Harmonic Detection From Noisy Speech With Auditory Frame Gain for Intelligibility Enhancement

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

***Abstract*—This paper introduces a novel (HDAG - Harmonic Detection for Auditory Gain) method for speech intelligibility enhancement in noisy scenarios. In the proposed scheme, a series of selective Gammachirp filters are adopted to emphasize the harmonic components of speech reducing the masking effects of acoustic noises. The fundamental frequency values are estimated by the HHT-Amp (Amplitude-based Hilbert Huang Transform) technique. Harmonic components estimated with low accuracy are detected and adjusted according to the FSFFE (Frequency Separation for Fundamental Frequency Estimation) low/high pitch separation. The central frequencies of the filterbank are defined considering the third-octave subbands which are best suited to cover the regions most relevant to intelligibility. Before signal reconstruction, the gammachirp filtered components are amplified by gain factors regulated by FSFFE classification. The proposed HDAG solution and three baseline techniques are examined considering six background noises with four signal-to-noise ratios. Three objective measures are adopted for the evaluation of speech intelligibility and quality. Several experiments are conducted to demonstrate that the proposed scheme achieves better speech intelligibility improvement when compared to the competing approaches. A perceptual listening test is further considered and corroborates the objective results.**

***Index Terms*—Gammachirp filtering, harmonic detection, low/high frequency separation, noisy speech.**

I. Introduction

ACOUSTIC noise is a strong masking effect that impairs speech intelligibility [1], [2]. This interference underlies several research studies such as speech enhancement [3], [4], [5], source localization [6], [7], robot audition [8], speech and speaker recognition [9], [10]. Thus, its mitigation is a relevant element of interest for intelligibility and quality enhancement. Several signal processing methods are described in the literature to attenuate noise interference for speech quality assessment [11]. However, this achievement not necessarily lead to speech intelligibility improvement [12]. On the other hand, acoustic masks [13], [14], [15] are defined to emulate the *cocktail party* effect. These solutions provide intelligibility enhancement for the target speech signal.

In the last years, the analysis of harmonic components of noisy speech [16], [17] has encouraged the proposal of new strategies for intelligibility gain [18], [19]. For these, harmonic components such as fundamental frequency (F0) and formants [20] play an interesting role in intelligibility in noisy conditions [16], [21], [22]. Time-domain adaptive solutions are designed to deal with the harmonics of the speech signal to reduce the noise effects. In [23], the formant center frequencies from voiced segments of speech are shifted away from the region of noise. This formant shifting procedure [24] simulates the human strategy to provide a more audible signal in a noisy environment, i.e., the Lombard effect [25]. Results showed that the Smoothed Shifting of Formants for Voiced segments (SSFV) can improve the intelligibility of speech signals in a car noise environment. A different approach was proposed in [26], where the HHT-Amp [27] F0 estimation technique was applied to the harmonic components of noisy speech. The F0-based Gammatone Filtering (GTF_{F0}) method considered integer multiples of the estimated F0 as center frequencies of a time-domain auditory filterbank. Finally, the outputs are amplified to emphasize the harmonics of the speech signal leading to intelligibility gain.

The use of the Gammatone filterbank in the GTF_{F0} method may be limited by the high-level masking effects [28]. To overcome this issue, the Gammachirp proposed in [29] produces a filter with an asymmetric amplitude spectrum. This auditory filter provides an interesting fit to various sets of noise masking data. The center frequencies of its filterbank must be well-defined considering the relevant ones for intelligibility. In this context, the Extended Short-Time Objective Intelligibility (ESTOI) [30] performs an evaluation of noisy speech in third-octave subbands. ESTOI also considers the temporal modulation frequencies relevant to speech intelligibility, whose values range from 1–12.5 Hz [31], [32], [33]. These subbands and frequency modulation ranges can assist in regulating the bandwidth of filterbanks to cover the harmonic components of speech most relevant for intelligibility.

This paper introduces the HDAG (Harmonic Detection with Auditory Gain) method to attain intelligibility enhancement for harmonic components of noisy speech signals. The proposed solution is performed in four steps. Initially, the HHT-Amp method [27] is applied to estimate the F0 of speech frames. In the second step, these frames are separated into low-pitch or high-pitch ones with the FSFFE [34] technique. The separation leads to the detection and adjustment of the F0 values according

Manuscript received 30 October 2023; revised 6 February 2024; accepted 8 April 2024. Date of publication 25 April 2024; date of current version 3 May 2024. This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under Grant 305488/2022-8, in part by the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) under Grant 200518/2023, and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) under Grant 001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jan Skoglund. *(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: anderson.queiroz@ime.eb.br; coelho@ime.eb.br).

Digital Object Identifier 10.1109/TASLP.2024.3393727

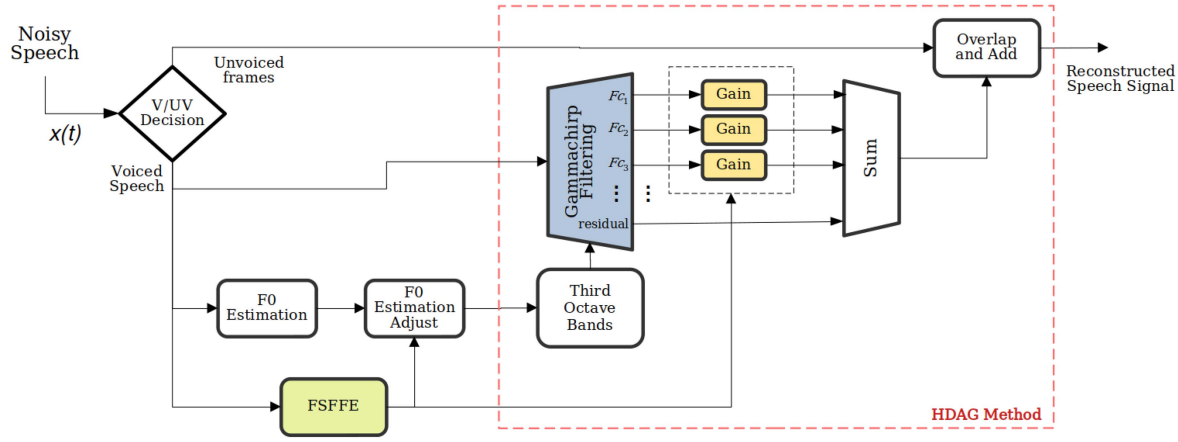

Fig. 1. Block diagram of the proposed HDAG method for improving the intelligibility of noisy speech signals.

to some typical errors [35] that may occur in estimation, improving its accuracy. In sequence, the third stage consists of filtering the harmonic components of noisy speech with Gammachirp. The central frequencies and bandwidths of the filterbank are selectively defined to cover the most relevant regions for speech intelligibility, as stated in [30]. Finally, the filtered components are amplified by a gain factor to highlight the harmonic components of speech. This amplification mitigates the masking effects of background noise leading to intelligibility enhancement.

Several experiments are conducted to examine the effectiveness of the HDAG method. For this purpose, speech utterances collected from the TIMIT [36] database are corrupted by six real acoustic noises, considering four SNR values: -10 dB, -5 dB, 0 dB, and 5 dB. The proposed method and three baseline approaches are examined in terms of intelligibility enhancement. To this end, ESTOI [30] and Short-Time Approximated Speech Intelligibility Index (ASII_{ST}) [37] are considered in the evaluation. Moreover, results for the Perceptual Evaluation of Speech Quality (PESQ) [38] demonstrate that HDAG also achieves quality assessment. Objective results indicate that the proposal outperforms the competitive approaches in terms of speech intelligibility, and also quality scores. These results are corroborated by a subjective listening evaluation test. The main contributions of this work are:

- Introduction of the HDAG method to improve the intelligibility and quality of acoustic noisy speech.
- Definition of the filterbank configuration using the third-octave bands and specific modulation frequencies, with higher resolution in regions most relevant to intelligibility.
- Adoption of the asymmetry coefficient from Gammachirp to adjust the filterbank to the noisy masked components of speech.
- Interesting intelligibility and quality assessment attained with adaptive gain factors defined according to FSFFE separation.

The remainder of this paper is organized as follows. Section II describes the steps of the proposed HDAG method for intelligibility enhancement. An explanation of the competitive approaches SSFV, PACO (pitch-adaptive complex-valued Kalman filter) [39] and GTF_{F0} is included in Section III. Section IV presents the evaluation experiments and results. Finally, Section V concludes this work.

II. The HDAG Method

The proposed method includes four main steps: harmonic detection, third-octave bands configuration, gammachirp filtering and output samples amplification by a gain factor. Finally, the overlap and add method is applied to achieve the reconstructed version of the target speech signal. Fig. 1 illustrates the block diagram of the HDAG method.

A. F0 Estimation

The fundamental frequency (F0) is estimated from noisy speech signals with the HHT-Amp method [27]. This F0 estimator ensures [27], [34] interesting accuracy results from noisy speech signals. HHT-Amp is evaluated in a wide range of noisy scenarios outperforming four competing estimators in terms of accuracy. It applies the time-frequency EEMD (Ensemble Empirical Mode Decomposition) [40], [41] to decompose a voiced sample sequence $x_q(t)$ such that

$$x_q(t) = \sum_{k=1}^{K} \text{IMF}_{k,q}(t) + r_q(t) \tag{1}$$

where $\text{IMF}_{k,q}(t)$ is the k-th mode of $x_q(t)$ and $r_q(t)$ is the last residual. Then, instantaneous amplitude functions are computed by

$$a_{k,q}(t) = |Z_{k,q}(t)|, k = 1, \ldots, K, \tag{2}$$

from the analytic signals defined as

$$Z_{k,q}(t) = \text{IMF}_{k,q}(t) + j\, H\{\text{IMF}_{k,q}(t)\}, \tag{3}$$

where $H\{\text{IMF}_{k,q}(t)\}$ refers to the Hilbert transform of $\text{IMF}_{k,q}(t)$. The Autocorrelation Function is calculated as

$$r_{k,q}(\tau) = \sum_{t} a_k(t)\, a_k(t+\tau). \tag{4}$$

For each decomposition mode k, let τ_0 be the lowest τ value that corresponds to an ACF peak, subject to $\tau_{\min} \leq \tau_0 \leq \tau_{\max}$.

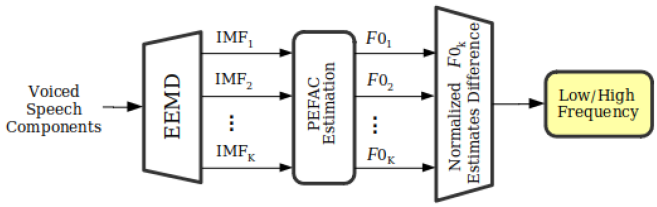

Fig. 2. Block diagram of the FSFFE technique for low/high pitch classification of speech frames.

The frequency restriction is applied according to the range $[F_{\min}, F_{\max}]$ of possible F0 values. The k-th F0 candidate is defined as τ_0/f_s, where f_s refers to the sampling rate. Finally, a decision criterion [27] is applied to select the best pitch candidate $\hat{T}_0$. Finally, the estimated F0 is given by $f_{est} = 1/\hat{T}_0$.

B. Harmonic Detection and Adjustment

Severe noise masking effects may impact the harmonic components of voiced speech leading to low accuracy F0 estimates. In order to detect and adjust the erroneous F0 values the FSFFE (Frequency Separation for Fundamental Frequency Estimation) [34] is applied to harmonic frames. This strategy separates the noisy speech frames into low-pitch or high-pitch ones. Possible errors in F0 estimates can be detected comparing its values with the separation. Fig. 2 illustrates the block diagram of the FSFFE method.

After the EEMD decomposition as in (1), pitch estimation is performed in voiced frames of each IMF using the PEFAC [42] algorithm. Let $\hat{F}0_{k,q}$ denote the pitch value estimated from frame q of $\text{IMF}_k(t)$, the $\hat{F}0_q$ vector is co mposed as

$$\hat{F}0_q = \left[\hat{F}0_{1,q}, \hat{F}0_{2,q}, \ldots, \hat{F}0_{K,q}\right]^T, \tag{5}$$

to express the tendency that the frame is placed in a low/high pitch region. Only the first four IMFs ($K = 4$) are considered to avoid the acoustic noise masking effect. The energy of these unwanted components is mostly concentrated at low frequencies ($K > 6$) [5], [11], [43].

A normalized distance is computed between IMFs for the successive frames to detect and overcome the differences in the estimated F0. Let k and k' denote IMF indexes, the distance is described as

$$\delta^q_{\hat{F}0}(k,k') = \left|\frac{\hat{F}0_{k,q} - \hat{F}0_{k',q}}{\hat{F}0_{k,q} + \hat{F}0_{k',q}}\right|. \tag{6}$$

The $\delta^q_{\hat{F}0}(k,k')$ values are computed for different indexes of k and k' resulting in a 4x4 distance matrix $\delta^q_{\hat{F}0}$. The row components of the matrix are summed to obtain the variation property for the k-th IMF. The frequency region is defined as the mean value of PEFAC F0 estimates ($\bar{F}0_q$) between the two IMFs with the smallest variation scores. Finally, the low/high pitch separation is performed according to the threshold γ as

$$\begin{cases} \bar{F}0_q \leq \gamma, & \text{low-frequency frame;} \\ \bar{F}0_q > \gamma, & \text{high-frequency frame.} \end{cases} \tag{7}$$

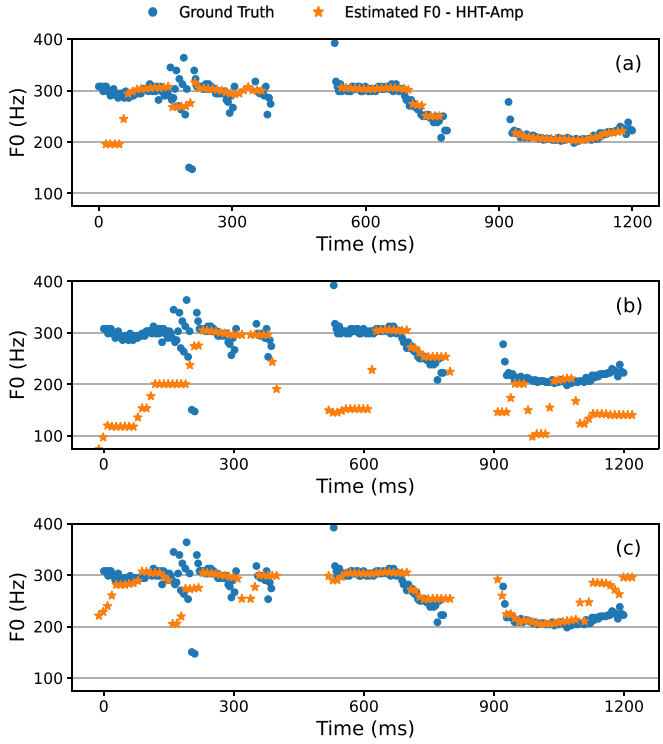

Fig. 3. Ground truth and F0 estimated with HHT-Amp technique for: (a) Clean speech segment, (b) Noisy signal with babble SNR $= -5$ dB, and (c) same noisy segment with estimates improved by FSFFE.

The threshold γ is fixed in 200 Hz which is related to the average values between male (50-200 Hz) and female (120-350 Hz) speakers [44].

The F0 adjustment is conducted according to the low/high pitch classification in (7). The F0 estimates are prone to doubling errors in low-pitch frames. Hence, a low-pitch frame that presents a F0 value ($f_{est,q}$) ranging from [200-400]Hz is adjusted to $f_{adj,q} = 0.5 f_{est,q}$. On the other hand, the high-pitch frame is adjusted to address possible halving and quartering [35] errors as follows:

$$f_{adj,q} = \begin{cases} 4f_{est,q}, & 50 \leq f_{est,q} \leq 100 \\ 2f_{est,q}, & 100 < f_{est,q} \leq 200 \end{cases}. \tag{8}$$

Fig. 3 illustrates the F0 adjustment procedure in frames of a 1200 ms speech signal. Fig. 3(a) refers to F0 attained with the HHT-Amp method for clean speech. The estimated values match the ground truth in the high-pitch region. Fig. 3(b) presents the F0 estimates related to the noisy version of the same speech segment for the babble noise [45] with SNR = -5 dB. Note that accuracy decreases significantly and halving errors appear in harmonic components, e.g., around 100 ms or 600 ms. These regions are adjusted with FSFFE as can be seen in Fig. 3(b). Observe that the proposed adjustment leads to accuracy improvement even in severe noisy conditions. The correction in harmonic detection is important especially in this case. Particularly, due to the fact that important components for speech intelligibility are placed in higher frequencies.

C. Third-Octave Bands Configuration

Third-octave filter banks have been shown to closely approximate the measured bands of the auditory filters [46]. Objective speech metrics consider the analysis of clean and noisy speech with third-octave subspaces. This is the case of the ESTOI [30] intelligibility measure, which gives a prediction through the correlation of third-order spectrograms from the reference and processed signal.

This work proposes the definition of auditory filtering based on the third-octave bands. The accurate harmonic detection $f_{adj,q}$ is adopted as the center frequency of the first band of the filter bank ($k = 0$). The center frequencies for the following k bands are attained adaptively in each frame q by

$$f_c(k, q) = 2^{\frac{k}{3}} f_{adj,q}. \tag{9}$$

The proposed set of filters provides better resolution in the frequencies near the harmonics of speech, which are the significant for speech intelligibility [30]. For instance, considering a fundamental frequency value of 200 Hz, in [26] it is assumed filters with center frequencies as integer multiples of F0, i.e., $f_c = [200, 400, 600, 800, 1000, \cdots]$ Hz. For the same F0 value, the proposed third-octave configuration presents $f_c = [200, 252, 317, 400, 503, \cdots]$ Hz.

D. Gammachirp Filtering

In this step, a set of L Gammachirp filters [29] $\{h_k(t), k = 1 \ldots, L\}$ are applied to successively filter the input sample sequence $x_q(t)$. Each filter $h_k(t)$ is implemented to the noisy signal considering frames of 32 ms, order $n = 4$, center frequencies given by (9). In order to align the impulse response functions, phase compensation is applied to all filters, which correspond to the non-causal filters

$$h_k(t) = a(t + t_c)^{n-1} \cos(2\pi f_c t + c \ln t) e^{-2\pi b(t+t_c)}\,,\; t \geq -t_c\,, \tag{10}$$

where c is the Gammachirp coefficient of the filter and

$$t_c = \frac{n-1}{2\pi b}, \tag{11}$$

which ensures that peaks of all filters occur at $t = 0$.

The bandwidth b is defined here according to the frequencies of the modulation transfer function considered in [31], [32], [33]. The results presented in [33] demonstrated that the frequency range relevant to the intelligibility of male speech sentences ranges from [1–12.5] Hz. Nevertheless, female sentences presented a larger range, with noticeable relevance for frequencies ≤ 20 Hz. Therefore, this work proposes a harmonic-adaptive bandwidth, given by

$$b = 0.15 f_{adj,q}. \tag{12}$$

Let $x_q^0(t) = x_q(t)$, the filtered signals $y_q^k(t), k = 1, \ldots, L$, are recursively computed by

$$\begin{cases} y_q^k(t) = x_q^{k-1}(t) * h_k(t) \\ x_q^k(t) = x_q^{k-1}(t) - y_q^k(t) \end{cases}, \quad k = 1, \ldots, L\,. \tag{13}$$

TABLE I
ESTOI [$\times 10^{-2}$] SCORES FOR DIFFERENT ASYMMETRY COEFFICIENTS OF THE GAMMACHIRP FILTER

Noise		Gammachirp Coefficient – c								
		2.0	1.5	1.0	0.5	0.0	–0.5	–1.0	–1.5	–2.0
Babble	–10 dB	18.5	19.3	18.5	18.9	19.0	18.4	**19.4**	18.4	19.1
	–5 dB	28.8	29.7	28.9	29.2	29.5	28.8	**29.8**	28.8	29.6
	0 dB	41.1	42.0	41.3	41.6	41.8	41.1	**42.2**	41.1	42.0
	5 dB	54.5	55.4	54.7	54.9	55.2	54.5	**55.6**	54.4	55.5
	Average	35.7	36.6	35.9	36.1	36.4	35.7	**36.8**	35.7	36.5
SSN	–10 dB	20.5	21.3	20.6	20.9	21.0	20.4	**21.5**	20.4	21.2
	–5 dB	30.2	31.1	30.4	30.7	30.8	30.3	**31.3**	30.1	31.1
	0 dB	41.6	42.4	41.8	42.0	42.2	41.6	**42.6**	41.5	42.5
	5 dB	54.2	55.0	54.4	54.6	54.8	54.2	**55.3**	54.1	55.2
	Average	36.6	37.4	36.8	37.0	37.2	36.6	**37.7**	36.5	37.5

The residual signal is defined as $r_q(t) = x_q^L(t)$ to guarantee the completeness of the input sequence, i.e.,

$$x_q(t) = \sum_{k=1}^{L} y_q^k(t) + r_q(t). \tag{14}$$

Table I presents the ESTOI scores for different asymmetry coefficients c of the Gammachirp filter. The intelligibility is predicted for a training subset of 48 speech signals of TIMIT [36] defined in [47]. The ESTOI scores with different values of c are computed for Babble [45] and SSN [48] noisy scenarios. Note that the coefficient c = -1 achieves the highest intelligibility rates for all the noisy conditions. This can be justified by the fact that acoustic noises might shift the harmonic detection. Therefore, the asymmetry of Gammachirp has the role of fine-tuning these harmonic components.

E. Frames Reconstruction With a Gain Factor

After the Gammachirp filtering, the amplitudes of the output samples $y_q^k(t), k = 1, \ldots, L$, are amplified by a gain factor $G_k \geq 1$. The idea is to emphasize the presence of the harmonic features of speech, which will lead to 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

$$\hat{x}_q(t) = \left[\sum_{k=1}^{L} G_k\, y_q^k(t)\right] + r_q(t)\,. \tag{15}$$

The reconstructed voiced frames in S_v 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 multiplies 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 each frame is reconstructed considering $G_k = 1$ for every $k \in \{1, \ldots, L\}$.

The set of gains G_k are empirically determined in each filter using the same training subset of 48 speech signals attained

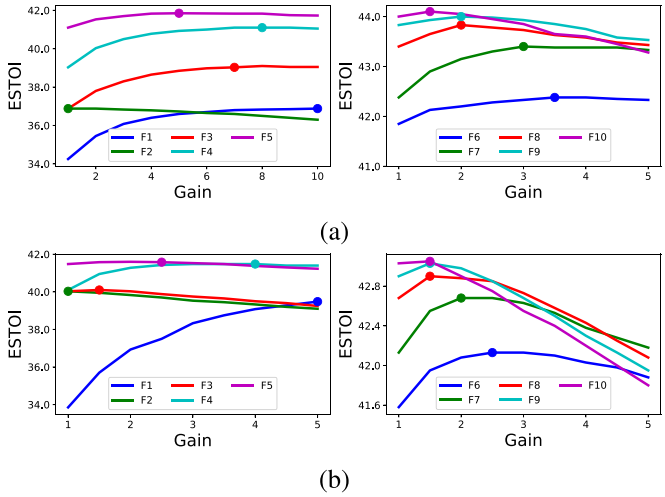

Fig. 4. ESTOI curves of (a) low pitch and (b) high pitch frames averaged for SNR values: −10 dB, −5 dB, 0 dB and 5 dB of Babble noise according the gain factor G_k for each gammachirp filter.

from TIMIT database. Fig. 4 illustrates the ESTOI curves for noisy speech signals with Babble and averaged to four SNR values. The configuration starts from the first filter (F1), and the gain is incremented until ESTOI reaches its maximum value (highlighted point). This gain is fixed, and the process is repeated for the subsequent filters. Observe that two different sets of gain are presented: one for low pitch (Fig. 4(a)) and the other for high pitch frames (Fig. 4(b)). Therefore, the G_k values for $L = 10$ filters that lead to the highest intelligibility ESTOI scores are defined as

$$G_k = \begin{cases} \{14, 1, 4, 8, 4, 3.5, 3, 2, 2, 1.5\}, & \text{low-pitch;} \\ \{14, 1, 1, 4.5, 2, 3.5, 2.5, 2, 1.5, 1.5\}, & \text{high-pitch.} \end{cases} \tag{16}$$

The proposed HDAG method is summarized in Algorithm 1. This algorithm is tailored to the harmonic detection scheme considered in this paper. However, Algorithm 1 can be also used with any other F0 estimation technique.

Algorithm 1: Intelligibility Enhancement Scheme HDAG.

for q **do**
  Input: $x_q(t)$
  Harmonic Detection
  $f_{est,q} \leftarrow$ F0 estimation with HHT-Amp as in Section II-A.
  $\hat{F}0_q \leftarrow$ PEFAC (5) for K=4 decomposed modes of (1).
  $\delta^q_{\hat{F}0} \leftarrow$ normalized distance matrix using (6)
  low/high pitch classification (7) according to $\bar{F}0_q$.
  Gammachirp Filtering
  for k **do**
    $h_k(t) \leftarrow$ impulse response of non-causal filters (10)
    $y^k_q(t) = x^{k-1}_q(t) * h_k(t)$
    $x^k_q(t) = x^{k-1}_q(t) - y^k_q(t)$
  end for
  $r_q(t) = x^L_q(t) \leftarrow$ residual components
  $\hat{x}_q(t) \leftarrow$ voiced frames reconstruction as in (15) and G_k from (16).
  $\hat{x}(t) \leftarrow$ overlap and add technique.
end for
return $\hat{x}(t)$

III. Harmonic-Based Comparative Methods

This Section briefly describes the baseline methods SSFV, PACO and GTF$_{F0}$. They also consider the harmonic components of noisy speech to attain intelligibility and quality improvement.

A. SSFV

The main idea of this solution consists of transforming the original signal adopting a Lombard effect strategy [25], [49]. In this effect the central frequencies of the formants are shifted (Formant Shifting). It moves away the energy from these frequencies from the region of spectral action of the noise. The formant shifting process is described in [23] and optimized to operate in environments with the presence of Car noise (composed by radio, message alert and telephone). Initially, LPC (Linear Prediction Coding) is used to estimate the poles and formant frequencies of the voiced speech signal. In the LPC model, a 25 ms frame of the signal $s(n, m)$ can be represented by linear predictions of order p [50], that is

$$s(n, m) = \sum_{j=1}^{p} a_j s(n - j, m) + e(n, m), \tag{17}$$

where a_j are the linear prediction coefficients, $e(n, m)$ indicates the residual error and $p = 12$. The variables n and m represent the signal sample and time frame indices, respectively. The LP filter A(z) is obtained from the coefficients a_j, so that

$$A(z) = 1 + \sum_{j=1}^{p} a_j z^j. \tag{18}$$

The poles **P** are obtained by the roots of the LP coefficients, and the formant frequencies **F** are defined as the estimated pole angles.

The formants obtained are shifted according to a function $\delta(F)$ [24] related to the characteristics of the acoustic noise. The displacement of formants is carried out according to the criterion

$$\hat{F}(f) = \begin{cases} F(f) + \delta(f), & f_1 < f < f_3 \\ F(f), & \text{otherwise.} \end{cases} \tag{19}$$

where f_1 and f_3 are the first and third formants, respectively. Finally, the resulting set of formants $\hat{\mathbf{F}}$ is obtained from these modifications.

B. PACO

The pitch-adaptive complex-valued Kalman filter (PACO) [39] is also adopted as a competitive technique for the proposed HDAG method. It applies the harmonic signal modeling for estimating the complex-valued speech AR parameters required for the Kalman filter. To this end, fundamental frequency estimation f for each 32 ms signal frame $y(n, l)$ is performed and phase

progression $\hat{\psi}(l)$ is recursively estimated for the harmonic h according to

$$\psi_h(l) = \psi_h(l-1) + \frac{\pi L}{f_s}(f_h(l) + f_h(l-1)). \quad (20)$$

Successive speech DFT bins of $y(n,l)$ are computed by incorporating the harmonic phase progression into a state-transition model. The AR coefficients $\hat{\mathbf{a}}(l)$ are defined from the DFT bins [39], which are the input for the Kalman filter gain G_K and obtain an estimation of $\hat{\mathbf{X}}(k,l)$ such as

$$\hat{X}(k,l) = G_k(k,l)(Y(k,l) - \hat{X}_{prop}(k,l)) \quad (21)$$

where $\hat{X}_{prop}$ is the state propagation estimate for the k-th bin. Finally, inverse DFT is applied and the processed speech signal is reconstructed by performing overlap and add.

C. GTF_{F0}

In the GTF_{F0} [26] 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] in frames of 32 ms considering order $n = 4$, and center frequency

$$f_c = kF0 \quad (22)$$

and bandwidth $b = 0.25F0$. The time-domain impulse response function described in (10) is applied for GTF_{F0} without the asymmetry coefficient. Thus, it can be considered a specific case of Gammachirp filterbank, in which $c = 0$.

After the Gammatone filtering, the amplitudes of the output samples $y_q^k(t), k = 1, \ldots, L$ are amplified by the following a gain factor $G_k \geq 1$. The integer multiples of F0 are amplified as in [26] with the following linear gains: $G_1 = G_2 = 5.0$, $G_3 = 4.0$ and $G_4 = 2.5$.

IV. Results and Discussion

This section presents objective results for the intelligibility and quality of acoustic signals processed by the HDAG method in comparison to SSFV, PACO and GTF_{F0} baseline techniques. ESTOI [30] and $ASII_{ST}$ [37] are considered to evaluate the speech intelligibility improvement and PESQ [38] compares the quality assessment of competitive methods. Following, results for a perceptual test are presented in order to corroborate the objective evaluation.

The experimental scenario considers a subset[2] of the TIMIT [36] database to evaluate the competitive methods. The set considered 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 F0 reference values and voiced/unvoiced information for the training and test datasets are obtained from [47]. Six noises are used to corrupt the speech utterances: acoustic Babble and Traffic attained from RSG-10 [45], Cafeteria, Train and Helicopter from Freesound.org,[3] and Speech Shaped Noise (SSN) from DEMAND [48] database. Experiments are conducted considering noisy signals with four SNR values (-10 dB, -5 dB, 0 dB and 5 dB), i.e., 288.000 test experiments in a frame by frame basis. In this study, it is assumed that the FSFFE separation into high-pitch and low-pitch speech frames is considered perfect and generates no errors in the whole system.

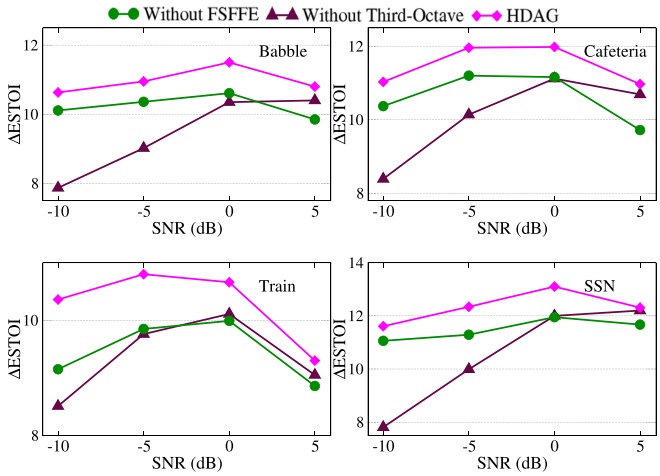

Fig. 5. ESTOI improvement (ΔESTOI) for the proposed HDAG solution and its versions without FSFFE or third-octave bands.

Fig 5 shows the ESTOI improvement (ΔESTOI) achieved by the proposed HDAG in comparison with its version without the FSFFE or third-octave bands configuration. Note that the HDAG attains the highest ESTOI improvement in the four noises, particularly in lower SNR values ($\leq$ 0 dB). In these cases, the third-octave band configuration has an interesting significance, since it provides a better resolution to the filterbank. Thus, the third-octave bands allow a fine adjustment of the gain factors in the filters whose frequency ranges are relevant to intelligibility. For instance, observe the ESTOI improvement in Train noise with SNR = -10 dB. While the version without a third-octave shows an improvement of 7.81 p.p. in ESTOI, the proposed HDAG presents an ESTOI improvement of 11.61 p.p.. Moreover, in the same case the FSFFE separation and the adjustment of estimated harmonic components of speech signals increased the ESTOI gain from 9.15 p.p. to 10.36 p.p..

A. Intelligibility and Quality Objective Evaluation

Table II shows the intelligibility and quality objective results with ESTOI and PESQ measures, respectively. Note that Babble and SSN noises present the most challenging scenarios among those evaluated in terms of intelligibility. For instance, the ESTOI averaged for the SNR values of UNP speech signals are 0.36 and 0.35 for the respective noises. Moreover, observe that the HDAG method achieves the best results for all 24 noise conditions even in the most challenging scenarios with negative SNR values. The scores of HDAG are particularly interesting for the non-stationary noises, i.e., Babble and Cafeteria. For these noise sources the ESTOI attained are considerably higher than all the competing solutions for all SNR values. The highest ESTOI accomplished by HDAG is 13 p.p can be observed for Helicopter

[1]Code available at staffwww.dcs.shef.ac.uk/people/n.ma/resources/gammatone/

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

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

TABLE II
INTELLIGIBILITY AND QUALITY RESULTS WITH THE PROPOSED HDAG AND COMPETITIVE METHODS

Noise	SNR	ESTOI					PESQ				
		UNP	SSFV	PACO	GTF_{F0}	HDAG	UNP	SSFV	PACO	GTF_{F0}	HDAG
Babble	–10 dB	0.18	0.17	0.17	0.24	**0.28**	0.56	0.99	1.25	1.77	**2.06**
	–5 dB	0.29	0.29	0.30	0.37	**0.40**	1.52	1.54	1.94	2.30	**2.50**
	0 dB	0.41	0.42	0.43	0.50	**0.53**	1.90	1.92	2.45	2.72	**2.86**
	5 dB	0.55	0.55	0.58	0.64	**0.66**	2.35	2.36	2.94	3.12	**3.22**
	Average	0.36	0.36	0.37	0.44	**0.47**	1.58	1.70	2.14	2.48	**2.66**
Cafeteria	–10 dB	0.20	0.19	0.19	0.27	**0.31**	1.00	1.30	1.50	1.97	**2.20**
	–5 dB	0.31	0.31	0.31	0.40	**0.43**	1.63	1.67	2.01	2.48	**2.63**
	0 dB	0.44	0.44	0.45	0.54	**0.56**	2.07	2.09	2.52	2.90	**3.02**
	5 dB	0.58	0.58	0.61	0.67	**0.69**	2.50	2.51	2.97	3.29	**3.37**
	Average	0.38	0.38	0.39	0.47	**0.50**	1.80	1.89	2.25	2.66	**2.81**
Traffic	–10 dB	0.38	0.38	0.44	0.44	**0.47**	1.59	1.58	**2.82**	2.34	2.51
	–5 dB	0.51	0.50	0.56	0.58	**0.60**	2.04	2.04	**3.27**	2.73	2.88
	0 dB	0.63	0.63	0.68	0.70	**0.71**	2.55	2.55	**3.62**	3.12	3.26
	5 dB	0.74	0.74	0.79	0.79	**0.80**	3.06	3.06	**3.86**	3.52	3.62
	Average	0.57	0.56	0.62	0.63	**0.65**	2.31	2.31	**3.39**	2.93	3.07
Train	–10 dB	0.32	0.30	0.36	0.38	**0.42**	1.33	1.38	1.92	2.09	**2.29**
	–5 dB	0.43	0.43	0.47	0.51	**0.54**	1.82	1.83	2.55	2.61	**2.75**
	0 dB	0.55	0.54	0.58	0.63	**0.65**	2.33	2.34	3.03	3.06	**3.18**
	5 dB	0.65	0.65	0.69	0.73	**0.75**	2.78	2.79	3.37	3.42	**3.54**
	Average	0.49	0.48	0.53	0.56	**0.59**	2.06	2.08	2.72	2.79	**2.94**
Helicopter	–10 dB	0.30	0.30	0.33	0.39	**0.43**	1.55	1.59	2.21	2.34	**2.54**
	–5 dB	0.41	0.41	0.45	0.52	**0.54**	1.89	1.91	2.71	2.74	**2.87**
	0 dB	0.53	0.53	0.59	0.64	**0.66**	2.33	2.34	3.17	3.15	**3.26**
	5 dB	0.66	0.65	0.72	0.75	**0.76**	2.76	2.76	3.53	3.51	**3.60**
	Average	0.47	0.47	0.52	0.58	**0.60**	2.13	2.15	2.91	2.93	**3.06**
SSN	–10 dB	0.17	0.16	0.20	0.24	**0.29**	1.22	1.41	1.95	1.88	**2.17**
	–5 dB	0.28	0.28	0.32	0.37	**0.41**	1.45	1.47	2.41	2.25	**2.41**
	0 dB	0.41	0.41	0.45	0.51	**0.54**	1.84	1.85	**2.89**	2.68	2.80
	5 dB	0.54	0.54	0.59	0.64	**0.66**	2.32	2.33	**3.29**	3.11	3.20
	Average	0.35	0.35	0.39	0.44	**0.47**	1.70	1.77	2.63	2.48	**2.65**
Overall		0.44	0.43	0.47	0.52	**0.54**	1.93	1.98	2.67	2.71	**2.86**

noise with SNR = -10 dB. According to the overall average, the proposed solution outperforms the competitive approaches with ESTOI of 0.54, against 0.52, 0.47 and 0.43 for GTF_{F0}, PACO and SSFV, respectively.

The PESQ score is here computed from 30% of the most relevant harmonic frames of noisy speech. These frames are selected from those with the lowest signal-to-noise ratio values. Note that HDAG outperforms the competing approaches for most of the noisy speech conditions in terms of quality assessment. The proposed solution achieves the highest PESQ, except for Traffic and SSN (0 dB and 5 dB) noises. In these cases PACO approach presents superior results since it is a speech enhancement method whose main focus is the gain of quality. In Helicopter with SNR = -10 dB the PESQ score attained by HDAG is 1.02 higher than UNP followed by increments of 0.79, 0.66 and 0.04 presented by GTF_{F0}, PACO and SSFV, respectively. In summary, the overall PESQ obtained with HDAG is 2.86, against 2.71 for the competing approach GTF_{F0}. Therefore, these results indicate that the proposed solution also provides quality assessment, outperforming even speech enhancement methods in the overall average.

Table III presents the average $ASII_{ST}$ results for the unprocessed (UNP) noisy speech signals. Here the SSN and Babble noises attained the lowest scores for SNR value of -10 dB, with $ASII_{ST}$ of 19.3 and 23.1, respectively. The $ASII_{ST}$ values incremented by each competitive method (Δ $ASII_{ST}$) are depicted in Fig. 6 for the six acoustic noises. Observe that the proposed solution accomplishes the highest scores for most conditions, except for Traffic (SNR = -10 dB). The best $\Delta ASII_{ST}$ (10.1×10^{-2}) is achieved by the challenging SSN noise in -10 dB. As can be seen in ESTOI, the SSFV approach does not present a noticeable $ASII_{ST}$ increment. Moreover, for the non-stationary Cafeteria noise the proposed solution attains an average intelligibility enhancement of 5.4×10^{-2}, compared with 3.5×10^{-2}, 1.8×10^{-2} and 0.3×10^{-2} for baselines GTF_{F0}, PACO and SSFV. Therefore, these results reinforce the robustness of HDAG against several noisy masking effects.

TABLE III
$ASII_{ST}$ [$\times10^{-2}$] SCORES FOR UNP NOISY SPEECH

SNR	Babble	Cafeteria	Traffic	Train	Helicopter	SSN
–10 dB	23.1	24.3	35.8	34.6	34.1	19.3
–5 dB	26.6	27.9	39.2	39.9	40.2	23.7
0 dB	33.0	34.3	43.7	47.2	43.5	30.0
5 dB	40.9	42.3	47.5	54.9	51.0	37.9
Average	30.9	32.2	41.6	44.2	42.2	27.7

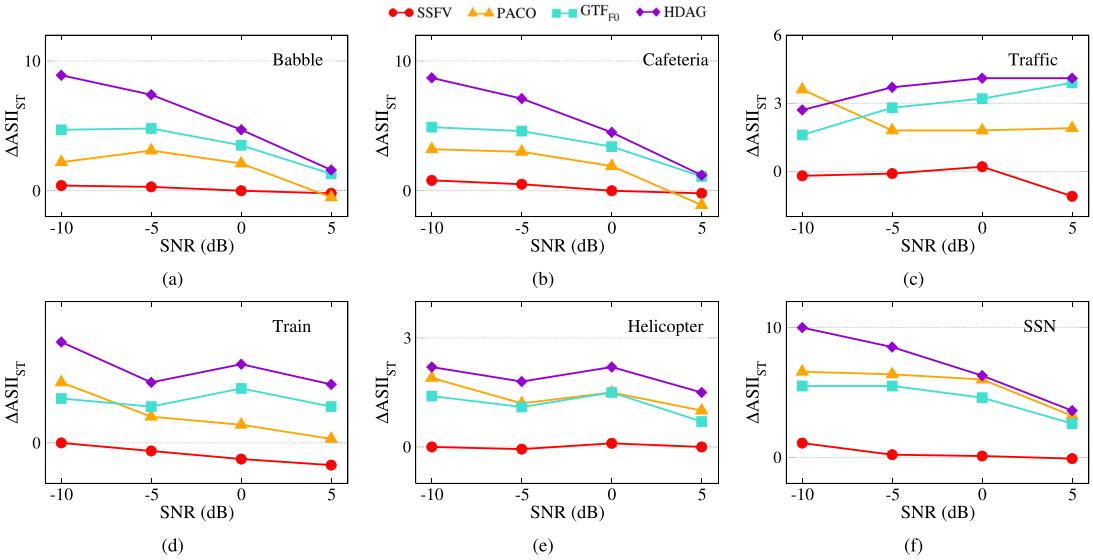

Fig. 6. ΔASII_{ST} intelligibility enhancement $[\times 10^{-2}]$ averaged for speech signals corrupted by noises: (a) Babble, (b) Cafeteria, (c) Traffic, (d) Train, (e) Helicopter, and (f) SSN.

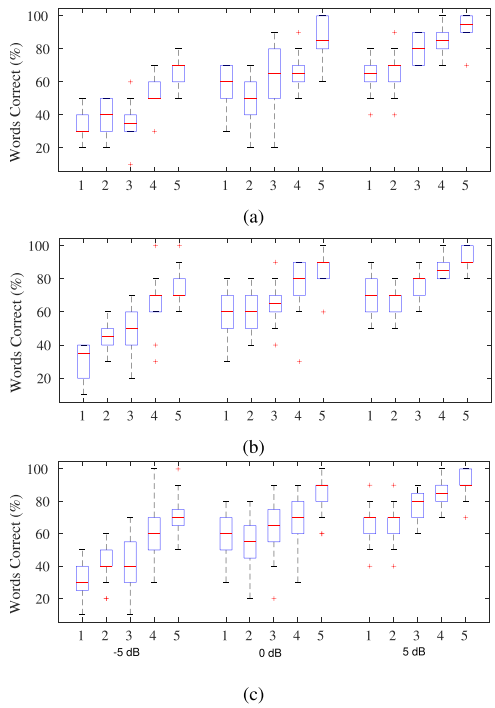

Fig. 7. Perceptual intelligibility evaluation with SSN additive acoustic noise for (a) male, (b) female volunteers, and (c) overall scores. Each case denotes: 1-UNP, 2-SSFV, 3-PACO, 4-GTF$_{F0}$, and 5-HDAG.

B. Perceptual Intelligibility Evaluation

A subjective listening test [51] is conducted considering a scenario of phonetically balanced words.[4] Ten native male and ten female Brazilian volunteers perform the test, with ages ranging from 19 to 57 years with an average of 32. The SSN noise is adopted with SNRs of -5 dB, 0 dB and 5 dB. Ten words are applied for each of the 15 test conditions, i.e., three SNR levels and four methods plus the unprocessed case. Participants are introduced to the task in a training session with 4 words. The material is diotically presented using a pair of Roland RH-200S headphones. Listeners hear each word once in an arbitrary presentation order and are asked to indicate the word in a sheet list.

The intelligibility results for each method are presented in Fig. 7. Each boxplot depicts the median and deviation values scores (%) for one scenario, separating the (a) male, (b) female volunteers, and (c) the overall scores. The proposed method accomplishes intelligibility under all conditions over the competing approaches. For male listeners the HDAG obtained average intelligibility scores of 66%, 85% and 93% compared to 52%, 66% and 86% in the GTF$_{F0}$ technique for SNR values of -5 dB, 0 dB and 5 dB, respectively. Furthermore, female volunteers presented higher intelligibility rates than males, mainly for -5 dB with 75% and 65% for HDAG and GTF$_{F0}$. The overall results show again the superiority of HDAG with average scores of 71%, 86% and 92%, surpassing GTF$_{F0}$ (59%, 71% and 86%) and PACO (43%, 64% and 78%). In accordance with findings in the objective measures ESTOI and ASII$_{ST}$, SSFV attains scores less or equal to the UNP case.

TABLE IV
NORMALIZED MEAN PROCESSING TIME

SSFV	PACO	GTF$_{F0}$	HDAG
0.32	0.67	0.89	1.00

C. Normalized Processing Time

Table IV indicates the computational complexity which refers to the normalized processing time required for each method

[4]The complete test database is available at lasp.ime.eb.br.

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 proposed HDAG solution. The processing time required for F0 estimation and accurate harmonic adjustment is also considered here. Note that the HDAG and GTF_{F0} schemes present a longer processing time since the FSFFE low/high pitch classification and HHT-Amp estimation are based on the EEMD, and demand a relevant computational cost.

V. Conclusion

This paper introduced the HDAG method for speech intelligibility enhancement in harmonic components of noisy speech. It is composed of four main steps. First, the HHT-Amp technique is adopted to estimate the F0 from voiced frames. The FSFFE separation was used for the detection and adjustment of these estimates, improving their accuracy. Then, a selective Gammachirp filterbank was applied to the frames considering third-octave bands to best cover the regions most relevant to intelligibility. Finally, the filtered components were amplified by gain factors regulated by low/high pitch classification. Extensive experiments were conducted to evaluate the intelligibility enhancement provided by the HDAG method and competitive approaches. Six acoustic noises were considered with four SNR values. Three measures are adopted for the objective evaluation of speech intelligibility and quality. The results demonstrate that the HDAG method outperformed the competitive approaches, with higher intelligibility and quality assessment in most noisy environments. A perceptual test for male and female listeners corroborated the objective results. Future research includes the investigation of the proposed method for other conditions, such as intelligibility enhancement for noisy reverberant speech.

References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.

[2] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*. Berlin, Germany: Springer, 2004, pp. 231–308.

[3] T. Gerkmann and R. C. 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.

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

[5] 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, 2021.

[6] E. Dranka and R. Coelho, "Robust maximum likelihood acoustic energy based source localization in correlated noisy sensing environments," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 2, pp. 259–267, Mar. 2015.

[7] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.

[8] J. Martinez-Carranza and C. Rascon, "A review on auditory perception for unmanned aerial vehicles," *Sensors*, vol. 20, no. 24, pp. 1–24, 2020.

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

[10] A. Venturini, L. Zao, and R. Coelho, "On speech features fusion, integration Gaussian modeling and multi-style training for noise robust speaker classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1951–1964, Dec. 2014.

[11] 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.

[12] P. C. 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.

[13] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Commun.*, vol. 51, no. 3, pp. 230–239, 2009.

[14] G. Kim and P. C. Loizou, "Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints," *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1010–1013, Dec. 2010.

[15] F. Farias and R. Coelho, "Blind adaptive mask to improve intelligibility of non-stationary noisy speech," *IEEE Signal Process. Lett.*, vol. 28, pp. 1170–1174, 2021.

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

[17] 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.

[18] T. Wang, W. Zhu, Y. Gao, S. Zhang, and J. Feng, "Harmonic attention for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2424–2436, 2023.

[19] S. Y. Barysenka and V. I. Vorobiov, "SNR-based inter-component phase estimation using bi-phase prior statistics for single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2365–2381, 2023.

[20] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Speech Audio Process.*, vol. 6, no. 1, pp. 36–48, Jan. 1998.

[21] 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.

[22] 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.

[23] 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.

[24] 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, 2017.

[25] E. Lombard, "Le signe de l'elevation de la voix," *Mal. Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 25, pp. 101–119, 1911.

[26] 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, 2021.

[27] 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.

[28] R. A. Lutfi and R. D. Patterson, "On the growth of masking asymmetry with stimulus intensity," *J. Acoust. Soc. Amer.*, vol. 76, pp. 739–745, 1984.

[29] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," *J. Acoust. Soc. Amer.*, vol. 101, no. 1, pp. 412–419, 1997.

[30] J. Jensen and C. H. 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.

[31] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1053–1064, 1994.

[32] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulation on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 2670–2680, 1994.

[33] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech iintelligibility," *PLoS Comput. Biol.*, vol. 5, no. 3, pp. 1–14, 2009.

[34] A. Queiroz and R. Coelho, "Noisy speech based temporal decomposition to improve fundamental frequency estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2504–2513, 2022.

[35] 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, 2020.

[36] J. Garofolo et al., "Timit acoustic-phonetic continuous speech corpus," in *Proc. Linguist. Data Consortium*, Philadelphia, PA, USA, 1993.
[37] R. C. Hendriks, J. B. 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.
[38] 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," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, pp. 749–752.
[39] J. Stahl and P. Mowlaee, "Exploiting temporal correlation in pitch-adaptive speech enhancement," *Speech Commun.*, vol. 111, pp. 1–13, 2019.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] I. R. Titze, *Principles of Voice Production*. Englewood Cliffs, NJ, USA: Prentice Hall, 1994.
[45] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise-database," *Rep. IZF*, vol. 3, pp. 1–12, 1988.
[46] *Electroacoustics Octave-Band and Fractional-Octave-Band Filters–Part 1: Specifications*," Standard IEC 61260-1: 2014, International Electrotechnical Commission, Geneva, Switzerland, 2014.
[47] S. Gonzalez and M. Brookes, "Pitch of the core TIMIT database set," 2014.
[48] 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.
[49] J. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Amer.*, vol. 93, no. 1, pp. 510–524, 1993.
[50] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ, USA: Prentice Hall, 1978.
[51] S. Ghimire, "Speech intelligibility measurement on the basis of ITU-T recommendation P.863," 2012.

Anderson Queiroz (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from the Regional Integrated University of High Uruguay and Missions - URI, Uruguay, in 2018, and the M.Sc. degree from the Military Institute of Engineering (IME), Rio de Janeiro, Brazil, in 2021. He is currently working toward the Ph.D. degree in defense engineering at IME, under the supervision of Prof. Rosângela Coelho. His research interests include acoustic signal processing, speech enhancement in acoustic scene environments, and objective intelligibility prediction.

Rosângela Coelho (Senior Member, IEEE) received the M.Sc. degree in electrical engineering from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil, in 1991, and the Ph.D. degree from the Ecole Nationale Supérieure des Télécommunications (Télécom Paris/IP Paris), France, in 1995. She was a Postdoctoral Researcher and an Assistant Professor at PUC-Rio and also a Visiting Professor at Télécom Paris (IP/Paris), France, from 1996 to 2001. In 2002, she joined the Military Institute of Engineering, Brazil, where she is an Associate Professor with Electrical Engineering Department. Prof. Coelho founded and heads the Laboratory of Acoustic Signal Processing. Her main research interests include acoustic signal processing, speech intelligibility and quality enhancement, time-frequency methods, drone audition, and acoustic signal processing with perceptual analysis and personalized solution for individuals with ASD condition. In 2003, she was the recipient of the University Research Program Grant Award from CISCO/USA. Prof. Coelho is currently a Senior Member of Signal Processing Society, and an Affiliate Member of the Technical Committee Audio and Acoustic Signal Processing. Since 2019, she has been an Associate Editor and currently as a Senior Associate Editor for IEEE SIGNAL PROCESSING LETTERS.