,2) & \cdots & \delta^q_{\hat{F}0}(1,K) \\ \delta^q_{\hat{F}0}(2,1) & 0 & \cdots & \delta^q_{\hat{F}0}(2,K) \\ \vdots & \vdots & \ddots & \vdots \\ \delta^q_{\hat{F}0}(K,1) & \delta^q_{\hat{F}0}(K,2) & \cdots & 0 \end{bmatrix}. \quad (5)$$

The row components of the matrix are summed, and the resulting values obtained express the variation property for the k-th IMF. The two IMFs with the smallest variation scores are selected, and the frequency region is defined as the mean value of PEFAC F0 estimates ($\bar{F}0_q$). Finally, a low-frequency to high-frequency threshold γ is adopted, and the separation is performed such as

$$\begin{cases} \bar{F}0_q \leq \gamma, & \text{low-frequency frame;} \\ \bar{F}0_q > \gamma, & \text{high-frequency frame.} \end{cases} \quad (6)$$

In [41] it is shown that the variability of speech F0 is between 50-200 Hz for men and 120-350 Hz for women. Therefore, in this study, a threshold of $\gamma = 200$ Hz is considered, since this is an average value for both genders of speakers.

D. Extraction and Correction of Candidates

The F0 estimation techniques are prone to different types of errors. Typical errors in F0 estimation, e.g., halving and doubling errors, are caused by the detection of harmonics other than the first, resulting in F0 estimates that are multiples of the true F0 [13]. In order to overcome this issue, the proposed method aims to correct the frequency candidates extracted by a F0 detection algorithm according to these types of errors. The FSFFE method states that, the F0 candidates (f_{cand}) must lie in [50,200] Hz or [200,400] Hz, for low-frequency and high-frequency frames, respectively. Thus, considering a low-frequency frame, f_{cand} is corrected following the criteria

$$f_c = \begin{cases} f_{cand}, & 50 \leq f_{cand} \leq 200 \\ 0.5f_{cand}, & 200 < f_{cand} \leq 400 \\ 0.25f_{cand}, & f_{cand} > 400 \end{cases}. \quad (7)$$

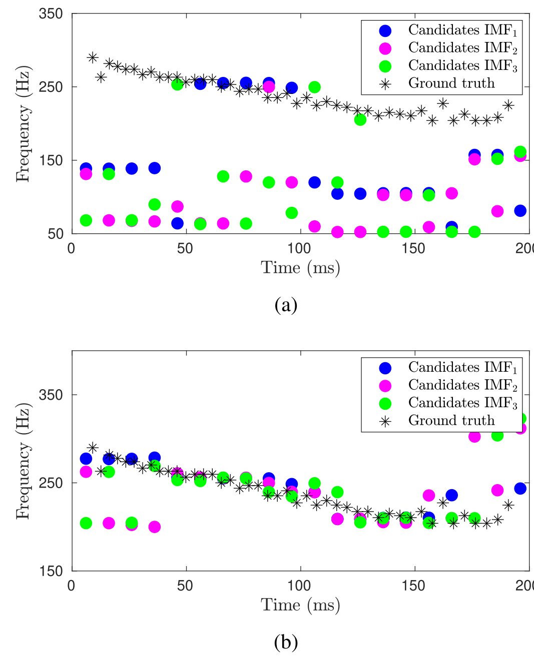

Fig. 4. (a) Original F0 Candidates from HHT-Amp method for a high-frequency speech segment corrupted by Babble noise with SNR = 0 dB, and (b) candidates adjusted by the proposed separation and correction algorithm.

where f_c is the corrected F0 candidate. Finally, the high-frequency frame is adjusted as follows:

$$f_c = \begin{cases} 4f_{cand}, & 50 \leq f_{cand} \leq 100 \\ 2f_{cand}, & 100 < f_{cand} \leq 200 \\ f_{cand}, & 200 < f_{cand} \leq 400 \\ 0.5f_{cand}, & f_{cand} > 400 \end{cases}. \quad (8)$$

Fig. 4 illustrates the F0 candidates correction procedure for a speech segment corrupted by Babble noise with SNR = 0 dB, with 200 ms duration. Fig. 4(a) denotes the F0 candidates extracted using the HHT-Amp [25] estimation technique, which computes a set of three candidates (one candidate from each of the first three IMFs) for a 10 ms time interval. Ground truth indicates that the entire speech segment is classified as high-frequency. Note that most candidates are located in the low-frequency region ($f_{cand} \leq 200$ Hz). Fig. 4(b) shows the new candidates, which are corrected according to the criteria described in (7) and (8). Moreover, observe that the adjusted F0 candidates take place in high-frequency region, matching the ground truth.

E. F0 Estimation Techniques Adopted in Evaluation

The SHR [15], SWIPE [16], and HHT-Amp [25] methods are considered in the experiments. The main idea is to evaluate the proposed separation and following correction of candidates to effectively improve the accuracy of the F0 estimators. A brief description of each method is presented below:

1) SHR: The SHR method [15] is based on the definition of a parameter measure called Sub-Harmonic-Harmonic Ratio, designed to describe the amplitude ratio between subharmonics and harmonics. Let $A(f)$ denote the amplitude spectrum for each short-term signal, the sum of harmonic amplitude is described as

$$SH = \sum_{n=1}^{N} A(nF0), \tag{9}$$

where N and $F0$ are the maximum number of harmonics considered in the spectrum and fundamental frequency, respectively. In other hand, the sum of subharmonic amplitude is achieved assuming at one half of F0

$$SS = \sum_{n=1}^{N} A\left(\left(n - \frac{1}{2}\right) F0\right). \tag{10}$$

Finally, SHR is the ratio between SS and SH:

$$SHR = \frac{SS}{SH}. \tag{11}$$

In [15], this ratio is computed on a logarithmic frequency scale. If the obtained SHR value is greater than a certain threshold, subharmonics frequencies are considered in the analysis. Otherwise, the harmonic frequencies are adopted in the F0 detection.

2) SWIPE: The core idea of SWIPE [16] is that if a signal is periodic with fundamental frequency f, its spectrum must contain peaks at multiples of f and valleys in between. Since each peak is surrounded by two valleys, the Average Peak-to-Valley Distance (APVD) for the n-th peak is defined as

$$d_n(f) = |X(nf)| - \frac{1}{2}\left[\left|X\left(\left(n - \frac{1}{2}\right) f\right)\right| + \left|X\left(\left(n + \frac{1}{2}\right) f\right)\right|\right] \tag{12}$$

where X is the estimated spectrum of the signal which takes frequency as input and outputs corresponding density. The global APVD is achieved by averaging over the first p peaks

$$D_n(f) = \frac{1}{p} \sum_{n=1}^{p} d_n(f). \tag{13}$$

The F0 estimated is the f value that maximizes this function, searching f in the range [50 500]Hz candidates, with samples distributed every $1/48$ units on a base-2 logarithmic scale.

3) HHT-Amp: The HHT-Amp method is summarized as follows:

- Apply the EEMD as described in Section II-A to decompose the voiced sample sequence $x_q(t)$.
- Compute the instantaneous amplitude functions $a_{k,q}(t) = |Z_{k,q}(t)|, k = 1, \ldots, K$, from the analytic signals defined as

$$Z_{k,q}(t) = \text{IMF}_{k,q}(t) + j\, H\{\text{IMF}_{k,q}(t)\}, \tag{14}$$

  where $H\{\text{IMF}_{k,q}(t)\}$ refers to the Hilbert transform of $\text{IMF}_{k,q}(t)$.
- Calculate the ACF $r_{k,q}(\tau) = \sum_t a_k(t)\, a_k(t+\tau)$ of the amplitude functions $a_{k,q}(t), k = 1, \ldots, K$.
- For each decomposition mode k, let τ_0 be the lowest τ value that correspond to an ACF peak, subject to $\tau_{min} \leq \tau_0 \leq \tau_{max}$. The restriction is applied according to the range $[F_{min}, F_{max}]$ of possible F_0 values. The k-th pitch candidate is defined as τ_0/f_s, where f_s refers to the sampling rate.
- Apply the decision criterion defined in [25] to select the best pitch candidate $\hat{T}_0$. The estimated F_0 is given by $\hat{F}_0 = 1/\hat{T}_0$.

In [25], it was shown that the HHT-Amp method achieves interesting results in estimating the fundamental frequency of noisy speech signals. HHT-Amp was evaluated in a wide range of noisy scenarios, including five acoustic noises, outperforming four competing estimators in terms of GE and MAE.

III. DCNN-BPS Technique

This Section describes the DCNN-BPS technique [30] adopted as baseline solution for the low/high frequency separation experiments. This technique consists on training a Deep Convolutional Neural Network (DCNN), in order to classify a voiced speech frame into low ($F0 \leq 200$ Hz) or high ($F0 > 200$ Hz) frequency. The DCNN architecture adopted is based on VGGNet [42], with six convolutional layers, three Fully-Connected (FC) layers and an output classification layer (Softmax). The DCNN input sequence is a 60 ms extracted directly from the voiced speech samples with a sampling rate of 16 kHz, or 960 samples.

The DCNN is trained considering 30% of the voiced frames of the CSTR [32] database and 25% of a subset [1] of the TIMIT [33] database. The training dataset is composed of 140000 speech frames, considering seven types of noises: Babble, SSN (Speech Shaped Noise), Cafeteria, Train, Helicopter, Traffic and Volvo. The SNR values (-10 dB and 0 dB) and architecture of the DCNN are the same as adopted in [30].

According to the low- or high-frequency classification, new F0 candidates f_j are extracted based on initial estimation attained by a conventional method f_p. The improved F0 estimate value is selected from the set of new candidates by exploiting a restrained selection procedure. Two spectral attributes are introduced to assist in selecting the improved F0 among candidates: Weighted Euclidean Deviation ($d_{j,q}$) and Weighted Comb Filtering ($y_{j,q}$). The first attribute is computed observing the first five frequency peaks positions in each frame, resulting in a peak vector $P_{i,q} = [P_{1,q}, P_{2,q}, \ldots, P_{5,q}]$. For a pitch candidate $f_{j,q}$, a candidate vector is defined as $V_{j,q} = [v_{j1,q}, v_{j2,q}, \ldots, v_{j5,q}]$, where $v_{ji,q}$ is the point-wise ratio between the peak vector and the candidate $f_{j,q}$. Thus, the Weighted Euclidean Deviation is described as

$$d_{j,q} = \|(V_j - T) \odot U\|_2, \tag{15}$$

where $T = [1, 2, 3, 4, 5]$ denotes the multiple harmonics of pitch, $U_i = 1/T_i$ and $\odot$ is the point-wise multiplication. The Weighted Comb Filtering for $f_{j,q}$ is defined as

$$y_{j,q} = \sum_f X_q(f) C\left(f/f_{j,q}\right), \tag{16}$$

where $X_q(f)$ is the power spectrum for the qth frame, and $C\left(f/f_{j,q}\right) = \left(\frac{1}{2} + \frac{1}{2}\cos\left(2\pi f/f_{j,q}\right) \exp\left(-f/f_{j,q}\right)\right)$. Finally, a

[1] [Online] Available: http://www.ee.ic.ac.uk/hp/staff/dmb/data/TIMITfxv.zip

TABLE I
LOW/HIGH FREQUENCY SEPARATION ERRORS (%) FOR THE WORST-CASE SCENARIO ASSUMPTION

		Babble					Cafeteria					SSN					Volvo					
	SNR(dB)	-15	-10	-5	0	5	-15	-10	-5	0	5	-15	-10	-5	0	5	-15	-10	-5	0	5	Avg.
CSTR	DCNN-BPS	42.3	37.1	30.3	24.3	19.5	44.1	36.6	29.3	22.3	18.6	45.1	39.8	32.1	24.2	18.2	22.9	18.5	16.3	15.8	16.0	27.7
	FSFFE	**39.7**	**25.1**	**14.5**	**7.7**	**5.0**	**36.2**	**25.0**	**13.8**	**7.8**	**4.9**	**42.9**	**31.0**	**15.7**	**7.6**	**3.8**	**2.5**	**2.8**	**2.6**	**2.9**	**3.0**	**14.7**
TIMIT	DCNN-BPS	26.2	19.6	17.3	15.3	14.0	26.3	19.9	17.5	15.2	14.0	26.3	20.1	17.7	15.4	14.3	18.9	15.5	13.7	12.7	12.6	17.0
	FSFFE	**26.1**	**18.5**	**12.7**	**8.4**	**6.2**	**25.2**	**18.8**	**13.0**	**8.9**	**6.2**	**23.7**	**17.0**	**10.5**	**7.5**	**5.6**	**5.4**	**5.5**	**4.7**	**4.0**	**3.7**	**11.6**

cost function is defined as

$$\text{cost}_q = |\log f_{j,q} - \log f_{i,q+1}| + \frac{\lambda}{pr_q\left(\frac{y_{j,q}}{\alpha} + \frac{1}{d_{j,q}} + \varepsilon_q\right)}, \tag{17}$$

where $|\log f_{j,q} - \log f_{i,q+1}|$ is a F0 smoothness feature, $\lambda > 0$ and α are regularization parameters, pr_q is the low/high frequency probability of the softmax layer of DCNN, and $\varepsilon_q = 1$ if $f_j = f_p$, or zero otherwise. The smaller cost function value for the F0 candidates is more likely to be the true F0 value.

IV. Results and Discussion

This Section presents results for low/high F0 separation errors attained by FSFFE method in comparison to DCNN-BPS baseline technique. Following, GE and MAE metrics are adopted to evaluate the errors for the competitive approaches in F0 estimation improvement experiments, considering several noisy environments. The F0 estimation errors are evaluated for an ideal case such as in [30], in which the low/high F0 separation is perfect and generates no error in the whole system. Finally, this work performs an analysis of GE and MAE considering the worst-case scenario, where separation errors propagate, affecting the accuracy of the F0 estimates.

A. Speech and Noise Databases

The experiments consider the CSTR [32] and a subset of TIMIT [33] databases to evaluate the competitive methods. CSTR is composed of 100 English utterances spoken by male (50) and female (50) speakers, sampled at 20 KHz. The reference F0 values are available based on the recordings of laryngograph data. The TIMIT subset is composed of 128 speech signals spoken by 8 male and 8 female speakers, sampled at 16 KHz and with 3 s average duration. The reference F0 values are obtained from [43].

Six noises are used to corrupt the speech utterances: acoustic Babble and Volvo attained from RSG-10 [40], Cafeteria, Train and Helicopter from Freesound.org,[2] and Speech Shaped Noise (SSN) from DEMAND [44] database. Experiments are conducted considering noisy signals with five SNR values.

B. Evaluation Metrics

For the evaluation, it is adopted the Gross Error rate (GE) and Mean Absolute Error (MAE). GE [34] is defined as

$$GE = \frac{P_{error}}{P} \times 100, \tag{18}$$

[2][Online]. Available: https://freesound.org.

where P denotes the total number of voiced frames, and P_{error} is the number of voiced frames for which the deviation estimated F0 from the ground truth is more than 20%. MAE [35] is computed as

$$MAE = \frac{1}{Q}\sum_{q=1}^{Q}\left|\hat{F}0(q) - F0(q)\right|, \tag{19}$$

where Q is the total number of frames, $\hat{F}0(q)$ the estimate and $F0(q)$ is the reference. This metric provides a greater perception of the error, since it indicates an absolute distance (in Hz) between F0 reference and estimation.

C. Low/High Frequency Separation Accuracy

Table I presents the error results with the FSFFE and the competing DCNN-BPS technique for low/high frequency separation of voiced speech frames. These results denote the mean values for the utterances of CSTR and TIMIT databases, respectively, considering the speech signals corrupted by four noises and five SNR values. Note that FSFFE outperforms the DCNN-BPS for all of the noisy scenarios. The proposed method achieves the lowest error values in all the 20 noisy conditions and the two databases. For instance, it can be seen that FSFFE attains interesting results for the most severely conditions, for SNR $= -15$ dB. In this case, the error scores are on average 8.3 percentage points (p.p.) smaller than those achieved by DCNN-BPS approach in CSTR database. The overall errors obtained with FSFFE are 14.7% and 11.6% for CSTR and TIMIT databases, against 27.7% and 17.0% for DCNN-BPS, respectively.

The interesting accuracy results attained by FSFFE method are particularly important, and can be justified by the fact that it is considered only the first four IMFs of EEMD. This attenuates the noise masking effects, since the most part of noisy energy is concentrated in low frequencies, e.g., IMF_6, IMF_7 and IMF_8. Moreover, PEFAC also contributes to the separation accuracy, due to the filter introduced in the algorithm, which rejects high-level narrow-band noise and favors the correct F0 estimates, even in severe noisy environments.

D. GE and MAE Results

The proposed method and DCNN-BPS baseline technique for improvement of F0 detection accuracy are compared in terms of GE and MAE, considering three F0 estimation approaches: SHR, SWIPE and HHT-Amp. In this study, the voiced/unvoiced detection is based on the energy and Zero Crossing (ZC) rate associated with the signal waveform as described in [45].

Fig. 5 shows the GE results of the F0 estimates for speech signals of CSTR database assuming the ideal case, with no

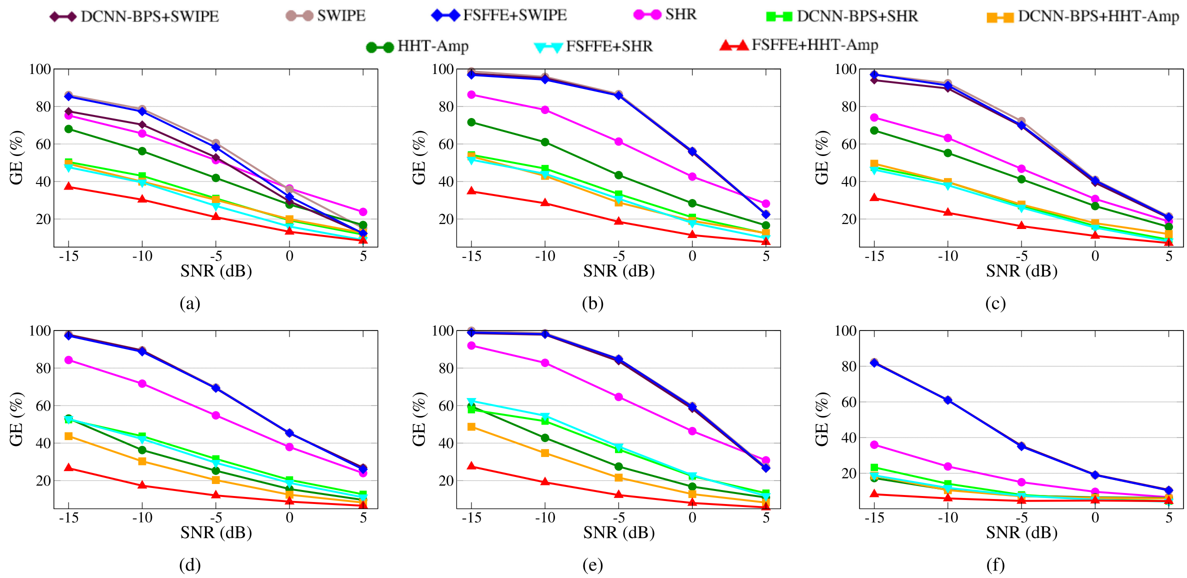

Fig. 5. The average GE with CSTR database considering six acoustic noises: (a) Babble, (b) SSN, (c) Cafeteria, (d) Train, (e) Helicopter, and (f) Volvo. Results are obtained assuming no errors in low/high F0 separation.

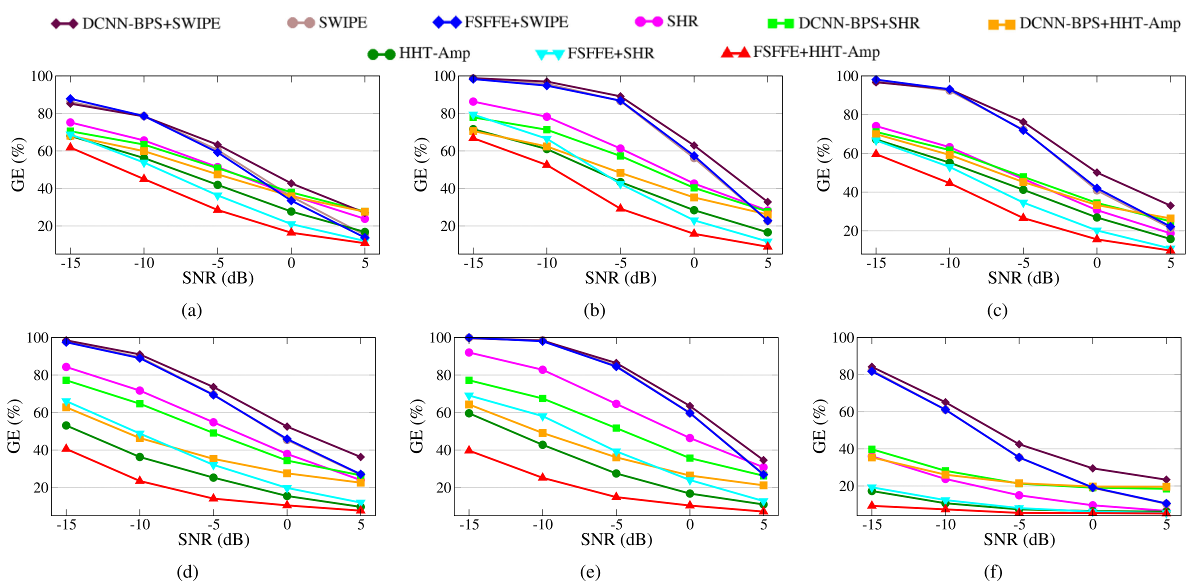

Fig. 6. The average GE with CSTR database considering six acoustic noises: (a) Babble, (b) SSN, (c) Cafeteria, (d) Train, (e) Helicopter, and (f) Volvo. F0 estimation is improved considering the low/high frequency separation errors (the worst-case scenario).

errors in the low/high frequency separation. The noisy scenarios are composed of six types of noises with SNR values between -15 dB and 5 dB. It may be observed that both separation methods are able to reduce GE errors from the F0 estimation algorithms, even in severe noisy conditions. Note that the SHR F0 estimation technique attains a GE score of 54.8% for Train noise (Fig. 5(d)) with SNR $= -5$ dB, and its value is reduced to 31.6% for the composition DCNN-BPS+SHR. Furthermore, the FSFFE+HHT-Amp combination presents the lowest scores in most of the noisy scenarios, outperforming the competitive compositions in terms of F0 estimation accuracy.

Fig. 6 depicts the GE results for the F0 estimation approaches improved by the FSFFE and DCNN-BPS solutions for the CSTR database, assuming that separation errors propagate to whole system (the worst-case scenario). It is interesting to mention that the FSFFE+HHT-Amp composition achieves the lowest error results, being the most accurate method when compared to competitive solutions. In addition, note that FSFFE method attains interesting F0 improvement for SHR and HHT-Amp F0 estimators, even assuming the low/high frequency separation errors. For instance, the HHT-Amp approach attained a GE of 59.6% for Helicopter noise (Fig. 6(e)) with SNR $= -15$ dB, in contrast with 39.7% of FSFFE+HHT-Amp scheme. For the SHR method, DCNN-BPS improved the F0 estimates in some cases, e.g., Train and Helicopter noises, but it is still surpassed by FSFFE. However, for SWIPE and HHT-Amp, the

TABLE II
GROSS ERROR (GE) RESULTS (%) WITH THE PROPOSED AND BASELINE METHOD FOR TIMIT DATABASE

		DCNN-BPS			FSFFE		
Noise	SNR	SHR	SWIPE	HHT-Amp	SHR	SWIPE	HHT-Amp
Babble	-15 dB	71.2	85.9	65.9	69.5	84.7	**57.3**
	-10 dB	66.3	83.7	57.9	60.3	81.7	**44.1**
	-5 dB	57.0	78.5	46.0	46.7	74.9	**29.5**
	0 dB	45.9	64.6	36.1	32.4	56.5	**16.3**
	5 dB	36.6	48.0	28.8	20.7	34.8	**9.4**
	Average	55.4	72.2	46.9	45.9	66.5	**31.3**
SSN	-15 dB	82.1	98.9	70.1	81.9	98.0	**64.4**
	-10 dB	78.9	98.7	60.6	71.8	97.6	**48.5**
	-5 dB	67.5	97.0	47.5	53.2	94.8	**27.6**
	0 dB	52.8	89.1	35.6	34.4	83.4	**15.5**
	5 dB	40.2	71.8	28.1	21.1	59.3	**8.3**
	Average	64.3	91.1	48.4	52.5	86.6	**32.9**
Cafeteria	-15 dB	70.9	97.4	64.5	66.9	98.7	**56.8**
	-10 dB	65.7	97.3	56.1	57.1	97.8	**39.8**
	-5 dB	56.5	93.1	45.5	43.4	92.0	**25.0**
	0 dB	44.5	81.0	35.1	29.1	73.9	**15.3**
	5 dB	35.0	62.8	27.7	18.4	49.8	**8.4**
	Average	54.5	86.3	45.8	43.0	82.4	**29.1**
Train	-15 dB	82.3	99.4	56.2	70.4	99.0	**37.9**
	-10 dB	72.6	96.1	46.5	56.3	94.8	**24.9**
	-5 dB	60.6	89.5	36.4	42.3	84.1	**14.5**
	0 dB	48.7	79.7	29.9	30.1	71.4	**10.3**
	5 dB	38.9	67.7	26.3	20.2	56.2	**8.1**
	Average	60.6	86.5	39.0	43.9	81.1	**19.1**
Helicopter	-15 dB	86.0	99.8	63.3	71.4	99.9	**42.9**
	-10 dB	80.5	99.5	49.5	63.0	99.9	**27.2**
	-5 dB	67.0	97.4	37.6	48.4	97.3	**17.1**
	0 dB	53.3	90.6	29.7	32.7	87.3	**9.7**
	5 dB	41.9	76.9	24.5	20.8	65.0	**7.1**
	Average	65.7	92.8	40.9	47.3	89.9	**20.8**
Volvo	-15 dB	49.3	95.7	33.9	30.9	95.1	**7.9**
	-10 dB	39.7	90.7	27.8	20.7	84.7	**7.0**
	-5 dB	31.9	78.0	24.1	13.5	65.9	**5.9**
	0 dB	27.0	58.7	22.2	9.2	42.4	**4.8**
	5 dB	24.1	44.2	21.8	6.6	26.2	**4.3**
	Average	34.4	73.5	26.0	16.2	62.8	**6.0**
Overall		55.8	83.7	41.2	41.4	78.2	**23.2**

errors of DCNN-BPS separation affect whole system, causing an increasing in the GE scores. Note that, in severe noisy condition, the SWIPE method obtained the highest GE values. This fact may be explained by its inner voiced/unvoiced detection that is impaired by the background situation.

Table II presents GE results of the F0 estimation techniques combined with FSFFE and DCNN-BPS for the TIMIT database. Note that FSFFE outperforms DCNN-BPS for SHR and HHT-Amp detectors. For example, considering the Helicopter noise with SNR $= -15$ dB, the GE for the SHR F0 estimation scheme decreased from 86.0% with DCNN-BPS [30] to 71.4% with FSFFE, i.e., a reduction of 14.6 p.p. Likewise the results exposed for CSTR, FSFFE+HHT-Amp achieves the best accuracy results for TIMIT database, with lowest GE values in all the 30 noisy conditions. For the Babble noise, this strategy attains the average GE of 31.3% against 46.9% for FSFFE+SHR. On overall average, FSFFE+HHT-Amp presents a GE score of 23.2%, which is 18.0 p.p., 18.2 p.p and 32.6 p.p smaller than

TABLE III
OVERALL AVERAGE GE RESULTS (%)

Separation	F0 Estimator	CSTR No Errors	CSTR With Errors	TIMIT No Errors	TIMIT With Errors
DCNN-BPS	SHR	28.7	46.7	39.9	55.8
	SWIPE	60.4	66.9	78.0	83.7
	HHT-Amp	24.8	41.0	24.7	41.2
FSFFE	SHR	27.3	34.5	36.1	41.4
	SWIPE	61.5	62.2	61.5	78.2
	HHT-Amp	**15.7**	**23.9**	**17.1**	**23.2**

TABLE IV
NORMALIZED MEAN PROCESSING TIME

F0 Estimator			DCNN-BPS			FSFFE		
SHR	SWIPE	HHT-Amp	SHR	SWIPE	HHT-Amp	SHR	SWIPE	HHT-Amp
0.03	0.03	0.93	0.22	0.20	1.07	0.94	0.95	1.00

DCNN-BPS+HHT-Amp, FSFFE+SHR and DCNN-BPS+SHR approaches, respectively.

Table III presents the overall GE results for F0 estimation experiments with CSTR and TIMIT databases, averaged over the six noises and five SNR values. The SHR, SWIPE and HHT-Amp estimation methods attain average GE values of 48.7%, 62.3% and 32.8% for CSTR, and 51.4%, 78.9% and 26.3% for TIMIT database, respectively. Note that for the worst-case scenario, FSFFE method outperforms DCNN-BPS, providing an interesting improvement in the F0 estimation accuracy. Furthermore, the proposed FSFFE method combined with the HHT-Amp F0 estimator outperforms other composition techniques, with GE of 15.7% and 17.1% in the ideal scenario.

Figs. 7 and 8 illustrate the MAE results obtained for six noisy conditions with the CSTR and TIMIT databases, respectively. Each box-plot denotes the MAE values achieved with five SNR values: -15 dB, -10 dB, -5 dB, 0 dB and 5 dB, assuming the worst-case scenario. Note that again the proposed method achieved the lowest MAE values. For CSTR database, the DCNN-BPS strategy attains improvement in the F0 estimates in some cases of SHR and SWIPE approaches, particularly for SSN, Cafeteria, and Helicopter noises. The MAE values for the TIMIT database indicate that FSFFE improved the F0 estimation accuracy, overcoming the DCNN-BPS baseline solution. Once again, the FSFFE method leads the HHT-Amp estimator to achieve best accuracy results for both databases, with exception for Volvo noise with TIMIT (Fig. 6(f)).

In summary, the proposed method leads to the best results when compared to the solution DCNN-BPS, in terms of low/high frequency separation accuracy, and improvement of F0 estimation accuracy. In addition, the combination of FSFFE with the HHT-Amp algorithm attained the lowest GE and MAE scores, for both the CSTR and TIMIT databases. Furthermore, SHR approach are also outperformed by the HHT-Amp. The fact that SHR adopts two F0 candidates, against the three candidates considered in HHT-Amp, can favor this last one in the true F0 selection. In contrast, competitive techniques based on the SWIPE estimation reach the highest error values. This occurs because SWIPE is very sensitive to the severe noise disturbances, which

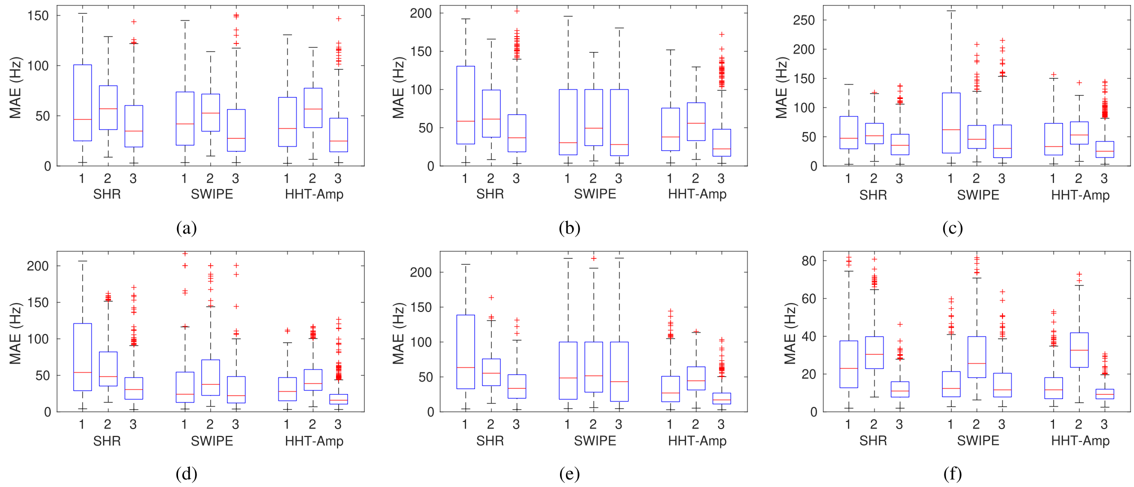

Fig. 7. The MAE box-plots for CSTR database speech signals with five SNR values: −15 dB, −10 dB, −5 dB, 0 dB and 5 dB; and for six noisy conditions: (a) Babble, (b) SSN, (c) Cafeteria, (d) Train, (e) Helicopter, and (f) Volvo. Case 1 refers to F0 estimation method, case 2 the improved F0 by DCNN-BPS and case 3 improved F0 by FSFFE. Results obtained considering the worst-case scenario, with low/high frequency separation errors.

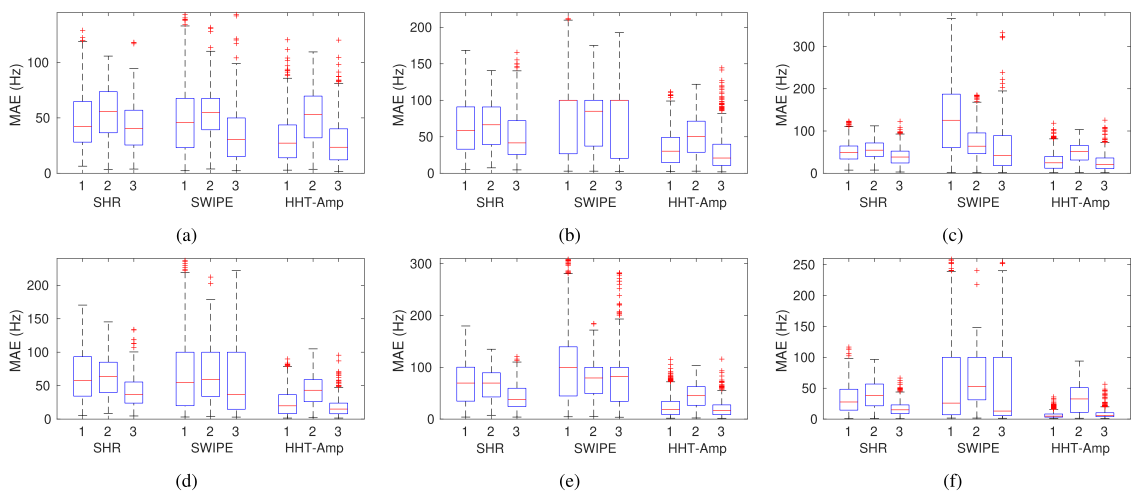

Fig. 8. The MAE box-plots for TIMIT database speech signals with five SNR values: −15 dB, −10 dB, −5 dB, 0 dB and 5 dB; and for six noisy conditions: (a) Babble, (b) SSN, (c) Cafeteria, (d) Train, (e) Helicopter, and (f) Volvo. Case 1 refers to F0 estimation method, case 2 the improved F0 by DCNN-BPS and case 3 improved F0 by FSFFE. Results obtained considering the worst-case scenario, with low/high frequency separation errors.

may cause expressive voiced/unvoiced detection errors. Finally, it is interesting to observe that even considering the classification errors in the estimation accuracy results, the proposed method improves F0 estimation of speech signals corrupted by different types of noise, and hence it can be used in real world applications.

Table IV indicates the computational complexity which refers to the normalized processing time required for each scheme evaluated for 512 samples per frame. These values are obtained with an Intel (R) Core (TM) i7-9700 CPU, 8 GB RAM, and are normalized by the execution time of the most accurate method FSFFE+HHT-Amp. The processing time required for the training process by DCNN is not considered here. Therefore, note that the HHT-Amp scheme and proposed method present a longer processing time, since they are based on the EEMD, and demands a relevant computational cost.

V. Conclusion

This article introduced the FSFFE method for low/high frequency separation of voiced frames, in order to improve the F0 estimation accuracy. The EEMD algorithm was applied to decompose the noisy speech signal. Then, PEFAC estimation and selection of analyzed decomposition modes was performed to detect the low or high frequency region of the voiced frames. According to this separation, the candidates of F0 estimation techniques were corrected, improving their accuracy. Extensive experiments were conducted to evaluate the improvement provided by the FSFFE method and the DCNN-BPS approach, with three F0 estimation techniques. Two speech databases and six acoustic noises of different sources were adopted for this purpose. Results were examined for an ideal scenario with no errors in low/high F0 separation, and for the worst-case scenario, considering the separation errors. The proposed method outperformed the DCNN-BPS in terms of low/high frequency separation in all cases. Furthermore, results obtained for the F0 estimation demonstrated that, for the worst-case scenario, FSFFE leads to superior improvement in F0 estimation accuracy. Particularly, the FSFFE+HHT-Amp combination outperformed the competitive approaches, with interesting accuracy in F0 detection in various noisy environments. Future research includes the investigation of the low/high frequency separation in other tasks, such as for speech intelligibility improvement.

REFERENCES

[1] L. Geurts and J. Wouters, "Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants," *J. Acoust. Soc. Amer.*, vol. 109, no. 2, pp. 713–726, Feb. 2001.

[2] K. N. Ross and M. Ostendorf, "A dynamical system model for generating fundamental frequency for speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 295–309, May 1999.

[3] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 2, pp. 176–182, Apr. 1975.

[4] A. Ljolj, "Speech recognition using fundamental frequency and voicing in acoustic modeling," in *Proc. Int. Conf. Spoken Lang. Process*., 2002, pp. 2137–2140.

[5] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: Tracking nonstationary noises during speech," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 437–440.

[6] L. Wang and F. Chen, "Factors affecting the intelligibility of low-pass filtered speech," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 563–566.

[7] A. Queiroz and R. Coelho, "F0-based gammatone filtering for intelligibility gain of acoustic noisy signals," *IEEE Signal Process. Lett.*, vol. 28, pp. 1225–1229, May 2021.

[8] H. Traunmuller and A. Eriksson, "The perceptual evaluation of F0 excursions in speech as evidenced in liveliness estimations," *J. Acoust. Soc. Amer.*, vol. 97, no. 3, pp. 1905–1915, 1995.

[9] C. Brown and S. Bacon, "Fundamental frequency and speech intelligibility in background noise," *Hear. Res.*, vol. 266, pp. 52–59, 2010.

[10] L. Wang, D. Zheng, and F. Chen, "Understanding low-pass-filtered mandarin sentences: Effects of fundamental frequency contour and single-channel noise suppression," *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. 141–145, 2018.

[11] J. Dubnowski, R. Schafer, and L. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 24, no. 1, pp. 2–8, Feb. 1976.

[12] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 25, no. 1, pp. 24–33, Feb. 1977.

[13] D. Talkin, W. R. Kleiin, and K. K. Paliwal, "A robust algorithm for pitch tracking," in *Speech Coding and Synthsis*, Amsterdam, Netherlands: Elsevier, 1995, pp. 497–518.

[14] 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, 2002.

[15] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. I333–I336.

[16] 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, 2008.

[17] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. 19th Eur. Signal Process. Conf.*, 2011, pp. 451–455.

[18] N. Yang, H. Ba, W. Cai, I. Demirkol, and W. Heinzelman, "BaNa: A noise resilient fundamental frequency detection algorithm for speech and music," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1833–1848, Dec. 2014.

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

[20] Y. Liu, J. Tao, D. Zhang, and Y. Zheng, "A novel pitch extraction based on jointly trained deep BLSTM recurrent neural networks with bottleneck features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 336–340.

[21] T. Drugman, G. Huybrechts, V. Klimkov, and A. Moinet, "Traditional machine learning for pitch detection," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1745–1749, Nov. 2018.

[22] J. Kim, J. Salamon, P. Li, and J. Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 161–165.

[23] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirovic, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1118–1128, Mar. 2020.

[24] 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.

[25] L. Zão and R. Coelho, "On the estimation of fundamental frequency from nonstationary noisy speech signals based on Hilbert–Huang transform," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 248–252, Feb. 2018.

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

[27] 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. 897–909, May 2014.

[28] 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, ch. 5, pp. 123–153.

[29] A. Stallone, A. Cicone, and M. Materassi, "New insights and best practices for the successful use of empirical mode decomposition, iterative filtering and derived algorithms," *Nat. Sci. Rep.*, vol. 10, no. 15161, pp. 1–15, 2020.

[30] M. Khadem-hosseini, S. Ghaemmaghami, A. Abtahi, S. Gazor, and F. Marvasti, "Error correction in pitch detection using a deep learning based classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 990–999, Mar. 2020.

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

[32] P. Bagshaw, S. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1993, pp. 1003–1006.

[33] J. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus," in *Linguist. Data Consortium*, Philadelphia, PA, USA, 1993.

[34] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative study of several pitch detection algorithms," *IEEE/ACM Trans. Acout., Speech, Signal Process.*, vol. 24, no. 5, pp. 399–418, Oct. 1976.

[35] C. Willmott et al., "Statistics for the evaluation and comparison of models," *J. Geophysical Res.*, vol. 90, pp. 8995–9005, Sep. 1985.

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

[37] N. Huang, "Introduction to the hilbert-huang transform and its related mathematical problems," in *Hilbert-Huang Transform and its N. Applications Huang and S. Shen*, Eds., Singapore: World Sci. Publishing, 2014.

[38] C. Medina, R. Coelho, and L. Zão, "Impulsive noise detection for speech enhancement in HHT domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2244–2253, Jun. 2021.

[39] N. Chatlani and J. Soraghan, "EMD-based filtering (EMDF) of low-frequency noise for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1158–1166, May 2012.

[40] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise-database," *Rep. IZF*, vol. 3, pp. 1–12, 1988.

[41] I. R. Titze, *Principles of Voice Production*. Englewood Cliffs, NJ, USA: Prentice Hall, 1994.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*., 2015.

[43] S. Gonzalez, "Pitch of the core TIMIT database set," 2014.

[44] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust.*, 2013, pp. 1–6.

[45] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *Proc. ASEE Reg. Conf.*, 2008, pp. 1–7.