F0-Based Gammatone Filtering for Intelligibility Gain of Acoustic Noisy Signals

A. Queiroz, *Student Member, IEEE*, and R. Coelho, *Senior Member, IEEE*

***Abstract*—This letter proposes a time-domain method to improve speech intelligibility in noisy scenarios. In the proposed approach, a series of Gammatone filters are adopted to detect the harmonic components of speech. The filters outputs are amplified to emphasize the first harmonics, reducing the masking effects of acoustic noises. The proposed GTF$_{\mathbf{F0}}$ solution and two baseline techniques are examined considering four background noises with different non-stationarity degrees. Three intelligibility measures (ESTOI, ESII and ASII$_{\mathbf{ST}}$) are adopted for objective evaluation. The experiments results show that the proposed scheme leads to expressive speech intelligibility gain when compared to the competing approaches. Furthermore, the PESQ and OQCM objective scores demonstrate that the proposed technique also provides interesting quality improvement.**

***Index Terms*—Non-stationary noises, Gammatone filtering, intelligibility improvement.**

I. Introduction

Acoustic noise masking effect is a crucial cause of impairment and a key challenge for speech intelligibility improvement research area. This issue underlies many applications such as speech synthesis, source localization, robot audition, and speech and speaker recognition. A diversity of speech enhancement approaches is described in the literature to mitigate this interference outcome with interesting speech quality improvement [1]–[5]. However, this achievement does not necessarily leads to speech intelligibility improvement [6]. The investigation of harmonic noisy speech signals has gained significant attraction [7], [8] for the proposal of strategies to achieve intelligibility gain. Moreover, harmonic components such as fundamental frequency (F0), or pitch, and formants play a significant role for speech intelligibility in noise [9]–[12].

Recently, time-domain adaptive solutions have been designed to deal with the harmonics of the speech signal to reduce the noise effects. In [13], the formant center frequencies from voiced segments of speech are shifted away from the region of noise. This formant shifting procedure [14] simulates the human strategy to provide a more audible signal in noisy environment, i.e., the Lombard effect. Results showed that the Smoothed Shifting of Formants for Voiced segments (SSFV) is able to improve the intelligibility of speech signals in car noise environment. A different approach was proposed in [15], where linear harmonic models are applied to represent the voiced segments as sum of sinusoids. Each voiced frame is reconstructed as sum of harmonics whose frequencies correspond to the speech F0 and its first integer multiples. The amplitude and phase estimation filter [16] was applied with the harmonic models (APES$_{\text{HARM}}$) and led to improved signal-to-noise ratio (SNR) of the reconstructed speech signals [15].

The main objective of this Letter is to provide intelligibility gain to speech signals corrupted by non-stationary acoustic noises. The proposed solution, namely GTF$_{\text{F0}}$, first applies the HHT-Amp F0 estimation method [17] to the harmonic/voiced segments. Integer multiples of the estimated F0 are used as center frequencies of a time-domain Gammatone filterbank. Following, the outputs are amplified by a gain factor to emphasize the harmonics of the speech signal leading to intelligibility improvement. It is worth to mention that the proposed GTF$_{\text{F0}}$ requires no prior knowledge of the noise statistics which means that it is suitable to any kind of noisy environment.

Extensive experiments are conducted to evaluate the proposed scheme for speech intelligibility and quality improvement. For this purpose, four acoustic noises with different non-stationarity degrees are used to corrupt the speech signals considering five SNR values. The formant shifting approach (SSFV) [13], and the technique based on harmonic models (APES$_{\text{HARM}}$) [15] are adopted as baseline solutions. Three objective intelligibility measures are used to evaluate the proposed and baseline methods: ESTOI [18], ESII [19] and ASII$_{\text{ST}}$ [20]. PESQ [21] and OQCM [22] are selected to examine the speech quality. Results show that the proposed solution outperforms the competitive methods in terms of speech intelligibility, and also quality scores.

II. F0 Estimation in Non-Stationary Noisy Scenario

In urban environments, speech signals are usually distorted by acoustic background noises. Particularly, the F0 estimation accuracy can be highly affected by the presence of acoustic noises. This task may become even more challenging when the background noise is non-stationary [7], [15], [17].

A. Non-Stationarity of Noisy Speech Signals

The non-stationarity degrees of speech signals corrupted by acoustic noises are here examined according to the Index of Non-Stationarity (INS) [23]. The INS objectively compares the target signal with stationary references called *surrogates*. A stationarity threshold $\gamma \approx 1$ is defined for each window length T_h considering a confidence degree of 95%. The INS is computed for different time scales T_h/T, where T refers to the total duration of the analyzed signal. Therefore, the signal is considered non-stationary whenever INS $> \gamma$.

Fig. 1 depicts the INS values obtained for a clean speech signal and four noisy versions with SNR of 0 dB. The Babble and Volvo noises are

Manuscript received March 22, 2021; revised May 14, 2021; accepted May 14, 2021. Date of publication May 28, 2021; date of current version June 24, 2021. This work was supported in part by the National Council for Scientific and Technological Development (CNPq) 308155/2019 and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) under Grant 203075/2016 and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - under Grant Code 001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao. *(Corresponding author: R. Coelho.)*

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

Digital Object Identifier 10.1109/LSP.2021.3084561

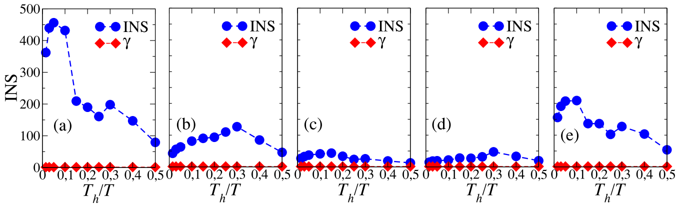


Fig. 1. INS values of (a) clean speech signal and four noisy versions with (b) Babble, (c) Cafeteria, (d) SSN, and (e) Volvo noises with SNR of 0 dB.

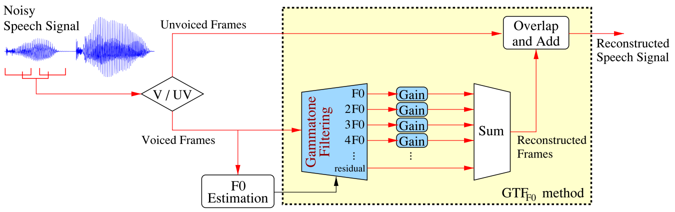


Fig. 2. Block diagram of the proposed GTF$_{\mathrm{F0}}$ method for speech intelligibility gain.

collected from the RSG-10 [25] database, while the speech shaped noise (SSN) and Cafeteria are respectively selected from the DEMAND [24] and Freesound.org databases to corrupt the speech signals. Note from Fig. 1(a) that the INS values reflect the highly non-stationary nature of the speech signal. Nevertheless, the noise interference considerably attenuates the non-stationary behavior of the clean speech signal. For instance, the maximum INS value (INS$_{\mathrm{max}}$) changes from 450 with clean speech to less than 100 when corrupted by the Cafeteria and SSN noises. These masking effects can modify the signal harmonic components (F0 and formants), which makes the F0 estimation in noise a challenging task.

B. HHT-Amp Estimation

The HHT-Amp estimator applies the Hilbert-Huang transform (HHT) to analyze the target speech signal. Instead of using the instantaneous frequencies as in [11], the F0 is estimated from the instantaneous amplitude functions of the target signal. Let $x(t)$ denote a speech signal divided into Q short-time frames $x_q(t), q = 1, 2, \ldots, Q$. The algorithm is summarized as follows:

1) For each sample sequence $x_q(t), q = 1, 2, \ldots, Q$, apply the ensemble empirical mode decomposition (EEMD) [26] and the Hilbert transform to define a series of K instantaneous amplitude functions $A_{k,q}(t), k = 1, \ldots, K$.
2) Compute the autocorrelation function (ACF) $r_{k,q}(\tau) = \sum_t A_k(t)\, A_k(t+\tau), k = 1, \ldots, K$, of the amplitude functions. Let τ_0 be the lowest τ value that correspond to an ACF peak of $r_{k,q}(\tau)$, subject to $\tau_{min} \leq \tau_0 \leq \tau_{max}$. The restriction is applied according to the range $[F_{min}, F_{max}]$ of possible F0 values. The k-th pitch candidate of frame q is defined as $P_{k,q} = \tau_0 / f_s$, where f_s refers to the sampling rate.
3) Select the true pitch value with error reduction using the following iterative post-processing procedure. The estimated pitch $\hat{T}_0(q)$ of each frame q is initially set to the first candidate $P_{1,q}$. Each $\hat{T}_0(q), q = 1, \ldots, Q$, is then compared to the average pitch value $m(q)$ computed from adjacent frames. The estimated pitch is updated to the next candidate whenever the value of $\hat{T}_0(q)$ deviates from $m(q)$ in more than 20% within a 40 ms interval. The estimated pitch is set to $m(q)$ if there is no candidate left for the corresponding frame.

In [17], it is shown that the HHT-Amp algorithm outperforms four competing estimators in different non-stationary noise scenarios. Moreover, the HHT-Amp iterative post-processing procedure guarantees the smoothness of the estimated pitch value. This is particularly important to the filtering method described in Section III since the speech signal harmonics are defined and amplified according to the estimated pitch of each frame.

III. Proposed GTF$_{\mathrm{F0}}$ Method

The block diagram of the GTF$_{\mathrm{F0}}$ method is exhibited in Fig. 2. The target noisy signal $x(t)$ is first split into Q overlapping short-time frames $x_q(t), q = 1, 2, \ldots, Q$, with 50% overlapping. In this work, it is assumed that the separation of voiced and unvoiced (V/UV) segments was previously applied to define two disjoint sets. S_v is formed by frames that contain voiced speech, and S_u consists of the unvoiced and noise-only segments. For each $q \in S_v$, the HHT-Amp method [17] is applied to estimate the F0 value from $x_q(t)$. A total of L Gammatone filters, with center frequencies set to $\hat{\mathrm{F0}}$, $2\,\hat{\mathrm{F0}}$, ..., $L\,\hat{\mathrm{F0}}$, are used to filter the sample sequence $x_q(t)$. Gain factors are employed to amplify the filters outputs before the reconstruction of the speech frame $\hat{x}_q(t)$. Finally, the overlap and add method is applied to achieve the reconstructed version $\hat{x}(t)$ of the target speech signal.

A. Gammatone Filtering

The Gammatone filter was introduced in [27] to describe the impulse response of the auditory system. The time-domain impulse response of the Gammatone filter is defined as

$$g(t) = at^{n-1}\cos(2\pi f_c t + \phi)e^{-2\pi bt}\,,\; t \geq 0\,, \qquad (1)$$

where a is the amplitude, n is the filter order, f_c is the center frequency, ϕ is the phase, and b is the bandwidth. In [28], it was shown that a set of fourth-order Gammatone filters are able to represent the magnitude characteristic of the human auditory system. In the Gammatone auditory filterbank, the bandwidth b presented in (1) is similar to the Equivalent Rectangular Bandwidth (ERB) derived in [29], i.e., $b = 1.019\,\mathrm{ERB}$.

In the proposed method, a set of L Gammatone filters $\{h_k(t), k = 1\ldots, L\}$ are applied to successively filter the input sample sequence $x_q(t)$. Each filter $h_k(t)$ is implemented[1] considering order $n = 4$, center frequency $f_c = k\,\hat{\mathrm{F0}}$, and bandwidth $b = 0.25\,\hat{\mathrm{F0}}$. In order to align the impulse response functions, phase compensation is applied to all filters, which correspond to the non-causal filters

$$h_k(t) = a(t+t_c)^{n-1}\cos(2\pi f_c t)e^{-2\pi b(t+t_c)}\,,\; t \geq -t_c\,, \qquad (2)$$

where $t_c = \frac{n-1}{2\pi b}$ ensures that peaks of all filters occur at $t = 0$. Let $x_q^0(t) = x_q(t)$, the filtered signals $y_q^k(t), k = 1, \ldots, L$, are recursively computed by

$$\begin{cases} y_q^k(t) = x_q^{k-1}(t) * h_k(t) \\ x_q^k(t) = x_q^{k-1}(t) - y_q^k(t) \end{cases}, \quad k = 1, \ldots, L\,. \qquad (3)$$

The residual signal is defined as $r_q(t) = x_q^L(t)$.

B. Speech Signal Reconstruction

After the Gammatone filtering, the amplitude of the output samples $y_q^k(t), k = 1, \ldots, L$, are amplified by a gain factor $G_k \geq 1$. The idea

[1] [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/N.Ma/

TABLE I
GROSS ERROR [%] ACHIEVED BY HHT-AMP AND COMPETING F0 ESTIMATORS

	Babble (INS$_{\max} = 34.6$)					Cafeteria (INS$_{\max} = 11.7$)					SSN (INS$_{\max} = 1.6$)					Volvo (INS$_{\max} = 0.9$)					Overall
SNR (dB)	-5	-3	0	3	5	-5	-3	0	3	5	-5	-3	0	3	5	-5	-3	0	3	5	Average
PRAAT	63.7	54.6	41.7	30.0	22.7	63.6	54.8	42.0	30.2	23.4	80.0	67.6	48.7	33.4	25.5	53.4	44.6	33.4	23.6	18.0	42.7
YIN	58.5	52.7	42.7	32.4	26.7	70.1	62.6	51.5	40.6	33.4	73.6	65.0	51.5	37.6	29.6	67.3	56.6	41.5	29.7	23.6	47.4
SWIPE	75.5	70.2	60.0	48.1	40.1	93.4	89.1	80.2	67.7	59.0	96.4	94.2	87.1	76.8	67.7	74.0	65.8	53.0	41.2	34.4	68.7
SFF	36.5	30.4	22.0	15.3	12.3	32.8	28.0	20.6	14.2	12.0	37.8	30.2	22.2	14.7	11.0	12.8	10.8	8.7	6.8	6.0	19.3
HHT-Amp	**28.8**	**23.7**	**16.6**	**11.1**	**9.1**	**24.9**	**20.1**	**13.6**	**8.8**	**6.9**	**30.3**	**23.8**	**15.7**	**10.5**	**7.7**	**3.3**	**3.0**	**2.2**	**1.9**	**1.9**	**13.2**

TABLE II
ESTOI, ESII, AND ASII$_{\mathrm{ST}}$ MEASURES [%] FOR UNP SPEECH SIGNALS

	ESTOI					ESII					ASII$_{\mathrm{ST}}$				
SNR (dB)	-5	-3	0	3	5	-5	-3	0	3	5	-5	-3	0	3	5
Babble	0.28	0.33	0.40	0.48	0.53	0.34	0.38	0.44	0.50	0.54	0.38	0.40	0.45	0.50	0.54
Cafeteria	0.30	0.35	0.43	0.51	0.57	0.36	0.39	0.45	0.52	0.56	0.39	0.41	0.46	0.51	0.55
SSN	0.28	0.33	0.40	0.47	0.53	0.31	0.34	0.40	0.46	0.50	0.35	0.37	0.42	0.47	0.50
Volvo	0.71	0.74	0.79	0.83	0.86	0.82	0.85	0.89	0.92	0.94	0.77	0.80	0.84	0.87	0.89

is to emphasize the presence of the first harmonics of the fundamental frequency. This will induce speech intelligibility improvement without introducing any noticeable distortion to the speech signal. The reconstruction of the voiced frame $q \in S_v$ leads to the sample sequence

$$\hat{x}_q(t) = \sum_{k=1}^{L} G_k \, y_q^k(t) + r_q(t) \, . \tag{4}$$

For the reconstruction of the entire speech signal, the voiced frames obtained in (4) and all the remaining frames in S_u are joined together keeping the original frames indices. Thus, all frames are overlap and added to reconstruct the modified version $\hat{x}(t)$ of the target speech signal. The completeness and continuity of $\hat{x}(t)$ is guaranteed by the adoption of the Hanning window that multiply all frames before the overlap and add method. This means that the reconstructed signal $\hat{x}(t)$ and the original signal $x(t)$ would be exactly the same if $G_k = 1$ for every $k \in \{1, \ldots, L\}$.

Fig. 3 illustrates an example[2] of the proposed GTF$_{\mathrm{F0}}$ to a speech signal selected from the TIMIT database [30]. The spectrogram of a clean speech segment and a noisy version are depicted in Figs. 3(a-b). The corrupted signal is obtained with SSN considering SNR of 0 dB. It can be noted that the presence of the acoustic noises clearly induce the F0 harmonics to blur, especially the first and second ones. The GTF$_{\mathrm{F0}}$ strategy is applied to the noisy signal considering frames of 32 ms and Gammatone filters bandwidth $b = 0.25\,\hat{\mathrm{F0}}$. The first $L = 4$ harmonics are amplified with the following gain factors: $G_1 = G_2 = 5.0\,\mathrm{dB}$, $G_3 = 4.0\,\mathrm{dB}$, and $G_4 = 2.5\,\mathrm{dB}$. These values were empirically determined considering the training subset of 72 speech signals of the TIMIT database defined in [31]. The resulting spectrogram is shown in Fig. 3(c). Note that harmonics are more prominent when compared to the noisy signal. This effect may reduce the impact of the acoustic noise to speech intelligibility.

IV. EXPERIMENTS AND RESULTS

Several evaluation experiments are conducted with the test subset defined in [31]. This set is composed of 192 speech signals from the TIMIT database [30] sampled at 16 kHz with 3 s average duration. The training and test subsets are independent in terms of speakers and different utterances content. V/UV segments and reference F0 values are obtained from [31]. The formant shifting approach (SSFV) considers the formant modification function that led to the best results in [14]. The harmonic models solution with the APES filter (APES$_{\mathrm{HARM}}$) is applied as described in [15]. Ideal V/UV separation is considered available for all experiments.

A. Comparison of F0 Estimation Methods

The HHT-Amp and four competing F0 estimators, namely PRAAT [32], YIN [33], SWIPE [34], and SFF [35], are here evaluated in terms of gross error (GE) [33]. For these experiments, the PRAAT software was set to estimate the F0 values using the ACF method. The YIN, SWIPE, and SFF methods are applied using the codes provided by the authors websites. The GE results are shown in Table III. It can be noted that the HHT-Amp leads to the lowest GE rates for all noisy scenarios. As expected, the noise sources have significant impact in the GE values. Considering SNR of 0 dB, the GE rate achieved by HHT-Amp varies from 16.6% with Babble to only 2.2% with Volvo. Additionally, the HHT-Amp yields to an overall GE reduction of 6.1 percentage points when compared to the SFF: from 19.3% to 13.2%. This reinforces the adoption of the HHT-Amp estimator for the proposed GTF$_{\mathrm{F0}}$.

B. Objective Intelligibility Evaluation

Table II presents the average ESTOI, ESII and ASII$_{\mathrm{ST}}$ scores obtained with the noisy unprocessed (UNP) speech signals. The intelligibility improvement achieved with the proposed and baseline solutions are depicted in Fig. 4. Note from the ESTOI results that the GTF$_{\mathrm{F0}}$ leads to the highest gain for all noisy scenarios. In average, it outperforms the SSFV approach in 10% for the Babble, Cafeteria and SSN noises. For the highly non-stationary Cafeteria noise, the proposed method attains an improvement of 10.1 at 0 dB, compared to 0.4 and -4.8 for the SSFV and APES$_{\mathrm{HARM}}$ techniques, respectively.

In terms of ESII and ASII$_{\mathrm{ST}}$ scores, it can be seen that the GTF$_{\mathrm{F0}}$ leads to the best results for three noise sources: Babble, Cafeteria and SSN. The only scenario where this solution does not achieve the highest rates is the Volvo noise. In this case, all approaches lead to negative intelligibility gain. It is due to the fact that the ESII and ASII$_{\mathrm{ST}}$ scores for Volvo are higher than 0.77 for the noisy signals (refer to Table II). The values are defined as very good intelligibility [36], [37]. Among all the scenarios, GTF$_{\mathrm{F0}}$ accomplishes the highest overall Δ ESII and ΔASII$_{\mathrm{ST}}$ of 8.4 and 6.6, respectively, for the non-stationary Babble noise with SNR of -3 dB. The APES$_{\mathrm{HARM}}$ baseline method is outperformed by GTF$_{\mathrm{F0}}$ and SSFV in all scenarios.

[2] Other noisy and processed examples are available at http://lasp.ime.eb.br/index.php?vPage=downloads.

TABLE III
PESQ Objective Scores for Noisy Conditions At Different SNRs

	Babble					Cafeteria					SSN					Volvo					Overall
SNR (dB)	-5	-3	0	3	5	-5	-3	0	3	5	-5	-3	0	3	5	-5	-3	0	3	5	Average
UNP	1.98	2.14	2.41	2.71	2.90	2.15	2.33	2.59	2.89	3.05	1.91	2.07	2.34	2.64	2.84	3.75	3.89	**4.08**	**4.25**	**4.35**	2.86
$\mathrm{GTF}_{\mathrm{F0}}$	**2.17**	**2.36**	**2.66**	**2.94**	**3.12**	**2.39**	**2.58**	**2.86**	**3.13**	**3.30**	**2.10**	**2.30**	**2.61**	**2.89**	**3.08**	**3.83**	**3.93**	4.06	4.17	4.23	**3.04**
SSFV	1.98	2.14	2.42	2.71	2.90	2.17	2.33	2.59	2.87	3.05	1.93	2.08	2.35	2.64	2.84	3.73	3.87	4.05	4.22	4.31	2.86
$\mathrm{APES}_{\mathrm{HARM}}$	2.01	2.18	2.47	2.75	2.91	2.17	2.35	2.62	2.89	3.05	1.95	2.14	2.44	2.72	2.90	3.36	3.47	3.64	3.77	3.84	2.78

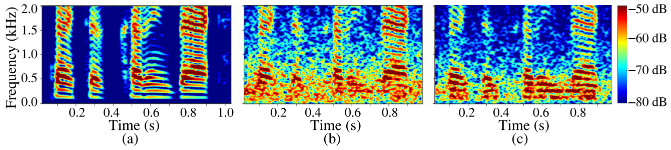
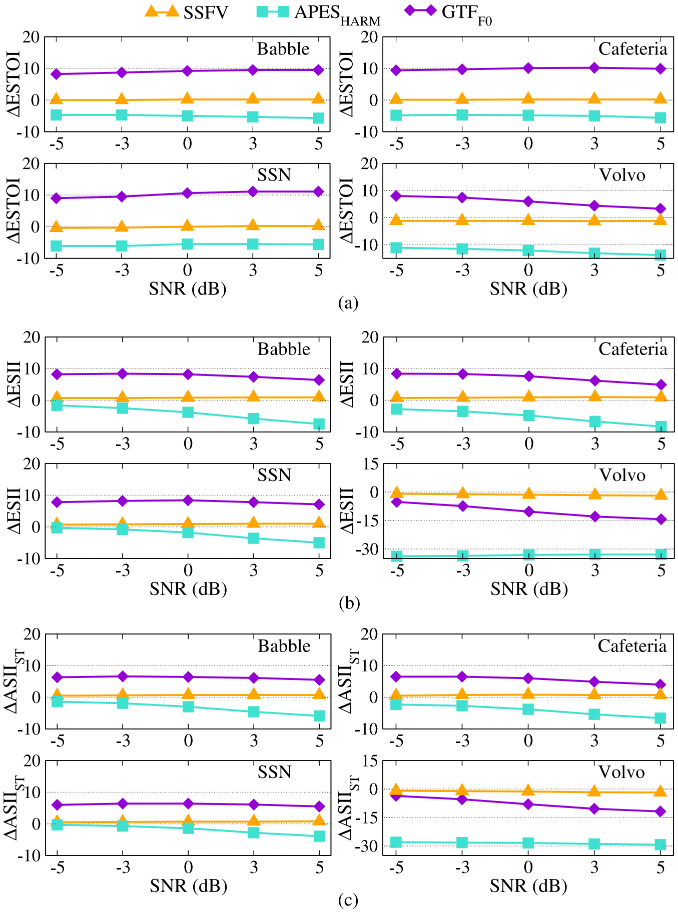


Fig. 3. Spectrogram of (a) a clean speech segment, (b) the same signal corrupted with SSN with SNR of 0 dB, and (c) the corresponding signal processed with the proposed $\mathrm{GTF}_{\mathrm{F0}}$ method.



(a)

(b)

(c)

Fig. 4. (a) ΔESTOI, (b) ΔESII, and (c) $\Delta\mathrm{ASII}_{\mathrm{ST}}$ intelligibility improvement $[\times 10^{-2}]$ in four noisy conditions.

C. Objective Quality Evaluation

The predicted quality scores computed with PESQ [21] are shown in Table III. As it can be seen, $\mathrm{GTF}_{\mathrm{F0}}$ attains the best PESQ results for three background noise sources: Babble, Cafeteria and SSN. Considering the Volvo noise, the unprocessed speech signals present good quality. It means that the highest PESQ scores are obtained by UNP with SNR $\geq$ 0 dB. The $\mathrm{GTF}_{\mathrm{F0}}$ attains the best average PESQ value of 3.06, which is 0.17 greater than the noisy signals result.

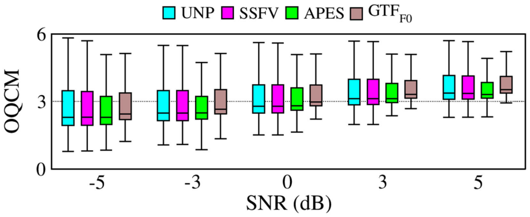


Fig. 5. Average OQCM scores for Babble, Cafeteria, SSN and Volvo noises.

The OQCM [22] is also adopted to objectively examine the speech signals in terms of quality. The OQCM is defined as $\mathrm{OQCM} = 1.594 + 0.805\,\mathrm{PESQ} - 0.512\,\mathrm{LLR} - 0.007\,\mathrm{WSS}$ to maximize the correlation with subjective scores in terms of overall speech quality. According to Fig. 5, $\mathrm{GTF}_{\mathrm{F0}}$ presents the greatest average OQCM values for all SNR levels. These results reinforce the capacity of the proposed solution to emphasize the harmonic components of speech signals, providing improvement in terms of both intelligibility and quality.

V. Conclusion

This letter introduced the time-domain $\mathrm{GTF}_{\mathrm{F0}}$ method to improve intelligibility of noisy speech signals. In this approach, F0 estimation and Gammatone filtering are applied to emphasize the first harmonics of the noisy speech signal. Four acoustic noises were considered to compose the evaluation scenario. Five objective prediction measures were applied to examine the proposed and competitive solutions. Results showed that $\mathrm{GTF}_{\mathrm{F0}}$ achieved the best intelligibility and quality scores considering ESTOI and PESQ prediction measures for all acoustic noises.

References

[1] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[2] T. Gerkmann and R. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[3] 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.

[4] 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.

[5] R. Tavares and R. Coelho, “Speech enhancement with nonstationary acoustic noise in time domain,” *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 6–10, Jan. 2016.

[6] P. Loizou and G. Kim, “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.

[7] 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.

[8] L. Wang and F. Chen, "Factors affecting the intelligibility of low-pass filtered speech," in *Proc. INTERSPEECH*, Aug. 2017, pp. 563–566.

[9] Y. Lu and M. Cooke, "The contribuition of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Commun.*, vol. 51, pp. 1253–1262, 2009.

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

[11] 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.

[12] J. Chen, H. Yang, and X. Wu, "The effect of F0 contour on the intelligibility of speech in the presence of interfering sounds for Mandarin Chinese," *J. Acoust. Soc. Amer.*, vol. 143, no. 2, pp. 864–877, 2018.

[13] K. Nathwani, G. Richard, B. David, P. Prablanc, and V. Roussarie, "Speech intelligibility improvement in car noise environment by voice transformation," *Speech Commun.*, vol. 91, pp. 17–27, May 2017.

[14] K. Nathwani, M. Daniel, G. Richard, B. David, and V. Roussarie, "Formant shifting for speech intelligibility improvement in car noise environment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5375–5379.

[15] S. Norholm, J. Jensen, and M. Christensen, "Enhancement and noise statistics estimation for non-stationary voiced speech," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 4, pp. 645–658, Apr. 2016.

[16] P. Stoica, H. Li, and J. Li, "A new derivation of the APES filter," *IEEE Signal Process. Lett.*, vol. 6, no. 8, pp. 205–206, Aug. 1999.

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

[18] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[19] K. Rhebergen and N. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[20] R. Hendriks, J. Crespo, J. Jensen, and C. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 851–862, May 2015.

[21] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.

[22] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[23] 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.

[24] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: A collection of multichannel recordings of acoustic noise in diverse environments," *Proc. 21st Int. Congr. Acoust.*, Jun. 2013.

[25] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise database," TNO Inst. Percep., Soesterberg, The Netherlands, Tech. Rep. IZF 3, 1988.

[26] 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.

[27] P. Johannesma, "The pre-response stimulus ensemble of neuron in the cochlear nucleus," *Proc. Symp. Hear. Theory*, Jun. 1972, pp. 58–69.

[28] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Proc. 9th Int. Symp. Hear., Audit. Physiol. Percep.*, 1992, pp. 429–446.

[29] R. D. Patterson and B. C. J. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," in *Proc. Freq. Selectiv. Hear.*, 1986, pp. 123–177.

[30] S. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," Philadelphia, PA, USA: NASA STI/Recon, Tech. Rep. N, vol. 24, 1993.

[31] S. Gonzalez, "Pitch of the core timit database set," May 2014. [Online] Available: http://www.ee.ic.ac.uk/hp/staff/dmb/data/TIMITfxv.zip

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

[33] 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.

[34] 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.

[35] 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.

[36] American National Standard, *Methods for Calculation of the Speech Intelligibility Index*. New York, NY, USA: Amer. Nat. Standards Inst., 1997.

[37] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," *Proc. IEEE Int. Conf. Acoust., Speech Signal*, vol. 1, 2006, pp. 493–496.