IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 November 2022, accepted 19 December 2022, date of publication 26 December 2022,
date of current version 5 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232549

RESEARCH ARTICLE

An Improved Noise Reduction Technique for Enhancing the Intelligibility of Sinewave Vocoded Speech: Implication in Cochlear Implants

VENKATESWARLU POLUBOINA[1], **APARNA PULIKALA**[1], **(Senior Member, IEEE), AND ARIVUDAI NAMBI PITCHAI MUTHU**[2,3]
[1]Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, Mangaluru 575025, India
[2]Department of Audiology and Speech Language Pathology, Kasturba Medical College, Mangaluru 575001, India
[3]Manipal Academy of Higher Education, Manipal 576104, India

Corresponding authors: Arivudai Nambi Pitchai Muthu (arivudai.nambi@manipal.edu), Aparna Pulikala (p.aparnadinesh@nitk.edu.in), and Venkateswarlu Poluboina (venki.187ec009@nitk.edu.in)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Research Panel Committee under Approval No. NITK/EC/Ph.D/284/2021.

ABSTRACT A cochlear implant (CI) is the most suitable option for individuals with severe profound hearing loss. CI restores the audibility to near perfection and offers good speech understanding in quiet. However, the speech perception in noise with CIs is less optimal as most speech coding strategies of CIs encode only the temporal envelope. Besides the current CI signal coding strategies lacks sophisticated pre-processing. In the current study, we proposed a novel pre-processing method to improve speech Intelligibility in noise and tested using the acoustic simulations of cochlear implants. The proposed noise reduction technique aims to minimize the mean square error (MSE) between the temporal envelopes of the enhanced speech and its clean speech. Therefore, the proposed method will be suitable for CI applications. This paper provides an analysis of the theoretical derivation of the noise suppression function and also the performance evaluation using objective and subjective tests. The effectiveness of the proposed method was objectively evaluated using the SRMR-CI and ESTOI. Additionally, speech recognition through the acoustic simulations of the cochlear implant was done for the subjective evaluation. Performance of the proposed method was compared with the Weiner filter (WF) and sigmoidal functions. The sinewave vocoder was used to simulate the cochlear implant perception. Both objective and subjective scores revealed that the performance of the proposed technique is superior to the WF and sigmoidal function.

INDEX TERMS Noise reduction, speech recognition, cochlear implant, vocoder simulation.

I. INTRODUCTION

For patients with profound hearing loss, cochlear implantation is a life-changing procedure [1]. Several sound-processing techniques have been developed over the past few decades to provide an improved auditory experience to the cochlear implant (CI) users [2]. The speech recognition performance of the CI users in quiet situation is satisfactory. However, their performance in noisy environments is suboptimal [3], [4], [5]. CIs recipients are less likely to identify speech in the existence of background noise than people with normal hearing (NH). CI user's speech recognition scores are reduced from 30 to 60% when signal-to-noise ratios are low in real life [6]. However, individuals with cochlear implants require a 25dB higher SNR to recognize at minimum 50% of the target speech given in the background talker noise [7]. These findings indicate that noise reduction strategies in CIs are a critical link in the signal processing pipeline because they help users maintain good speech intelligibility even in noisy conditions.

The associate editor coordinating the review of this manuscript and approving it for publication was Berdakh Abibullaev.

A variety of noise reduction (NR) strategies have been proposed to enhance speech intelligibility (SI), voice quality, and hearing comfort in poor listening situations to overcome issues with speech perception. The goal of NR is to remove as much noise as feasible from a noisy mixture while keeping the target signal distortions to a minimum. Time-frequency masking (TFM) is a type of NR technique frequently used in hearing aids and CIs [8], [9]. The ideal binary mask (IBM) [10], [11] and the Wiener Filter (WF) [9] are the most common methods [12]. When it comes to CI applications, general-purpose masks have their limitations [13].

CI users are generally encouraged to use suppression functions that are more aggressive than those used in hearing aids or those with normal hearing [8], [14], [15]. When the SNR is above a specified threshold value, IBM preserves the time-frequency points of the noisy signal and suppresses the remaining time-frequency points. Unlike IBM, WF is the method for providing masks with continuous weights. References [12], [16], and [17] states that in auditory prosthetics such as hearing aids and CIs, WF has an improvement over the IBM approach. The WF method provided a path to reduce the mean square error between the target and estimated signals.

In recent studies, machine learning techniques have been used in CIs to reduce noise [18], [19], [20]. WF-based techniques are commonly used for getting the desired target speech for the supervised learning process. Even though they provide attractive results, their accuracy is strongly correlated with the size of data sets for speech and noise, as this can result in over-fitting of the data (training), which may limit their ability to generalize to various acoustic environments.

In this study, we present a technique for reducing noise in acoustic simulations of CIs. Most cochlear implants encode temporal envelopes. Hence, Therefore, the current method intends to minimize the MSE between the estimated and target signal's squared envelopes. Simulations reveal that the proposed method (PM) performs better than the WF mask with the minimum MSE. In the current study, the efficacy of the PM was tested on the acoustic simulations of CI using sinewave vocoder [21], [22]. In the field of CI research, sinewave vocoders frequently serve to replicate some of the characteristics of CI processing [23]. Here the proposed noise reduction method and the traditional single microphone noise reduction method (i.e., WF) are compared in terms of speech recognition performance. As part of this work, we aim to examine the noise suppression capacity of NR methods under various challenging conditions. Using two noises with lower SNR levels, we synthesize the test data for evaluation. For confirming the effectiveness of PM on normal speech, we use an objective measurement (the extended short-time objective intelligibility (ESTOI) [24]). The ESTOI measure has shown to be highly accurate in predicting speech intelligibility under many conditions of degradation [24]. We conducted a listening experiment with NH subjects using vocoder speech to evaluate the performance further. Psychoacoustic experiments with NH people indicated that the proposed method yields higher speech intelligibility in a wide variety of SNRs.

The following sections of the paper are organized as follows: In Section II noise reduction methodology of the proposed method is explained. In Section III, we discuss the simulation results of the perceptual and objective evaluations. Section IV and V of the paper provide a discussion and conclusion.

II. NOISE REDUCTION METHODOLOGY

The typical noise reduction method applied to the CI simulator is as shown in Fig. 1. The general noise additive method defines the noisy speech signal as $y(n) = x(n) + b(n)$, where $b(n)$ is the additive noise, and $x(n)$ is the target speech signal. The spectral components are calculated by Fast Fourier Transform (FFT) with two times window length, whereas the window length is 20 msec of the sampling rate (f_s). The ω^{th} spectral component of noisy speech short time frame (τ_s) can be defined by

$$Y(\tau_s, \omega) = X(\tau_s, \omega) + B(\tau_s, \omega) \tag{1}$$

$$\bar{X}(\tau_s, \omega) = Y(\tau_s, \omega) * W(\tau_s, \omega) \tag{2}$$

where $W(\tau_s, \omega)$ is the noise reduction filter's coefficient vector. There are many methods for finding the coefficient vector. They define the filter coefficients based on the functions of noisy speech SNR estimations. The Wiener filter (WF) is an example of a time-frequency mask that has been effectively applied to CI. The WF [25] can be defined as

$$W(\tau_s, \omega) = \frac{\gamma(\tau_s, \omega)}{1 + \gamma(\tau_s, \omega)} \tag{3}$$

where $\gamma(\tau_s, \omega)$ is an a priori SNR defined as follows:

$$\gamma(\tau_s, \omega) = \frac{E[X^2(\tau_s, \omega)]}{E[B^2(\tau_s, \omega)]} \tag{4}$$

in which $E[X^2(\tau_s, \omega)]$, and $E[B^2(\tau_s, \omega)]$ represents clean speech and noise instantaneous powers respectively. E[ ] represent the expected value operator. From the decision direct approach method [26], as defined by a priori SNR,

$$\begin{aligned}\gamma(\tau_s, \omega) = \alpha &\frac{E[X^2(\tau_s - 1, \omega)]}{E[B^2(\tau_s, \omega)]} \\ &+ (1 - \alpha) Max\left[\frac{E[Y^2(\tau_s, \omega)]}{E[B^2(\tau_s, \omega)]} - 1, 0\right]\end{aligned} \tag{5}$$

where α is the weighting factor, for better performance of WF, we consider this value 0.98. Similarly, the sigmoid function [27] employed for noise reduction was also successfully used for CIs, defined as follows

$$g(\tau_s, \omega) = e^{-2/\gamma(\tau_s, \omega)} \tag{6}$$

A. SIGNAL PROCESSING

The noisy speech signal is windowed, after which the short-time Fourier transform (STFT) is applied. The estimated envelope is the absolute value of its STFT signal,

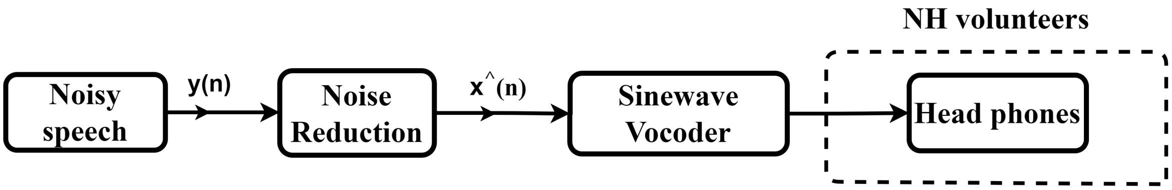

FIGURE 1. Block diagram representing noise reduction and vocoder-based simulation of cochlear implants.

defined as

$$Y_a(\tau_s, \omega) = Y_r(\tau_s, \omega) + iY_i(\tau_s, \omega) \tag{7}$$

$$Envelope = \sqrt{Y_r^2(\tau_s, \omega) + Y_i^2(\tau_s, \omega)} \tag{8}$$

phase information defined as

$$\phi(\tau_s, \omega) = \tan^{-1} \frac{Y_i(\tau_s, \omega)}{Y_r(\tau_s, \omega)} \tag{9}$$

B. THE PROPOSED NOISE SUPPRESSION FUNCTION

In this section, we introduce a novel optimization framework for obtaining noise suppression function and calculation of noise power. The proposed noise reduction method to calculate suppression function from the minimization of MSE between the desired speech and its enhanced speech, at each spectral band, is defined as

$$J(\tau_s, \omega) = \mathbb{E}[e^2(\tau_s, \omega)] \tag{10}$$

where $e(\tau_s, \omega)$ is the error between desired speech and its enhanced speech envelope, given by

$$e(\tau_s, \omega) = \mid X_a(\tau_s, \omega) \mid^2 - \mid \hat{X}_a(\tau_s, \omega) \mid^2 \tag{11}$$

$$\hat{X}_a(\tau_s, \omega) = V(\tau_s, \omega) * Y(\tau_s, \omega) \tag{12}$$

where $V(\tau_s, \omega)$ are filter coefficients.

$$e^2(\tau_s, \omega) = \left(\mid X(\tau_s, \omega) \mid^2 - \mid V(\tau_s, \omega) * Y(\tau_s, \omega) \mid^2 \right)^2 \tag{13}$$

Assuming that $X(\tau_s, \omega)$ and $B(\tau_s, \omega)$ both have zero means and are independent of each other [28], (10) can be written as

$$\begin{aligned} J(\tau_s, \omega) = {} & E\left[\mid X_a(\tau_s, \omega) \mid^4\right] \\ & + \mid V(\tau_s, \omega) \mid^4 E\left[\mid X_a(\tau_s, \omega) \mid^4\right] \\ & + 4 \mid V(\tau_s, \omega) \mid^4 E\left[\mid X_a(\tau_s, \omega) \mid^2\right] \\ & \times E\left[\mid B_a(\tau_s, \omega) \mid^2\right] \\ & + \mid V(\tau_s, \omega) \mid^4 E\left[\mid B_a(\tau_s, \omega) \mid^4\right] \\ & - 2 \mid V(\tau_s, \omega) \mid^2 E\left[\mid X_a(\tau_s, \omega) \mid^4\right] \\ & - 2 \mid V(\tau_s, \omega) \mid^2 E\left[\mid X_a(\tau_s, \omega) \mid^2\right] \\ & \times E[\mid B_a(\tau_s, \omega) \mid^2] \end{aligned} \tag{14}$$

For minimizing, the above equation (14) can be differentiated with respect to $V(\tau_s, \omega)$ and equated to zero [29] which gives

$$\begin{aligned} & \mid V(\tau_s, \omega) \mid^2 (E\left[\mid X_a(\tau_s, \omega) \mid^4\right] \\ & \quad + E\left[\mid X_a(\tau_s, \omega) \mid^2\right] E[\mid B_a(\tau_s, \omega) \mid^2] \\ & \quad + E\left[\mid B_a(\tau_s, \omega) \mid^4\right]) \\ & = E[\mid X_a(\tau_s, \omega) \mid^4] + E[\mid X_a(\tau_s, \omega) \mid^2] \\ & \quad \times E[\mid B_a(\tau_s, \omega) \mid^2] \end{aligned} \tag{15}$$

The above equation can be rewritten as (16), shown at the bottom of the page, where $\mid X_a^4(\tau_s, \omega) \mid$ can be written as

$$\begin{aligned} E[\mid X_a^4(\tau_s, \omega) \mid] = {} & E[X^4(\tau_s, \omega)] + E[X^4(\tau_s, \omega)] \\ & + 2E[X^2(\tau_s, \omega)]E[X^2(\tau_s, \omega)] \end{aligned} \tag{17}$$

where $\sigma_{ax}^2(\tau_s, \omega)$, $\sigma_{ab}^2(\tau_s, \omega)$ represent acoustic clean and noise signal powers respectively, which can be defined as

$$\sigma_{ax}^2(\tau_s, \omega) = 2\sigma_x^2(\tau_s, \omega) = E[X^2(\tau_s, \omega)] \tag{18}$$

$$\sigma_{ab}^2(\tau_s, \omega) = 2\sigma_b^2(\tau_s, \omega) = E[B^2(\tau_s, \omega)] \tag{19}$$

C. ESTIMATION OF NOISE POWER

The voice activity detection method is used to decide whether the input signal contains noise or speech based on the speech presence probability (SPP), with usual probability threshold (PTH) between 0 to 1. In this study, we considered the SPP greater than or equal to 0.6 for speech presence based on the pilot study [30], [31]. SPP of less than 0.6 is considered as the noise. The noise power spectral density is calculated using the typical recursive relation [23]

$$\begin{aligned} E[B_a^2(\tau_s, \omega)] = {} & \lambda E[B_a^2(\tau_s - 1, \omega)] \\ & + (1 - \lambda)E[\mid Y_a(\tau_s, \omega) \mid^2] \end{aligned} \tag{20}$$

$$\mid V(\tau_s, \omega) \mid = \sqrt{\frac{E\left[\mid X_a^4(\tau_s, \omega) \mid\right] + \sigma_{ax}^2(\tau_s, \omega) * \sigma_{ab}^2(\tau_s, \omega)}{E\left[\mid X_a^4(\tau_s, \omega) \mid\right] + E\left[\mid B_a^4(\tau_s, \omega) \mid\right] + 4\sigma_{ax}^2(\tau_s, \omega) * \sigma_{ab}^2(\tau_s, \omega)}} \tag{16}$$

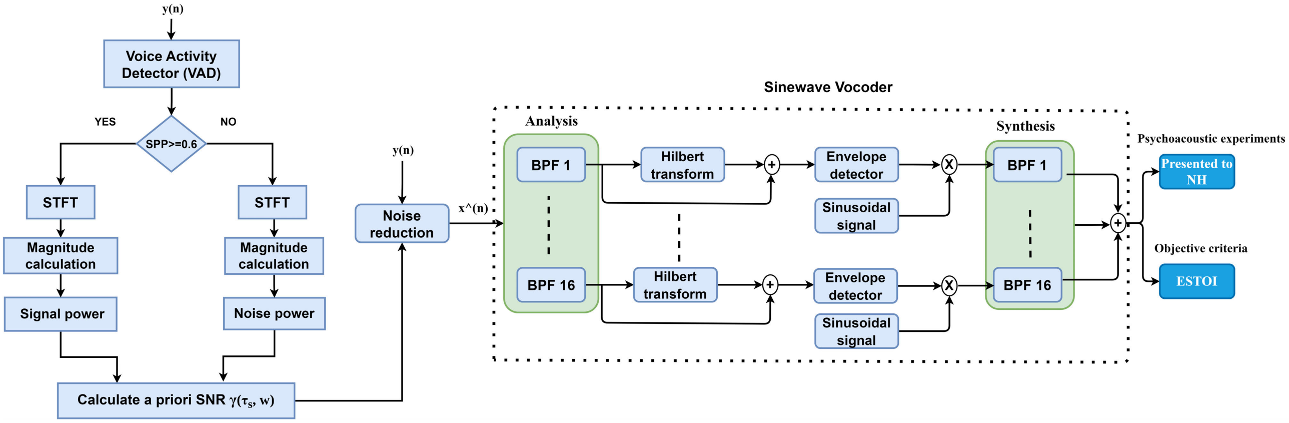

FIGURE 2. Block diagram of steps involved in psychoacoustic studies and objective assessment.

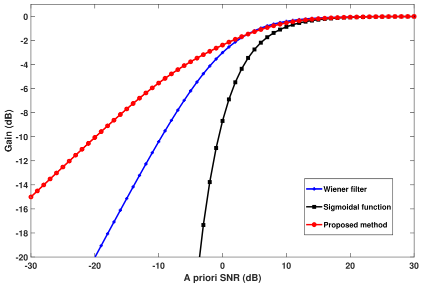

FIGURE 3. Comparison of the noise suppression functions with different a priori SNRs.

λ is the smoothing factor whose value ranges from 0 to 1. The variance of the acoustic signal can be defined as

$$\sigma^2_{a\omega}(\tau_s, \omega) = E[x^4(\tau_s, \omega)] = 2\sigma_x^2(\tau_s, \omega) \qquad (21)$$

For the analytic speech signal, the fourth order expected value in (14) could be expressed as (22) in accordance with [4]

$$E\left[\mid X_a^4(\tau_s, \omega) \mid\right] = 2\sigma_x^2(\tau_s, \omega) * \sigma_x^2(\tau_s, \omega) \qquad (22)$$

Using (4), (22), the final noise suppression function $V(\tau_s, \omega)$ that minimizes (16) is given by

$$V(\tau_s, \omega) = \sqrt{\frac{\gamma^2(\tau_s, \omega) + \gamma(\tau_s, \omega)}{\gamma^2(\tau_s, \omega) + 4\gamma(\tau_s, \omega) + 1}} \qquad (23)$$

Comparing $V(\tau_s, \omega)$ in (23) with wiener function $W(\tau_s, \omega)$ in (3) and sigmoidal function $g(\tau_s, \omega)$ in (6) and observing the same in Fig. 3, it is evident that the PM is the most allowable (less aggressive) suppression function.

D. PROCESSING STEPS

We sampled the noisy speech signal with a sampling rate of 44100 Hz and transformed it into a frequency domain with the FFT of two times the window length, considering a 20ms frame length. Additionally, a frame-shift of 12 ms was applied. The extracted speech was windowed and transferred to FFT for getting spectral analysis. The absolute value of the spectral bands served as the magnitude. Similarly, the phase was extracted for reconstructing the original signal.

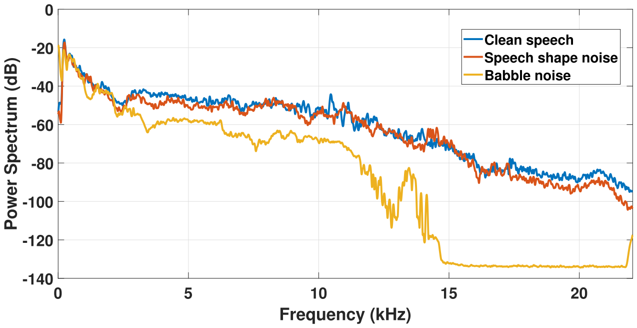

FIGURE 4. Power spectrum of speech shape noise and 4-talker babble noise. To synthesize the testing data, these two noises were used.

An a priori SNR ($\gamma(\tau_s, \omega)$) was calculated using a voice activity detector and a decision direct method [31], [32], and was used as the basis for evaluating the proposed and WF functions. In some studies [4], [33], $\gamma(\tau_s, \omega)$ was calculated from the available clean and noise samples individually. However, in this case, we calculated $\gamma(\tau_s, \omega)$ from the noisy-speech spectrum for practical purposes. We computed the enhanced speech from the proposed method along with the WF.

Verifying the efficacy of the PM on actual cochlear implants can be complicated by various factors such as the availability of neural survival, duration of the deafness, insertion depth, etc [34]. The above factors can confound the outcome, so it would be better to test it with acoustic simulation before testing it on actual CI users. If simulation results are positive, the algorithm can be tested on actual CIs as well. Vocoders are commonly used to replicate some of the characteristics of CI signal processing in CI research [23].

We processed the estimated speech through a 16-channel sinewave vocoder (CI simulator) as shown in Fig. 2. The frequency range of the channels was selected using a gamma-tone filter bank with a range of 80 to 7562 Hz [35], [36]. To estimate the performance of the PM, we conducted both perceptual and acoustic evaluations for speech in noise.

We selected two different noises here: speech shape noises and 4-talker babble noise, and their power spectrum was compared with clean speech spectra shown in Fig. 4.

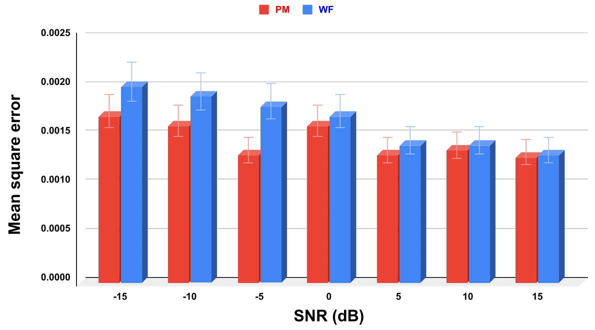

FIGURE 5. WF (blue) and PM (red) represent the mean square error concerning input SNR.

III. SIMULATION RESULTS

A. OBJECTIVE EVALUATION OF THE NOISE SUPPRESSED FUNCTIONS USING MSE

The proposed method was compared with the WF and sigmoidal function based on the mean square error between the clean signal and estimated signal envelopes.

$$MSE_{env}(s,c) = \frac{1}{T}\sum_{n=1}^{T}\Big[\mid X_a(n)\mid - \mid \hat{X}_a(n)\mid\Big]^2 \qquad (24)$$

and the MSE between desired and estimated signal

$$MSE_{signal}(s,c) = \frac{1}{T}\sum_{n=1}^{T}\Big[X(n) - \hat{X}(n)\Big]^2 \qquad (25)$$

where T represents the total number of samples present in each noisy speech, and s represents the total number of noisy sentences. Considering $c = 16$ channels and $s = 49$, a total of 784 samples were used for evaluating both (24) and (25). The evaluated mean square errors shown in Fig. 5 indicate that at different SNR levels, the PM (red) offers lower MSE values than WF (blue).

Moreover, we observe that the relative effectiveness of the proposed method for predicting the speech envelope (as compared to the WF) has increased when the SNR decreases. These results support the theory presented in the previous sections II and II.C. Since the WF attempts to reduce the MSE between clean and its estimated signal, the proposed method can be used to estimate the MSE between the clean and the estimated envelopes.

B. SPEECH INTELLIGIBILITY OF COCHLEAR IMPLANTS

In this study, we evaluated the performance of the proposed method objectively, using speech to reverberation modulation energy ratio (SRMR). Specifically, the SRMR-CI is optimized for evaluating the CI signal processing [37]. Fig. 6 shows that speech intelligibility decreases with unprocessed speech, WF, and sigmoidal functions at low SNR levels, particularly $SNR < -5dB$ compared to PM.

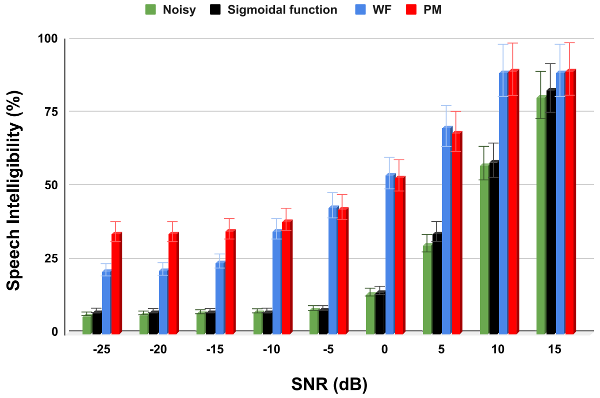

FIGURE 6. Speech intelligibility of CIs according to SRMR-CI metrics with different SNR levels.

TABLE 1. Normal hearing Participants details.

Participant	Age	Gender
NH1	21	Male
NH2	29	Female
NH3	25	Male
NH4	29	Female
NH5	22	Male
NH6	24	Female

C. PSYCHOACOUSTIC EXPERIMENTS

The output speech stimuli of the CI simulator were presented to the normal hearing (NH) volunteers through headphones at the most comfortable level (40 dB speech recognition threshold (SRT)) for conducting a psychoacoustic test. Similarly, ESTOI was used to determine the speech intelligibility of the estimated signal and target signal in acoustic evaluation tests.

1) PARTICIPANTS

This study included six NH individuals with no previous complaints of hearing problems. The sample size used in the present study fulfills the minimum required sample size for psycho-physical research [38]. The participants were 25 years old on average (with a 3.4-year standard deviation) as shown in Table 1. The individuals have given written permission before participating in this study, by following the Helsinki Declaration. The local Ethics Committee has given its approval to the study (Approval Number: NITK/EC/Ph.D/284/2021)

2) STIMULI PRESENTATION

For this experiment, the given input signal is a noisy speech signal. Two different noises, 4-talker babble noise and speech shape noise are added to clean speech at different SNRs (15, 10, 05, 0, −5, −10, −15) in dB. The pre-processed signals in MATLAB were given to NH volunteers via Sennheiser HD280pro headphones. A practice trial has been given to all participants to avoid potential learning effects. Once the individuals had become familiar with the task, they were subjected to the actual perceptual test. The sentence list for each signal processing condition was randomized for each participant. The testing sequence was also randomly assigned

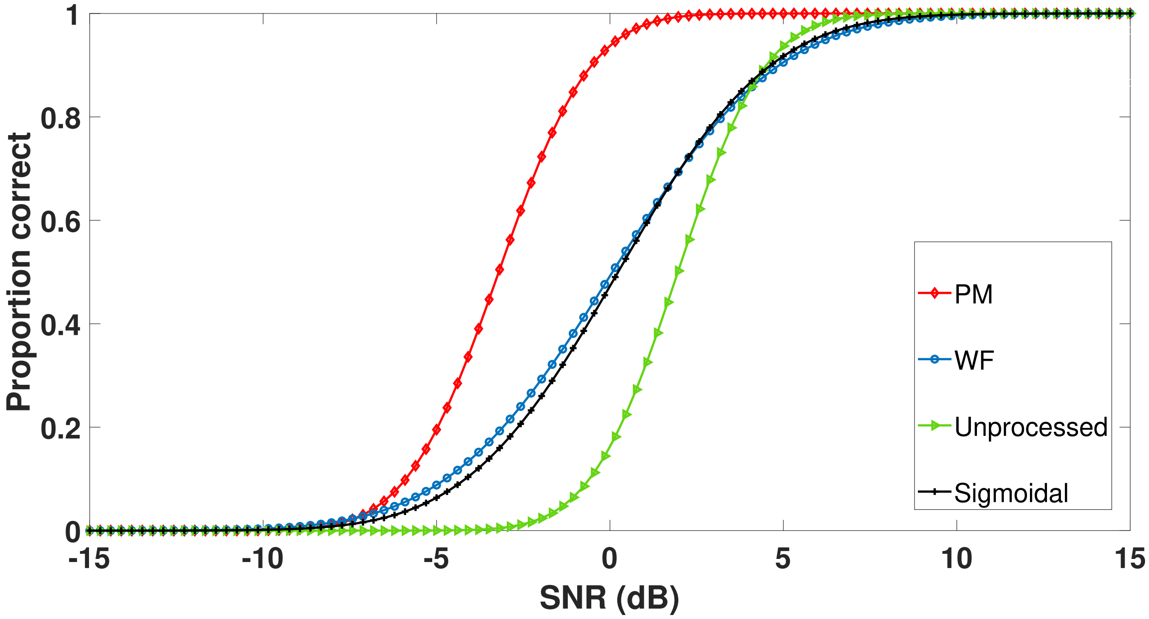

FIGURE 7. The psychometric plots with the speech recognition ability of volunteers for different conditions.

to each participant. The speech recognition test in noise had 7 different lists [39], with each list having 7 sentences, and each sentence having 5 keywords. These clean speech sentences have information up to 10kHz as shown in Fig. 8 (f). The standardized QuickSIN protocol [39] wherein, The first sentence of every list begins at +15 dB SNR, and the remaining sentences were given with decreasing SNR by +5dB sequentially. Participants were required to identify the words in the sentences as they heard. The responses were collected in written form for further evaluation. We calculated the total number of keywords identified correctly by every participant in each method. The speech recognition threshold in noise (SRTN) was calculated using Finney's (1952) Spearman Karber Equation given by:

$$SRTN = i + (d/2) - (d * correct/W) \tag{26}$$

where i = initial presenting SNR
d = step size (+5dB)
W= identified keywords per decrement with SNR
correct = number of key words correctly identified

3) PERCEPTUAL MEASURE WITH SPEECH SHAPE NOISE

SRTN was determined using the Spearman Karber equation (Finney, 1952) based on the total correct scores computed for each SNR. The calculated SRTN goes through statistical analysis, so one-way ANOVA with repeated measurements was used to investigate the noise reduction effect. The F-statistic from a repeated measures ANOVA is reported as:

$$F(df_{between},\ df_{within}) = F_{value},\ p = p_{value} \tag{27}$$

where df is degrees of freedom between the methods and within the methods. If the p-value is 0.05, the F value has a 5% chance of being incorrect and a 95% chance of being true. The hypothesis test is significant statistically if the P value is lower than the significance level (0.05). Noise reduction had a significant main effect on speech recognition in the presence of speech-shaped noise (F(3,15) = 9.56, p<0.001). Since ANOVA revealed a significant difference, we conducted pairwise comparisons. For pairwise comparisons, a series of one-tailed paired 't' tests with the alternate hypothesis of, "measure 1 is less than measure 2" was performed. Table 2 depicts the variables representing measure 1 and measure 2 for the comparisons as well as the corresponding 't' & 'p' values. A point noteworthy of statistical inference is that the smaller the SRTN value better is the performance. The analysis revealed that the SRTN of the PM is significantly better than without noise reduction (Unprocessed) with t (t-distribution [40]) = −5.809, p = 0.001. In addition, the SRTN with the WF is considerably better than the Unprocessed with t = −2.219, p = 0.039. Hence, the analysis revealed that both PM and WF have significantly improved speech recognition with speech shape noise as compared to that without noise reduction. However, the proposed method has shown significant improvement in SRTN results compared to the traditional wiener filter with t = −2.294, p = 0.035. As shown in Fig. 7, the PM (red) was significantly more effective in providing better SRTN compared to the WF (blue) and sigmoidal function (black). Hence, Table 3 indicates that the PM had a higher impact on improving SNR than the WF.

TABLE 2. Quality assessment based on statistical tests of psychoacoustic experiments.

Measure 1 vs. Measure 2	t	p
Proposed method vs. Unprocessed	-5.809	**0.001**
Wiener filter vs. Unprocessed	-2.219	**0.039**
Sigmoidal function vs. Unprocessed	-2.697	**0.021**
Proposed method vs. Sigmoidal function	-2.411	**0.03**
Proposed method vs. Wiener filter	-2.294	**0.035**

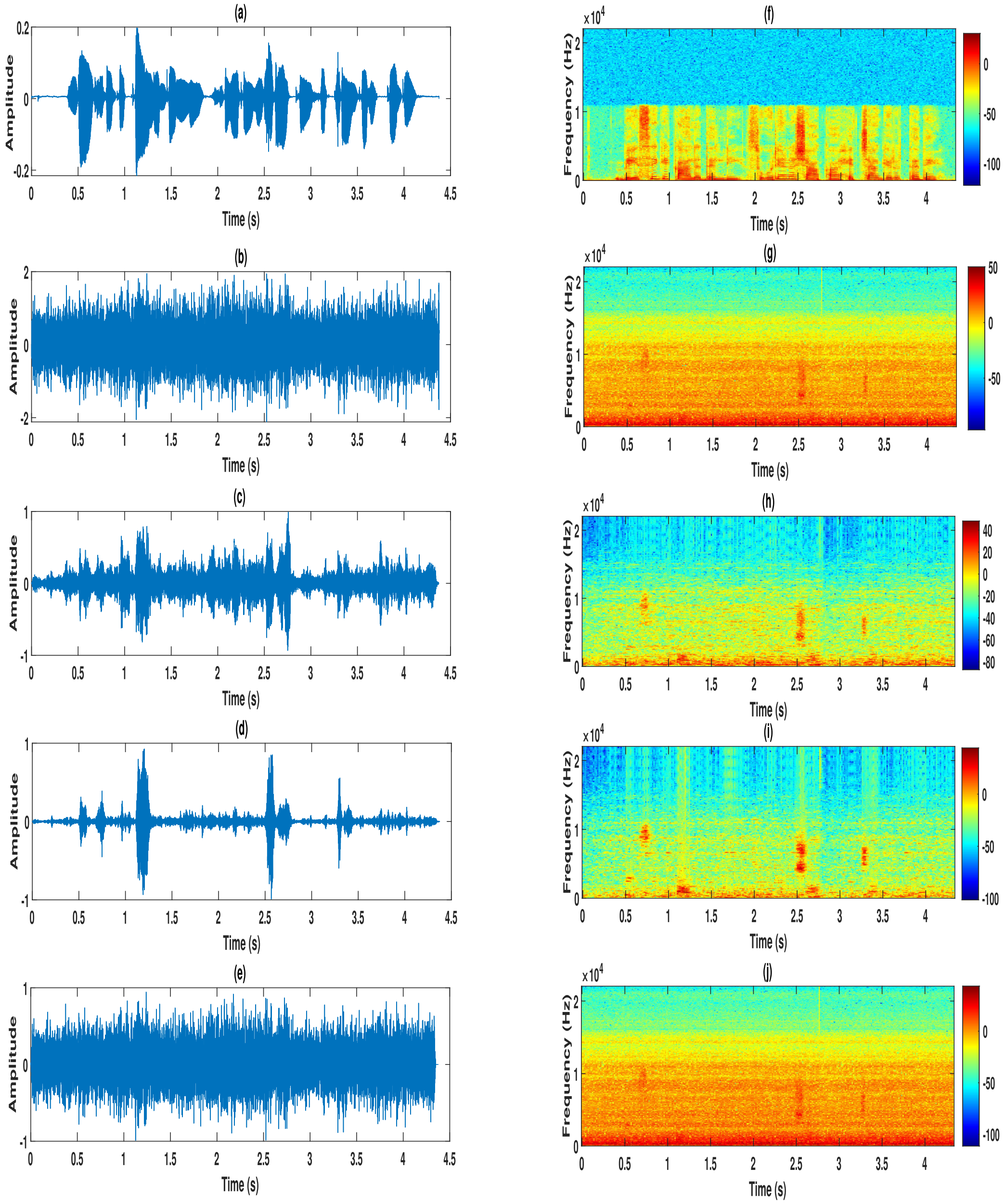

FIGURE 8. The waveform of (a) clean, (b) noisy, enhanced speech signals by (c) PM, (d) WF, and (e) Sigmoidal function. Spectrogram representation of (f) clean, (g) noisy, and enhanced speech signals by (h) PM, (i) WF, and (j) Sigmoidal function.

TABLE 3. The average SRTN and standard deviation of each noise-reduction method with speech shape noise.

Method	SRTN in dB	Standard deviation
Proposed method	**-3.167**	2.251
Wiener filter	0.167	2.338
Sigmoidal function	0.23	1.862
Unprocessed	2.833	1.033

Fig. 8 shows the waveform and the spectrogram of clean speech, noisy speech (noise at −5 dB), and noisy speech modified by the PM, WF, and sigmoidal function. Here, the magnitude of a (noisy/processed) speech envelope at every time interval is related to the intensity. The spectrogram obtained by the WF (i) and PM (h) seems to have improved the signal strength compared with the spectrogram obtained using the sigmoidal function (j) and the unprocessed (g).

4) PERCEPTUAL MEASURE WITH BABBLE NOISE

A masker of four-talker Kannada language babble was selected. This study intends to reflect the properties of actual listening. Further, babble efficiently reduces the intensity of amplitude modulation of speech over steady spectrum noise. A similar analysis was conducted for speech babble noise as well. The determined SRTN underwent statistical analysis, and the noise reduction effect was measured using a one-way ANOVA with repeated measurements. There was no significant main effect of noise reduction block on speech

TABLE 4. The average SRTN and standard deviation of each noise-reduction method with 4 talker babble noise.

Method	SRTN in dB	Standard deviation
Proposed method	**3.667**	1.472
Wiener filter	4.167	1.366
Sigmoidal function	4.16	1.211
Unprocessed	3.833	2.338

TABLE 5. ESTOI values (D) for each noise reduction method with speech shape noise.

SNR in dB	Unprocessed	Sigmoidal function	WF	PM
15	0.79	0.79	0.805	**0.83**
10	0.68	0.678	0.69	**0.721**
5	0.54	0.54	0.56	**0.59**
0	0.38	0.39	0.40	**0.44**
-5	0.23	0.239	0.25	**0.29**
-10	0.13	0.13	0.15	**0.17**
-15	0.06	0.066	0.074	**0.093**

recognition score with babble noise F(2,10) = 0.115, p = 0.893. However, the observation of mean values indicates that the SRTN with the proposed method is slightly better than the traditional WF results shown in Table 4. Hence, compared to WF and Unprocessed, PM offers a marginal improvement in speech recognition when speech babble is present.

D. ACOUSTIC ASSESSMENT OF SPEECH IN NOISE

The Extended Short-Time Objective Intelligibility (ESTOI) metric was applied to assess the objective evaluation of speech intelligibility in noise, based on the correlation between processed noisy speech and clean speech. ESTOI is highly relevant to human speech intelligibility, according to previous studies [21]. ESTOI values are evaluated in terms of speech intelligibility index values (D) at various SNRs. The D values range from 0 to 1, and higher values suggest better speech intelligibility. An ESTOI score is computed in three steps: (1) Each subband's temporal envelope is obtained after passing the signals through a filter bank of one-third octave; (2) the distance between the clean speech and processed speech short-time envelope spectrograms is estimated after time and frequency normalization, resulting in intermediate indices for short-time intelligibility; (3) the final intelligibility index D is derived by averaging the intermediate indices. [21] provides more information on the three steps of the ESTOI measurement.

1) OBJECTIVE EVALUATION OF SPEECH INTELLIGIBILITY

The speech ineligibility index (D) values were computed to six different lists with two noise conditions for three noise reduction strategies at different SNRs (+15 dB, +10 dB, +05 dB, 0, −5 dB, −10 dB, −15 dB). A general trend was observed by analyzing the D values. The D values also decreased when the SNR decreased from +15 dB to −15 dB in the three methods.

Fig. 9 shows the average ESTOI scores at seven different SNR levels for the speech shape noise. Table 5 shows the average ESTOI scores (D) at seven different SNR conditions with the speech shape noise. Herewith speech shape noise mask, maximum D values were obtained for all SNR levels of the proposed method compared to WF and sigmoid function. Similarly, the proposed method's speech intelligibility (D) values were nearly identical to the other methods when dealing with babble noise.

IV. DISCUSSION

Compared to the WF time-frequency mask, the proposed algorithm implementation requires extra 3 multiplications, 2 additions, and 1 square root calculation. This is due to the fact that the proposed time-frequency mask was derived by minimizing the MSE between the squared envelopes of the enhanced speech and its clean speech. We can understand the complexity of the proposed algorithm (23) by comparing its equation to that of the typical WF (3). However, the proposed method outperforms the WF method in terms of speech intelligibility at a price of comparable complexity.

In Fig. 3, the noise suppression of the PM is softer than that of the WF method. CIs requires more aggressive noise suppression [8], [41]. However, aggressive noise suppression should only be implemented if it preserves speech components. However, any implementation of an aggressive WF method would compromise the speech component as well, thereby reducing speech intelligibility. The listeners with CI rely on the envelope for speech intelligibility. Therefore the preservation of temporal envelope is essential for speech intelligibility. However, aggressive noise reduction [42], might have a negative impact on the envelope. Earlier studies also have shown that less aggressive noise reduction resulted in better speech intelligibility than the more aggressive gain suppression function [4], [43]. This supports our findings wherein the gain function derived in the current study is less aggressive than the WF as well. Therefore, a good noise reduction algorithm should provide an optimum trade-off between the magnitude of noise suppression and preservation of speech cues.

Perceptual data indicated that the speech recognition scores improved significantly with WF noise suppression, especially with PM. The pairwise comparisons revealed that the speech recognition scores were significantly better with PM than WF and sigmoidal function. The sigmoidal function does not work for negative SNR levels, which is a crucial region for people with hearing impairments, as shown in the noise suppression function and the perceptual data. It is well-known that the traditional WF can suppress noise, but due to its aggressive nature, some of the speech's spectral content is also lost during noise suppression, as shown in Fig. 8. Comparison of clean speech (a) and Enhanced Speech using WF (d) of Fig. 8 reveals the loss of target speech component which would have negatively affected the speech intelligibility. On the other hand, the PM provides an optimum trade-off between noise suppression and speech intelligibility. Hence, PM suppresses noise while preserving important cues for speech intelligibility, resulting in a better speech recognition score than the other two methods. We observed by the mean data (as shown in Table 3) that the proposed method improves the speech recognition threshold in noise (SRTN) by 6 dB

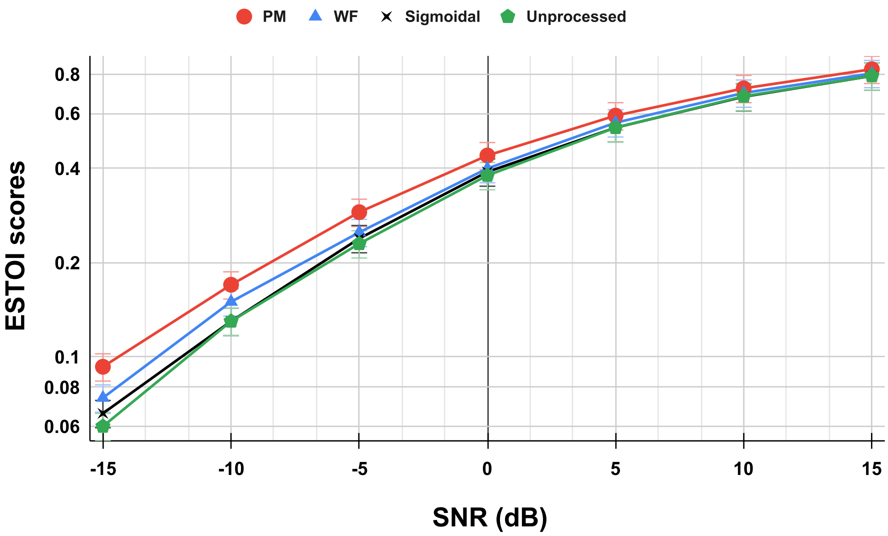

FIGURE 9. The plot of ESTOI scores for speech signals corrupted by speech shape noise at different SNR levels.

SNR compared to the unprocessed (noisy). This may lead to an improvement in speech recognition by almost 60% in real-life situations [44].

A similar analysis was done with speech babble noise. However, the statistical analysis and subjective analysis indicated that all three methods do not improve the SNR because all three methods have been implemented in conjunction with the voice activity detector. The background noise itself is speech, and the algorithm detected the noise segment based on the voice activity. However, the speech babble is voice-based, so the algorithm fails to distinguish target and mask signals. In real-time applications, most of the algorithms fail when background noise is speech itself. In such a scenario, it is ideal to provide more cues to the auditory system to segregate target speech and noise like temporal fine structures. Hence, in one of our previous investigations [36], we have proposed how effectively TFS cues can be encoded to improve speech recognition in noise, so we recommend these MSE minimization methods for improving SNR in non-speech noise scenarios. However, for speech noise situations there is a need to investigate another method or provide more cues for segregating speech and noise.

V. CONCLUSION

In this current study, we have proposed a novel noise reduction technique, and its performance was compared with traditional WF and sigmoidal functions. Overall perceptual and objective analyses indicated that PM is more efficient in improving speech intelligibility when compared to sigmoidal function and traditional WF. Thus, the proposed noise reduction technique has potential implication in CI and a further study can be conducted on actual CI users.

REFERENCES

[1] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, "Cochlear implants: System design, integration, and evaluation," *IEEE Rev. Biomed. Eng.*, vol. 1, pp. 115–142, 2008.

[2] J. Wouters, H. J. McDermott, and T. Francart, "Sound coding in cochlear implants: From electric pulses to hearing," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 67–80, Mar. 2015.

[3] K. Nie, G. Stickney, and F.-G. Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 1, pp. 64–73, Jan. 2005.

[4] R. A. Chiea, M. H. Costa, and J. A. Cordioli, "An optimal envelope-based noise reduction method for cochlear implants: An upper bound performance investigation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1729–1739, 2021.

[5] J. J. Remus and L. M. Collins, "The effects of noise on speech recognition in cochlear implant subjects: Predictions and analysis using acoustic models," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 18, pp. 1–12, Dec. 2005.

[6] A. J. Spahr, M. F. Dorman, and L. H. Loiselle, "Performance of patients using different cochlear implant systems: Effects of input dynamic range," *Ear Hearing*, vol. 28, no. 2, pp. 260–275, Apr. 2007.

[7] C. W. Turner, B. J. Gantz, C. Vidal, A. Behrens, and B. A. Henry, "Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing," *J. Acoust. Soc. Amer.*, vol. 115, no. 4, pp. 1729–1735, Apr. 2004.

[8] S. J. Mauger, P. W. Dawson, and A. A. Hersbach, "Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction," *J. Acoust. Soc. Amer.*, vol. 131, no. 1, pp. 327–336, Jan. 2012.

[9] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *J. Acoust. Soc. Amer.*, vol. 133, no. 3, pp. 1607–1614, Mar. 2013.

[10] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[11] R. Koning, N. Madhu, and J. Wouters, "Ideal time–frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 331–341, Jan. 2015.

[12] R. Koning, I. C. Bruce, S. Denys, and J. Wouters, "Perceptual and model-based evaluation of ideal time-frequency noise reduction in hearing-impaired listeners," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 687–697, Mar. 2018.

[13] F. Henry, M. Glavin, and E. Jones, "Noise reduction in cochlear implant signal processing: A review and recent developments," *IEEE Rev. Biomed. Eng.*, early access, Jul. 7, 2021, doi: 10.1109/RBME.2021.3095428.

[14] A. A. Hersbach, K. Arora, S. J. Mauger, and P. W. Dawson, "Combining directional microphone and single-channel noise reduction algorithms: A clinical evaluation in difficult listening conditions with cochlear implant users," *Ear Hearing*, vol. 33, no. 4, pp. e13–e23, Jul. 2012.

[15] G. L. Mourão, M. H. Costa, and S. Paul, "Speech intelligibility for cochlear implant users with the MMSE noise-reduction time-frequency mask," *Biomed. Signal Process. Control*, vol. 60, Jul. 2020, Art. no. 101982.

[16] O. ur Rehman Qazi, B. van Dijk, M. Moonen, and J. Wouters, "Speech understanding performance of cochlear implant subjects using time–frequency masking-based noise reduction," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1364–1373, May 2012.

[17] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 63–72, Jan. 2013.

[18] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1568–1578, Jul. 2017.

[19] N. Y.-H. Wang, H.-L.-S. Wang, T.-W. Wang, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao, "Improving the intelligibility of speech for simulated electric and acoustic stimulation using fully convolutional neural networks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 184–195, 2021.

[20] R.-Y. Tseng, T.-W. Wang, S.-W. Fu, C.-Y. Lee, and Y. Tsao, "A study of joint effect on denoising techniques and visual cues to improve speech intelligibility in cochlear implant simulation," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 4, pp. 984–994, Dec. 2021.

[21] J. D. Crew and J. J. Galvin, "Channel interaction limits melodic pitch perception in simulated cochlear implants," *J. Acoust. Soc. Amer.*, vol. 132, no. 5, Nov. 2012, Art. no. EL429.

[22] Q. Mesnildrey, G. Hilkhuysen, and O. Macherey, "Pulse-spreading harmonic complex as an alternative carrier for vocoder simulations of cochlear implants," *J. Acoust. Soc. Amer.*, vol. 139, no. 2, pp. 986–991, Feb. 2016.

[23] P. C. Loizou, "Speech processing in vocoder-centric cochlear implants," in *Cochlear and Brainstem Implants* (Advances in Oto-Rhino-Laryngology), vol. 64, 2006, pp. 109–143.

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

[25] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech Language Process.*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[26] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, p. 289.

[27] Y. Hu, P. C. Loizou, N. Li, and K. Kasturi, "Use of a sigmoidal-shaped function for noise attenuation in cochlear implants," *J. Acoust. Soc. Amer.*, vol. 122, no. 4, Oct. 2007, Art. no. EL128.

[28] Y. Lu and P. C. Loizou, "Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1123–1137, Jul. 2011.

[29] A. Hjorungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2740–2746, Jun. 2007.

[30] J. Sohn, N. Soo Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[31] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[32] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[33] Y. Hu and P. C. Loizou, "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," *J. Acoust. Soc. Amer.*, vol. 127, no. 6, pp. 3689–3695, Jun. 2010.

[34] A. Kan, C. Stoelb, R. Y. Litovsky, and M. J. Goupell, "Effect of mismatched place-of-stimulation on binaural fusion and lateralization in bilateral cochlear-implant users," *J. Acoust. Soc. Amer.*, vol. 134, no. 4, pp. 2923–2936, Oct. 2013.

[35] W. Ngamkham, C. Sawigun, S. Hiseni, and W. A. Serdijn, "Analog complex gammatone filter for cochlear implant channels," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 969–972.

[36] V. Poluboina, A. Pulikala, and A. N. P. Muthu, "Contribution of frequency compressed temporal fine structure cues to the speech recognition in noise: An implication in cochlear implant signal processing," *Appl. Acoust.*, vol. 189, Feb. 2022, Art. no. 108616.

[37] J. F. Santos and T. H. Falk, "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2197–2206, Dec. 2014.

[38] A. J. Anderson and A. J. Vingrys, "Small samples: Does size matter," *Investigative Ophthalmol. Visual Sci.*, vol. 42, no. 7, pp. 1411–1413, 2001.

[39] M. Avinash, R. Meti, and U. Kumar, "Development of sentences for quick speech-in-noise (QuickSin) test in Kannada," *J. Indian Speech Hear Assoc.*, vol. 24, pp. 59–65, Jan. 2010.

[40] NIST. *Critical Values of the Student's T Distribution*. Accessed: Jun. 2022. [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm

[41] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[42] K. Chung, "Effective compression and noise reduction configurations for hearing protectors," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. 1090–1101, Feb. 2007.

[43] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.

[44] T. Venema, "Three ways to fight noise: Directional mics, DSP algorithms, and expansion," *Hearing J.*, vol. 52, no. 10, p. 58, Oct. 1999.

VENKATESWARLU POLUBOINA received the B.Tech. degree in electronics and communication engineering and the M.Tech. degree in embedded systems from Jawaharlal Nehru Technological University, Anantapur, India, in 2013 and 2018, respectively. He is currently pursuing the Ph.D. degree with the National Institute of Technology Karnataka, Surathkal, India. He has two years of teaching experience and six months of experience in automation. His current research interests include speech signal processing, machine learning, and deep learning.

APARNA PULIKALA (Senior Member, IEEE) has been associated with NITK, Surathkal, since 2002, under various capacities, where she has been working as an Assistant Professor, since 2008. She has presented a number of research papers at various international conferences. She has published more than 28 research papers in various journals and conference proceedings. She is actively involved in research activities in the area of signal processing for ten years. Her research interests include biomedical signal processing, signal compression, audio and speech processing, computer architecture, and embedded systems.

ARIVUDAI NAMBI PITCHAI MUTHU has been associated with the Kasturba Medical College, Mangaluru, since 2008, under various capacities, where he has been working as an Associate Professor, since 2017. He has presented a number of research papers at various international conferences. He has published more than 25 research papers in various journals and conference proceedings. He has been actively involved in research activities in signal processing for ten years. His research interests include psychoacoustics, digital signal processing, auditory prosthesis, and implantable devices.