A Two-Stage Deep Neuroevolutionary Technique for Self-Adaptive Speech Enhancement

RYAN LEBLANC, (Student Member, IEEE), AND SID AHMED SELOUANI, (Senior Member, IEEE)
LARIHS Lab, Université de Moncton, Shippagan, NB E8S 1P6, Canada
Corresponding author: Ryan LeBlanc (el7765@umonton.ca)
This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2018-05221.

ABSTRACT This paper presents a novel self-adaptive approach for speech enhancement in the context of highly nonstationary noise. A two-stage deep neuroevolutionary technique for speech enhancement is proposed. The first stage is composed of a deep neural network (DNN) method for speech enhancement. Two DNN methods were tested at this stage, namely, both a deep complex convolution recurrent network (DCCRN) and a residual long short-term memory neural network (ResLSTM). The ResLSTM method was combined with a minimum mean-square error method to perform a preliminary enhancement. The ResLSTM network is used as an a priori signal-to-noise ratio (SNR) estimator. The second stage implements a self-adaptive multiband spectral subtraction enhancement method using tuning optimization based on a genetic algorithm. The proposed two-stage technique is evaluated using objective measures of speech quality and intelligibility. The experiments are carried out using the NOIZEUS noisy speech corpus using conditions of real-world stationary, colored, and nonstationary noise sources at multiple SNR levels. These experiments demonstrate the advantage of building a cooperative approach using evolutionary and deep learning-based techniques that are capable of achieving robust speech enhancement in adverse conditions. Indeed, the experimental tests show that the proposed two-stage technique outperformed a baseline implementation using a state-of-the-art deep learning approach by an average 13% and 6% improvement for six noise conditions at a −5 dB and a 0 dB input SNR, respectively.

INDEX TERMS Adaptive signal processing, artificial neural networks, evolutionary computation, speech enhancement.

I. INTRODUCTION
Increased adaptivity is an important subject of speech enhancement research, focusing on dealing with nonstationary noises. The difficulty of a speech enhancement task can largely be related to the stationarity of the noise. Noise signals are nondeterministic and can be better categorized as stationary or nonstationary. A stationary noise signal is generated by a process that has statistical properties that do not change over time (e.g., a fan blowing in the background). Inherently, suppressing nonstationary noise (e.g., multitalker babbles) is a more difficult task than suppressing stationary noise.

Very few methods of speech enhancement have been shown to be effective at reducing noise from highly nonstationary environments [1]. The main conventional categories of single-channel algorithms are spectral subtractive methods, statistical-model-based methods, and subspace methods [2]. Other approaches have also been investigated using empirical mode decomposition [3] and deep neural networks, which have been the focus of the most recent research efforts in the field [4]–[7].

Most approaches to single-channel speech enhancement are highly dependent on the adequate selection of training material (i.e., they are training-dependent). This training-dependency limits adaptivity, and results in poor performance for unknown noise conditions. While training dependency is a well-known challenge for DNN methods, it is also true for conventional methods as well. For most conventional methods, the tuning of the speech enhancement algorithm is a method of adapting to different noise conditions. The tuning entails selecting values for the algorithm’s parameters, which gives an optimal performance for the selected noise conditions. To perform the tuning, scenarios are chosen (i.e., noise conditions), on which the algorithm’s performance is optimized. However, to accommodate a large
number of scenarios, a compromise must be made on the algorithm’s performance. This means that tuning for highly nonstationary noise conditions, which encompasses a broad range of scenarios, can be difficult. This compromise results in a lack of adaptivity for speech enhancement algorithms. The same can be said for the training of deep learning models, where the adequacy of training material becomes critical to its robustness.

To our best knowledge, the use of a metaheuristic approach to tuning optimization has not been implemented in an adaptive control scheme for single-channel speech enhancement. A training-independent approach was shown in [8], where a self-adaptive tuning approach was shown to be advantageous. Building on this success, the present work seeks to expand further on the training-independent approach of self-adaptive control for single-channel speech enhancement algorithm tuning. In addition, no research has been done on collaborative approaches to single-channel speech enhancement using both a training-dependent method and a training-independent method.

The purpose of this paper is twofold. The first purpose is to evaluate a tuning optimization-based self-adaptive control scheme for single-channel speech enhancement. This adaptive control scheme combines the tuning optimization problem introduced in [9] for automatic tuning, the use of a feed-forward adaptive control scheme for single-channel speech enhancement shown in [10], and a metaheuristic approach shown for dual-channel adaptive control schemes [11]. This control scheme implementation is evaluated by proposing a generalized adaptive control framework, expanding on the work performed in [8]. The new framework is implemented and evaluated through multiband spectral subtraction (MBSS) and generalized subspace (GS) techniques. The second purpose is to evaluate a novel approach to increased adaptivity for training-dependent methods of speech enhancement. The proposed approach is a hybrid model for speech enhancement using both deep learning and the proposed self-adaptive framework. The purpose is to ascertain whether a training-dependent method could be complemented by a training-independent method to increase the adaptivity of a training-dependent method. This approach can be interpreted as a method to improve noise adaptivity with no modifications needed to the training-dependent method itself. This hybrid model was achieved by proposing a two-stage deep neuroevolutionary technique for speech enhancement. This novel two-stage deep neuroevolutionary technique was tested by evaluating its capacity to perform speech enhancement in multiple noise conditions, including highly nonstationary noise conditions.

The organization of the paper is as follows. Section II overviews related works that have investigated approaches to increase adaptivity for speech enhancement. Section III overviews two MBSS- and GS-based speech enhancement algorithms and methods used for their tuning. Section IV details the speech enhancement tuning optimization problem. Section V presents the proposed self-adaptive framework for speech enhancement algorithms. Section VI details the motivation behind an evolutionary approach to the self-adaptive framework and describes its implementation. Section VII overviews a deep learning speech enhancement technique and presents the proposed two-stage neuroevolutionary approach for speech enhancement. Section VIII presents the experimental setup used for evaluating the proposed speech enhancement techniques. Section IX presents and discusses the results of the performance evaluation of the proposed methods by using objective measures and comparison with state-of-the-art speech enhancement methods. Section X concludes and gives the perspective of this work.

II. RELATED WORK
A. ALGORITHM TUNING APPROACHES
Manual algorithm tuning is performed experimentally, as shown in [12]. In this work, optimal tuning parameters were determined empirically using the spectral subtraction algorithm for selected training material (i.e., noise conditions). While effective for unchanging noise conditions, manual tuning is time-consuming.

To reduce the time and effort required for manual tuning, automatic tuning for speech enhancement has been proposed [13], [14]. Speech enhancement tuning is treated as an optimization problem, where optimal tuning parameter values are determined automatically using a training set. While effective and efficient, good automatic tuning, such as good manual tuning, relies on the adequacy of the chosen training material.

In [13], a framework was presented for the automatic tuning and evaluation of a speech enhancement algorithm for a communication system integrated into a car. They tested their framework using the steepest gradient descent optimization algorithm and an optimization criterion using a weighted sum of objective speech measures. Their spectral subtraction and various minimum-mean-square error (MMSE) implementations showed that their proposed architecture and optimization methodology resulted in audible improvement and perceptual sound quality.

In [14], a methodology was presented for tuning speech enhancement algorithms automatically, employing the genetic algorithm to solve the tuning optimization problem. Their tuning methodology was employed for a single-channel system for hands-free devices. Their optimization criterion was an average of a composite objective speech measure. They compared an automatically tuned system to a system manually tuned by an expert, which showed great performance advantages for the automatically tuned system.

A summary of the different approaches to speech enhancement algorithm tuning is shown in Table 1. A self-adaptive approach to tuning optimization was proposed in this work to remove the need for manual intervention in retuning a speech enhancement algorithm.
TABLE 1. Summary of approaches to speech enhancement algorithm tuning.

| Tuning approach                | Highlights                                                                 | Limitations                                                                 |
|-------------------------------|---------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Manual tuning [12]            | uses empirically determined parameters                                     | time-consuming, assumes sufficient contextual knowledge of deployed algorithm (training-dependent), and lack of adaptivity imposes compromise on performance |
| Automatic Tuning [13], [14]   | automatic procedure for optimizing speech enhancement tuning; treat tuning as an optimization problem | assumes sufficient contextual knowledge of deployed algorithm (training-dependent), and lack of adaptivity imposes compromise on performance |

TABLE 2. Summary of approaches for adaptive single-channel speech enhancement.

| Single-channel approach          | Highlights                                                                 | Limitations                                                                 |
|---------------------------------|---------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Linear approximation for adaptive approach [12] | adapt tuning parameter to empirically determined linear function | assumes a linear relation to tuning parameter, thus shows a lack of nuance |
| Adaptive Wiener filter [10]    | adapt tuning parameters using feed-forward control scheme, using local context of input signal statistics | assumes relation between sample-to-sample local signal statistics and broader speech enhancement performance; missed opportunity for optimization |
| Noise-adaptive approaches to neural network methods [15] | adapt a trained neural network for untrained noise conditions | training-dependent; risk of catastrophic forgetting |

B. ADAPTIVE SINGLE-CHANNEL APPROACHES

In [12], an adaptive linear estimator was proposed. The linear estimator was guided by an empirically determined tuning parameter value and allowed the tuning parameter to vary frame by frame depending on the detected input SNR, as shown in Fig. 1. Limitations for this approach include assuming that the speech enhancement quality should vary as a function of the SNR and assuming that the tuning parameter should vary linearly.

In [10], an adaptive Wiener filtering method for speech enhancement was proposed. The approach employs a feed-forward control scheme where the filter transfer function is adapted from sample to sample based on the speech signal local statistics, as shown in Fig. 2. Their experimental results show the superiority of the proposed adaptive filtering method in the case of additive white Gaussian noise and colored noise. They did not, however, test for nonstationary noise conditions and did not employ an optimization approach.

Noise-adaptive methods for neural network methods of speech enhancement have been proposed for improving performance for untrained noise conditions. [15] proposed a regulation-based incremental learning strategy while attempting to minimize the possibility of catastrophic forgetting.

A summary of the different adaptive approaches to single-channel speech enhancement is shown in Table 2. The lack of nuance given by linear estimations and the training dependency of neural network methods motivated the pursuit of an optimization-based training-independent approach in this work. The success of a feed-forward control scheme served as inspiration for the basis of this work’s proposed self-adaptive framework.

C. ADAPTIVE NOISE CANCELLATION

Adaptive noise cancelling explores adaptive control schemes applied to dual-channel speech enhancement. The concept of adaptive noise cancellation was introduced in [20] to improve on fixed filter coefficient methods. In its basic form, it uses a reference input signal that contains noise with very little to none of the speech signals. An adaptive filter is then
used to adapt its weights for the noise input. This method of adaptive control minimizes an error signal to converge toward an optimal solution for the filter weights.

The optimization of the adaptive filter weights is performed by an adaptive algorithm. Many researchers have opted for gradient-based solutions such as the least mean square (LMS) algorithm, the normalized LMS (NLMS) algorithm, and the recursive least squares algorithm. Metaheuristic approaches to adaptive noise cancelling have also been shown using a dual-channel setup [11]. These stochastically driven optimization solutions have been shown to give performance increases compared to gradient-based approaches in given conditions. Examples of metaheuristic algorithms employed for adaptive noise cancellation are the particle swarm optimization (PSO) algorithm [16], [17], [19], the cuckoo search algorithm [18], and the bat algorithm [17].

In [16], an adaptive noise cancellation scheme using PSO was shown for a dual-channel setup. They succeeded in their approach by formulating the task of noise cancellation as a coefficient optimization problem. It was found that the proposed PSO approach outperformed the gradient-based LMS and NLMS under various noise conditions.

In [19], a speech enhancement method based on adaptive noise cancellation and PSO was shown. They used the traditional adaptive noise cancelling control scheme to optimize the tuning coefficients of an adaptive filter. They tested the algorithm for three variations of the PSO algorithm, employing a composite speech metric as the objective optimization function. It was found that the proposed approach gave promising results.

A summary of the different approaches to adaptive noise cancelling is shown in Table 3. The success seen with metaheuristic approaches to adaptive noise cancellation motivated the use of a metaheuristic approach to a single-channel speech enhancement adaptive control scheme.

### III. TUNING OF MBSS- AND GS-BASED SPEECH ENHANCEMENT TECHNIQUES

Speech enhancement tuning is performed by manipulating selected parameters to obtain an optimal enhanced speech quality. The ultimate goal of tuning is to adapt speech-enabled systems to real-life conditions where the characteristics of the noise might be changing constantly. Hence, the parameters need to be updated continuously over time, which can be done by using tuning algorithms such as those first proposed in [14]. The first premise for effectively achieving this tuning is to involve an enhancement technique with parameters that are accessible, quantifiable, and easy to modify. This premise justified the choice of the two enhancement techniques for this study, namely, MBSS and GS. The second premise is to perform speech enhancement on a per-frame basis, where tuning parameters can be used for improved performance with specific noise conditions.

Previous techniques for adaptive tuning have been used for both MBSS [21] and GS [22]. In these implementations, tuning parameters were adaptively controlled by a linear function of the input speech’s SNR. In this work, both MBSS and GS were modified to be implemented with self-adaptive control for their tuning parameters.

#### A. TUNING MULTIBAND SPECTRAL SUBTRACTION PARAMETERS

One of the most effective spectral subtraction (SS) algorithms for speech enhancement is the MBSS method [1], [21]. It divides the noisy speech spectrum into \( N \) nonoverlapping bands, and then spectral subtraction is performed on each band independently. The use of this approach was motivated by the fact that noise can affect the speech signal spectrum in a nonuniform manner. The MBSS noise reduction is performed as follows:

\[
|\hat{X}_i(w_k)|^2 = |Y_i(w_k)|^2 - \alpha_i \delta_i |\hat{D}_i(w_k)|^2 \quad b_i \leq w_k \leq e_i \tag{1}
\]

where \( w_k = 2\pi k/N \) (\( k \) varying from 0 to \( N - 1 \)) are the discrete frequencies; \( b_i \) and \( e_i \) are the starting and ending frequency bins of the \( i \)th frequency band; \( \alpha_i \) is the oversubtraction factor; and \( \delta_i \) is a weight for each individual band.

For the proposed self-adaptive approach, the MBSS was performed using (1), and spectral flooring for each individual band was performed as shown in (2). The initial noise estimate was performed using the mean of the initial frames of the noisy speech signal, assuming they were speech-absent. The noise estimate was updated using a voice-activity detector set with a minimum threshold.

\[
|\hat{X}_i(w_k)|^2 = \beta |\hat{D}_i(w_k)|^2 \quad \text{if} \quad |Y_i(w_k)|^2 < \alpha_i \delta_i |\hat{D}_i(w_k)|^2, \tag{2}
\]

where \( \beta \) is a flooring parameter, with \( b_i \leq w_k \leq e_i \).

The clean speech spectrum estimate of the \( i \)th band using MBSS was given by (1). The \( \alpha \) parameter (\( \alpha \geq 1 \)) is used to reduce the amplitude of the broadband peaks, whereas the \( \beta \)
parameter $\beta$ ($0 < \beta \ll 1$) is used to control the amount of residual noise and the amount of perceived musical noise.

The MBSS performance can be modified using $\alpha_i, \delta_i, \beta$ and the number of bands as tuning parameters.

**B. TUNING GENERALIZED SUBSPACE PARAMETERS**

The generalized subspace approach presented in [22] has been shown to be an effective subspace method of speech enhancement [1]. The self-adaptive tuning capacity is provided to the conventional GS by performing the following six steps. A visual representation of the GS method steps is shown in Fig. 3.

**Step 1:** Compute the covariance matrix of the noisy signal $R_y$ and estimate the matrix $\Sigma = R_y^{-1}R_i - I$ using singular value decomposition. Similar to MBSS, the noise covariance matrix $R_n$ is computed using noise samples collected during speech-absent frames. $y$, $n$, and $x$ are the $K$-dimensional noisy speech, noise, and clean speech vectors, respectively.

**Step 2:** Perform eigendecomposition of $\Sigma$ as follows:

$$\Sigma V = V \Lambda_x,$$

where $V$ is the eigenvector matrix, and $\Lambda_x$ is the diagonal eigenvalue matrix of $R_x$.

**Step 3:** Estimate the speech signal subspace dimension $M$ by assuming the eigenvalues of $\Sigma$ are ordered as $\lambda_1^x \geq \lambda_2^x \geq \ldots \geq \lambda_K^x$.

$$M = \arg \max_{1 \leq k \leq K} \{ \lambda_k^x \geq 0 \}$$

where $\lambda_k^x$ is the $k$th diagonal element of $\Lambda_x$.

**Step 4:** Compute the optimal linear estimator $H$ by assuming, as in the multiband approach, that $\mu_k$ can vary per spectral component:

$$q_{kk} = \begin{cases} \frac{\lambda_k^x}{\lambda_k^x + \mu_k}, & k = 1, 2, \ldots, M \\ \frac{\lambda_k^x}{\lambda_k^x + \mu_k}, & k = M + 1, \ldots, K \\ 0, & \end{cases}$$

$$Q = \text{diag}(q_{11}, q_{22}, \ldots, q_{KK})$$

$$H_{opt} = V^{-T}QV^T$$

where $H$ is a $K$-by-$K$ matrix.

**Step 5:** Estimate the enhanced speech signal as follows:

$$\hat{x} = H_{opt} \cdot y,$$

where $\hat{x}$ is an estimate of the clean speech.

The Lagrange multiplier $\mu$ controls the trade-off between residual noise and speech distortion, performing a comparable role to the $\alpha$ parameter in the MBSS method. This parameter was shown to be more effective when varied based on the current noise conditions (e.g., SNR) and thus can be used as a tuning parameter.

**IV. TUNING OPTIMIZATION**

The tuning procedure is an iterative process by which different tuning parameter values are passed to a speech enhancement algorithm and evaluated for their effectiveness on given noise conditions. In manual tuning, iterative optimization process was performed manually, and the evaluation portion was performed subjectively. In automatic tuning, the iterative optimization process was performed computationally, and the evaluation was performed objectively. This section details the tuning optimization, as performed with automatic tuning.

**A. OPTIMIZATION PROBLEM DEFINITION**

Giacobello et al. [9] showed that tuning speech enhancement processing can be viewed as an optimization problem. An optimization problem requires at least one objective function and at least one independent variable to be manipulated. For the speech enhancement tuning optimization problem, the objective function can be selected from the numerous speech quality measures, and the independent variable is the tuning parameter related to the speech enhancement algorithm. Withholding tuning parameter specifics, the speech enhancement tuning optimization problem can be formulated by:

$$\text{minimize } T(\hat{x}[n, v])$$

subject to $L \leq v \leq U$

where $T(\hat{x}[n, v])$ is an objective function used on the output of a speech enhancement algorithm $\hat{x}[n, v]$, $v$ is the tuning parameter vector to be manipulated, $L$ is the lower bound of $v$, and $U$ is the upper bound of $v$.

**B. OBJECTIVE FUNCTION(S)**

The objective function’s role is to measure the quality of the enhanced speech with the utilized tuning parameter value(s). A visual representation of the optimization algorithm’s objective function is shown in Fig. 4. In our tests, the speech
enhancement in this context was done using either MBSS (using (1) and (2)), or GS (using the steps listed in 3).

Multiple objective speech measures exist, each varying in its accuracy to predict speech quality and intelligibility [23]. Four aspects were considered for choosing an objective function:

- the speech metric bias;
- the time sensitivity of the application;
- the minimum number of frames for the objective function to be accurate;
- the need for a clean reference signal (signal-ended vs. double-ended).

Each objective measure can have a bias for certain speech characteristics, and each bias has a potential benefit for certain speech enhancement applications. Computationally intensive measures may not be appropriate for time-sensitive applications, despite being a strong speech quality or intelligibility predictor. Some measures can be more accurate over a larger duration of speech, whereas applying the measure to an individual speech component may lead to inaccuracies. Although double-ended measures are typically more accurate, they are only viable in a testing or training environment.

In this work, both speech quality and intelligibility were prioritized, low complexity measures were preferred, and only short-term measures were chosen (i.e., requiring fewer frames). Double-ended measures were used to represent an ideal scenario, whereas single-ended measures were used to represent a practical setting.

Experiments were carried out by comparing intelligibility-focused measures, quality-focused measures, a multiobjective configuration using both quality- and intelligibility-focused measures, and a single-ended measure. The results of these experiments are shown in section IX.

C. INDEPENDENT VARIABLE(S)

When choosing independent variables (a.k.a. tuning parameters), special consideration was given to the type of variable and their complementarity to other tuning parameters. Three aspects were considered for choosing the independent variables:

- benefit of per-frame variation (if more than one, complementarity);
- physical constraints;
- recommended range in literature.

In this work, per-frame-varying parameters were prioritized, i.e., those yielding better performance when varied per frame, as the noise signal varies. For MBSS, the \( \alpha \) \((\alpha \geq 1)\) and \( \beta \) \((0 < \beta \ll 1)\) parameters were used. Although only \( \alpha \) was recommended to be varied per frame, preliminary tests showed that \( \beta \) was complementary when also varied per frame (i.e., improved performance). For the GS method, the per-frame-varying \( \mu \) \((\mu \geq 1)\) parameter was used.

Constraint functions for independent variables can improve the optimization efficiency and can help avoid local optimums. In this work, upper and lower bounds were chosen in proximity to literature recommended ranges for the tuning parameters. Many experiments have been conducted to identify the optimal range of values for \( \alpha \) (less than 1 to 2) and for \( \beta \) (0.005 to 0.06) [24], [25]. In [22], a recommended range of 1-20 was given for the \( \mu \) with the spectral domain estimator.

Tests were made to determine the effect of increasing a constraint range from a recommended range for a tuning parameter. These tests were performed to verify whether increasing the range from the literature recommendation could yield improved adaptivity. The results of these tests are shown in section IX.

D. CONSIDERATIONS ON THE COMPLEXITY

Computational problems can be classified to better define them and to identify a valid approach to obtain a solution. The ability to solve an optimization problem can vary. Factors include the form of the objective and constraint functions, as well as the number of variables and constraints present. Giacobello et al. [9] described the speech enhancement tuning optimization problem as a nonlinear and nonconvex programming problem of combinatorial nature. This problem analysis referred to a specific, likely complex, objective measure. Whereas less complex objective speech measures may exist, using the more complex classification allows for a broader application.

To solve a computational problem, the proper computational method must be chosen. General nonlinear programming problems can become intractable with a few hundred variables, and no effective methods exist for solving them [26]. Therefore, approaches for these problems require some compromise. This can be done by selecting the appropriate class of algorithms (i.e., hard computing or soft computing). The general goal for speech enhancement is to optimize the perceptible qualities of speech. Speech perception is inherently inexact and subjective in nature. This effectively means that the optimization would likely not have an exact solution. In addition, when passing a certain threshold, a more exact solution might not improve the perception (i.e., there is fault tolerance).

V. SELF-ADAPTIVE TUNING FRAMEWORK FOR SPEECH ENHANCEMENT

A. SELF-ADAPTIVE CONTROL FOR SPEECH ENHANCEMENT TUNING

Speech enhancement processing, which can be described as a linear and time-invariant discrete-time system [2], can be
tuned and controlled. To tackle adverse nonstationary noise conditions, a self-adaptive control scheme was deployed, allowing an algorithm to tune itself for current noise conditions.

Tuning approaches are usually based on linear estimators constrained with bounded values for the key parameters, as shown in Eq. 8, for a tuning parameter $p$. This method of using a linear relation estimate shows limitations in adaptivity since adapting to different noise conditions requires modifying the bounds and the linear estimator’s parameters manually, as well as being influenced by the SNR, a metric with low correlation to speech quality and intelligibility.

$$p = p_0 - \frac{(\text{SNR})}{s} \text{ for lowBound < SNR < highBound} \quad (8)$$

where $p_0$ and $s$ are constants chosen experimentally.

In an attempt to eliminate the need for manual tuning, a self-adaptive framework for speech enhancement is proposed. This framework is proposed to further the work performed in [8], which showed improvement for a self-adaptive implementation of MBSS. The proposed self-adaptive control scheme automatically updates the tuning parameters with optimal values for the speech and noise conditions being actively processed. Optimal tuning parameter values are obtained by performing speech enhancement tuning optimization. The proposed framework uses tuning optimization in a feed-forward control scheme, as shown in Fig. 5. The tuning optimization is performed using noisy speech frames and guidance criteria, such as constraints for the tuning parameters. The tuning optimization was performed synchronously with the speech enhancement algorithm (i.e., no tuning parameters were stored in memory and used for speech enhancement).

**FIGURE 5.** Model for generalized self-adaptive framework.

B. FRAME BUFFERING FOR EFFICIENT SPEECH ENHANCEMENT

Buffering was used to treat a group of neighboring frames in an identical way to improve efficiency. A buffer was used to perform speech enhancement using identical tuning parameters for each frame in the buffer. Speech enhancement is performed in small segments (i.e., frames) so that noise is relatively consistent across the frame. It was thought that using identical tuning parameters for a small number of neighboring frames would increase efficiency while maintaining satisfactory speech enhancement results. The number of frames in the buffer is here referred to as the buffer size.

Tests were performed to determine the ideal number of frames for the buffer. An ideal buffer size of 100 ms was determined, as shown in Fig. 6. Larger buffer sizes yielded less residual noise at the cost of suppressing more speech characteristics. This result coincides with the average duration of a spoken vowel (i.e., 99 ms), determined in [27].

**FIGURE 6.** Spectrograms of speech processed with a self-adaptive framework for various buffer sizes.

C. VOICE ACTIVITY CRITERIA FOR FRAME SELECTION IN TUNING OPTIMIZATION

A voice-activity-based approach was used to improve the tuning optimization process. Whereas tuning optimization could be performed on a single frame, a single frame might not be indicative of speech quality. In [8], tuning optimization was performed using only the frames in the current buffer. It was thought that sufficient voice-active and voice-inactive frames for each tuning optimization would lead to better results. The tuning optimization was then performed on evaluation blocks consisting of the frames in the current buffer and frames from memory (i.e., the previously enhanced frames).
Fig. 7 shows an example of an evaluation block creation process for speech enhancement tuning. This example depicts a three-frame buffer and a minimal criterion of 3 voice-active frames in the evaluation block. Performed tests showed that imposing a minimum of voice-active and voice-inactive frames improved the tuning results. In this work, a minimum evaluation block of 4 voice-active frames, a minimum of 4 voice-inactive frames, and a maximum of 20 frames were used.

VI. EVOLUTIONARY APPROACH TO SELF-ADAPTIVE FRAMEWORK

A. EVOLUTIONARY OPTIMIZATION APPROACH

For speech enhancement tuning optimization, i.e., a possibly intractable problem, a soft computing algorithm was considered. Soft computing methods can solve problems that cannot be solved with hard computing methods [28]. In addition, as discussed in subsection IV-D, given fault tolerance for the problem (i.e., subjectivity of speech quality), a heuristic approach to optimization was considered. Heuristic methods trade optimality for practicality, sacrificing some accuracy for speed. Different soft computing methods have been developed, namely, evolutionary-based methods, fuzzy logic, and swarm intelligence. Evolutionary algorithms are a powerful method to find a good solution for large multimodal problems. Evolutionary optimization algorithms (EOAs) are effective methods of metaheuristic optimization [29].

EOAs are a population-based optimization approach, where each potential solution of a problem is referred to as an individual and a group of individuals is a population. EOAs generate a population from which each individual is tested via a fitness function (i.e., an objective function). The best individuals dictate modifications to the population to converge toward an optimal solution. The evaluation and modifications are repeated over a fixed number of iterations or until the relative change to the solution is below a certain threshold.

Tests were made to evaluate the effectiveness of different EOAs for speech enhancement tuning optimization. The tested algorithms were PSO, genetic algorithm (GA) and differential evolution (DE). The results of these tests are shown in section VIII.

B. IMPLEMENTATION OF EVOLUTIONARY OPTIMIZATION APPROACH

Fig. 8 shows a visualization of the tuning optimization process using an EOA. To better explain this process, an example is given for MBSS. The process begins with generating an initial population, which pools different values of $\alpha_i$ and $\beta$. The population then begins its first iteration, and is tested using the objective function, as explained in section IV.B. The MBSS here, is performed using equations (1), and (2). Evaluating the results from the objective function, the population is then modified to improve on the last population by using the selection, crossover, and mutation functions. These iterations are then repeated until the stopping criteria are met, and optimal tuning parameter values are obtained.

C. OPTIMIZATION PARAMETERS/ZZZZZ/WWW/STOPPING CRITERIA

To further improve the efficiency of solving the speech enhancement tuning problem, optimization stopping criteria were investigated. Optimization stopping criteria provide more insight to the optimization algorithm on what an optimal solution might be. There are no universally optimal stopping criteria values. Tests considering speech enhancement performance and computational efficiency were carried out to determine suitable values for stopping criteria. Values for the function tolerance, the maximum number of iterations and the objective limit were determined. The results of these tests are shown in section VIII.

VII. DEEP NEUROEVOLUTIONARY APPROACH TO SPEECH ENHANCEMENT

In an attempt to improve neural network-based speech enhancement methods, a two-stage hybrid system is proposed. The proposed two-stage hybrid system uses both a neural network-based method and the proposed self-adaptive method in succession. This was attempted to observe whether a training-dependent method and a training-independent (self-adaptive) method could complement each other for improved speech enhancement performance in the context of nonstationary noise.
A. NEURAL NETWORK-BASED SPEECH ENHANCEMENT

Numerous recent studies have shown that neural networks can be used effectively for speech enhancement. They are typically used as nonlinear filters to achieve speech denoising. Early applications saw the use of shallow neural networks [30]. More recent approaches show the advances of deep neural networks being applied to speech enhancement [31]–[35]. Deep learning approaches to speech enhancement have been shown to be very effective.

With a speech enhancement method, one must consider variations in noise levels, noise types and speakers. This can be referred to as dynamicity or robustness. Similar to speech enhancement tuning, a model must be trained considering multiple noisy scenarios. Typically, when there are significant differences in the training material and the test material, there will be degradation in a model’s performance.

Techniques have been proposed to increase noise adaptivity in these methods, such as a regularization-based incremental learning strategy for deep learning methods of speech enhancement [15]. Some approaches, however, pose a risk of model performance loss due to catastrophic forgetting. To the authors’ best knowledge, little research has been done on neuroevolutionary approaches for adaptive speech enhancement in adverse conditions.

B. DEEPPMSE

In the proposed framework, the neuroadaptive method [33] is used. DeepMMSE is an improvement of the system presented in [36], which is an application of deep learning to MMSE (i.e., the DeepXi framework) that was found to be effective for speech enhancement. The DeepXi framework used a residual long short-term memory (ResLSTM) network as an a priori SNR estimator. In the DeepMMSE configuration, an improvement was achieved by employing a temporal convolutional network (TCN) for a priori SNR estimation instead of the ResLSTM network used in DeepXi. Furthermore, a novel DeepMMSE method was proposed using the DeepXi-TCN framework with an MMSE noise periodogram estimator. This later DeepMMSE configuration was shown to be an effective method of speech enhancement, outperforming many deep learning approaches to speech enhancement. The DeepMMSE method was made available at https://github.com/anicolson/DeepXi

The method was trained as detailed in [33]. A large variety of noises were used in the training, including environmental background noise and colored noise recordings. The method was trained for noise conditions ranging from −10 to 20 dB input SNR, in 1 dB increments. In this work, the DeepMMSE method was performed for 16 kHz speech and then downsampled to 8 kHz when taking the objective measures.

C. DCCRN

A second deep learning-based method was used in the proposed framework, namely, the deep complex convolution recurrent network (DCCRN) [37]. This method employs a new network structure design that simulates complex-valued operations, where CNN and RNN structures are employed for complex-value operations. The DCCRN method uses a complex LSTM, complex CNN and complex batch normalization layer in the encoder/decoder. The correlation between magnitude and phase is modeled by the complex module with the simulation of complex multiplication.

The method was trained using the Libri1mix train-360 dataset [38]. The noises used in the training, sampled from WHAM! [39] primarily consists of nonstationary ambient environments such as coffee shops, restaurants, and bars. The method was trained for noise conditions ranging from 0 to 20 dB input SNR, in 5 dB increments. In this work, the...
FIGURE 12. DCCRN complex module [37].

DCCRN method was performed for 16 kHz speech and then downsampled to 8 kHz when taking the objective measures. The method was executed by using a nonofficial implementation from [40].

D. TWO-STAGE HYBRID CONFIGURATION

In approaching unknown noise conditions with a DNN method for speech enhancement, as is typical with nonstationary noise, an imperfect performance is expected. Given possible complications for adaptive approaches to DNN methods discussed in II, a novel two-stage approach is proposed. An overview of the proposed configuration for a two-stage deep neuroevolutionary approach is shown in Fig. 13.

In the first pass, a speech enhancement algorithm will reduce background noise for noisy speech. This process is imperfect and leaves remnant background noise, as well as introduces some distortions, which can be considered newly added noise. A two-stage configuration can be used to further filter noise from noisy speech and help eliminate added noise from a first speech enhancement process.

To evaluate the effectiveness of the two-stage approach, three configurations were tested. In the first baseline configuration, a DNN method was run in succession twice on the same speech flow. The second and third configurations used the DNN method and the proposed self-adaptive framework (MBSS-GA), whose positions were swapped in the first and second stages. Both the DeepMMSE and DCCRN methods were tested in these configurations.

VIII. EXPERIMENTAL SETUP

A. DATA

Several experiments were carried out to evaluate the proposed methods. The experiments were conducted using 13 sentences from the NOIZEUS noisy speech corpus [1]. The corpus uses IEEE sentences downsampled from 25 kHz to 8 kHz [1], [41]. Two real-world nonstationary noise sources were used from the AURORA database, namely, babble and street noise [42]. Four real-world colored noise sources were used. Factory, white and pink noise recordings were taken from the NOISEX-92 noise database [43], and car noise was taken from the AURORA database [42]. Noise was introduced to the clean speech files at −5 dB, 0 dB, 5 dB and 10 dB overall SNR.

B. TEST METRICS

Objective measures were used to evaluate the performance of the speech enhancement methods. The measures were used to evaluate the enhanced speech quality and intelligibility with respect to the corresponding clean speech. The objective measures that were used included:

- The frequency-weighted segmental signal-to-noise ratio (fwSegSNR) [44] was used for objective quality and intelligibility evaluation.
- The log-likelihood ratio (LLR) [45] was used for objective quality evaluation.
- The perceptual evaluation of speech quality (PESQ) [46], [47] was used for objective quality and intelligibility evaluation.
- The normalized-covariance measure (NCM) [48] was used for objective intelligibility evaluation.
- Three objective composite measures were also used to evaluate the perception of speech [23]. These measures evaluate the predicted rating of signal distortion (SIG), the predicted rating of background noise distortion (BAK), and the predicted rating of overall quality (OVL). Each of the three composite measures uses a five-point scale (1-5).

The fwSegSNR, LLR, and PESQ were chosen for their high correlation with the overall quality of the signal and signal distortion [23]. The fwSegSNR, NCM, and PESQ were chosen for their good performance in predicting speech intelligibility [48]. fwSegSNR and LLR were mainly included since they were used as optimization targets in the proposed speech enhancement methods. The composite measures SIG, BAK, and OVL were used to better observe individual aspects of the enhanced speech.

Average objective scores were obtained over the tested speech files under the same noise conditions. Apart from the LLR, which is best minimized, all objective measures were to be maximized for the best results. To further evaluate the speech enhancement performance, a visual representation of the enhanced speech spectrogram was compared to the original and clean spectrograms.
TABLE 4. Comparison of evolutionary approaches to tuning optimization in a self-adaptive framework, averaged over 30 sentences.

| Babbled Noise | Objective measure | MBSS-GA | MBSS-DE |
|---------------|-------------------|---------|---------|
| Input SNR (dB) |                   |         |         |
| 0             | PESQ              | 1.28    | 1.37    | 1.27    |
| 0             | LLR               | 1.02    | 1.02    | 1.19    |

C. SPEECH ENHANCEMENT BASELINES
The single-channel speech enhancement methods used for comparison with the proposed self-adaptive algorithms were the MBSS [21], SS [12], and GS [22]. Three recent single-channel deep neural network approaches to speech enhancement were also used. These methods are the Deep-MMSE method [33], the DCCRN method [37], and a speech enhancement application NVIDIA RTX-Voice Beta [49]. The NVIDIA application leverages NVIDIA RTX GPUs and their AI capabilities for removing background noise from various audio application recordings. Technical specifications for the application were scarce and not readily found, only indicating the use of artificial intelligence. The NVIDIA application for speech enhancement was in beta format at the time of use and may have been subject to change since the tests were performed. For the NVIDIA method, the signal alignment could not be consistently optimal for the measurements on the enhanced speech; as such, the scores shown may be lower than those in an ideal usage.

D. METHOD CONFIGURATION
For the proposed method, a frame size of 25 milliseconds, a 50% overlap and a Hann window were used.

D. METHOD CONFIGURATION

For the proposed method, a frame size of 25 milliseconds, a 50% overlap and a Hann window were used.

Table 4 shows different EOAs that were tested with the proposed framework for speech enhancement tuning optimization. The GA performed better than PSO and DE in our preliminary testing. It was also noted that the GA was on average slower than the PSO and DE methods. This was thought to be a result of PSO and DE falling into local optima more often, reducing execution time at the cost of performance. Given that these tests were performed using default configurations, GA was kept for further tests since reducing execution time was thought to be easier than improving performance.

Table 5 shows the GA parameters that were determined best through preliminary tests for use with the proposed framework. Listed are the function tolerance, the maximum number of iterations, the objective limit, and the other configuration parameters used for GA.

Two self-adaptive techniques shown are herein referred to as MBSS-GA and GS-GA.

E. SYSTEM USED
To measure the runtime, tests were performed using a laptop computer with the following specifications:

- An Intel Core i7-8750H processor (2.2GHz);
- 16 GB of RAM;
- An NVIDIA GeForce GTX 1050Ti graphics card. The self-adaptive method was coded in Matlab.

IX. RESULTS AND DISCUSSION

A. INDEPENDENT VARIABLES

The speech enhancement performance of the self-adaptive framework (MBSS-GA) for varying constraint bounds of $\alpha_i$ and $\beta_i$ is shown in Fig. 14. Fig. 15 shows the frequency distribution of both $\alpha_i$ and $\beta_i$ for these performances. The highest performances were obtained from tests 3 and 4. These tests allowed for a larger range of $\alpha_i$, while maintaining the literature recommended bounds for $\beta_i$. From the frequency distribution, it was deduced that this increase in performance was from the increased frequency of higher $\alpha_i$ values. It can also be seen by comparing tests 4 and 8 that increasing the allowed $\beta_i$ range drastically reduced the performance. This performance reduction can be attributed to the increased frequency of higher $\beta_i$ values.

Increasing the tuning parameter range resulted in an increase in speech enhancement performance. This was interpreted as an advantage of using an evolutionary optimization approach. By comparison, increasing the range for a linear estimation would likely reduce the performance, showing a lack of nuance. It is thought that the metaheuristic nature of evolutionary algorithms permits better adaptivity than a linear estimator.

B. OBJECTIVE FUNCTIONS

Tests were made comparing intelligibility-focused measure NCM, quality-focused measure LLR, intelligibility-and quality-focused measure fwSegSNR, a multiobjective configuration using both fwSegSNR and LLR, and a single-ended measure, namely, the speech-to-reverberation modulation energy ratio (SRMR), which is a nonintrusive metric for speech quality and intelligibility [50], [51]. The multiobjective configuration was achieved using the median of the LLR values, chosen from the Pareto set.

Spectrograms for the self-adaptive framework’s (MBSS-GA) performance when optimizing different objective measures are shown in Fig. 16. The spectrograms showed that all objective measures, although of low complexity, successfully reduced noise. The LLR, fwSegSNR, and multiobjective

TABLE 5. Genetic algorithm tuning parameters used in MBSS-GA experimental implementation.

| GA parameters | Implementation |
|---------------|----------------|
| initial population number | 50 |
| initial population assignment | uniform |
| crossover fraction | 0.8 |
| selection function | uniformly stochastic |
| mutation function | adaptive feasible |
| max. stall generations | 50 |
| max. generations | 60 |
| function tolerance | 0.1 (variable for metric) |
| objective limit | -150 |
| population size | 20 |
configuration yielded similar noise reduction; the SRMR and NCM, however, were visibly distinct. The SRMR had visibly more noise, and the NCM seemed better at retaining speech characteristics in certain time intervals (as seen between 1.25 and 2 seconds). NCM, however, retained more noise in other areas (as seen after 2.5 seconds and before 0.25 seconds). This is logical since NCM is intelligibility-focused, and residual noise is more consequential for quality measures.

Objective measures showed that the multiobjective configuration using both fwSegSNR and LLR yielded the best results. The multiobjective configuration performed better than both objective measures used individually, indicating some complementarity between the two. The visual similarity between both individual LLR and fwSegSNR performances could help to explain the complementarity between the two.

It is thought that the combination of good estimators for speech quality and speech intelligibility resulted in better performance.

The double-ended measures reduced significantly more noise than the blind version, using a single-ended measure (SRMR). This is logical considering the increased precision for objective functions having access to a clean reference signal. The use of the single-ended measure proved to achieve noise reduction, despite not having access to a reference signal, indicating promise for adaptive single-ended speech enhancement. In future work, the use of single-ended measures that are better estimators of speech quality and intelligibility could be used to achieve better results. In addition, the use of speech characteristics as the objective function, as seen in [10], could be researched for embedded applications of speech enhancement.
FIGURE 16. Self-adaptive speech enhancement framework performance for speech corrupted with babble noise at 0 dB SNR, a comparison for various objective measures averaged over 13 sentences.

C. EVOLUTIONARY APPROACH

Comparisons of speech enhancement performance for MBSS-GA, GS-GA, MBSS [21], GS [22], and SS [12] are shown in Table 6 and Table 7. The MBSS-GA performed better than or equal to the traditionally tuned MBSS in 11 out of 21 cases for babble noise and in 4 out of 21 cases for street noise. The MBSS-GA consistently outperformed the MBSS for fwsSegSNR and performed comparably for the LLR, PESQ, and NCM, surpassing it for 6 out of 9 cases for babble noise. The blind implementation of MBSS-GA, using single-ended SRMR, gave a lower performance than MBSS in general. It did, however, perform comparably well for the PESQ and NCM for both noise types, as well as the fwsSegSNR for the babble noise. The GS-GA performed better than or equal to the traditionally tuned GS in 8 out of 21 cases for babble noise and in 10 out of 21 cases for street noise. The GS-GA outperformed GS for the optimized predictors LLR and fwsSegSNR and performed comparably well for the other objective measures in all noise types and levels.

For both self-adaptive framework implementations (MBSS-GA and GS-GA), a general performance increase was shown for the optimized objective measures when compared to their baseline methods (MBSS and GS). The framework implementations did not, however, result in increases for all objective measures. From these results, it was seen that it is possible to optimize certain speech characteristics over others. This ability to selectively optimize speech characteristics could prove useful for targeting specific postspeech-enhancement applications that prefer a certain speech characteristic.

The performance increase from the traditionally tuned methods was higher with MBSS-GA for babble noise and was higher with GS-GA for street noise. Despite inherent algorithm differences, this is thought to be associated with the considerably larger range for $\mu$ than that of $\alpha$. Thus, a lower variation range was beneficial for babble noise, whereas a higher range was beneficial for street noise. Increasing the range, however, must be done cautiously, as it increases the likelihood of falling into local optima. In future work, stationary noise will be studied, and noise type detection will be evaluated to adapt the optimization range for increased adaptivity.

The MBSS-GA outperformed the GS-GA in 20 out of 21 cases for babble noise and in 13 out of 21 cases for street noise, despite GS-GA showing better improvement over traditional tuning. This shows that the performance of the proposed framework is tied to the capabilities of the implemented speech enhancement algorithm. Given that MBSS-GA outperformed GS-GA almost consistently, only MBSS-GA was retained for further comparisons.

An evaluation of the execution time for the self-adaptive method using different objective functions is shown in Table 8. Following tests, the GA runtime was improved by reducing the max. stall generations to 1, this configuration was used for this evaluation. It was seen that the objective function had a considerable impact on the execution time, where the SNR and the segSNR were the fastest and the SRMR was the slowest. It is noteworthy, however, that both the SNR and segSNR implementations were optimized (i.e., reducing unnecessary processes in their coding), whereas the fwsSegSNR and SRMR were not optimized. It was also seen that the improvement of quality can come at a compromise of a higher execution time, as was seen with the fwsSegSNR implementation’s performance. It was also noted that the fastest implementations of the self-adaptive method (SNR and SegSNR) were consistently faster than the average recording duration tested (at 2.69 seconds). These results are promising for possible real-time implementations of this method.

D. EVOLUTIONARY VS. NEURAL APPROACH

Comparisons of speech enhancement performance for Deep-MMSE [33], NVIDIA RTX-voice [49], MBSS-GA, and MBSS-GA(blind) are shown in Table 9 and Table 10. The
DeepMMSE method outperformed the other methods in 17 out of 21 cases for babble noise and in 18 out of 21 cases for street noise. In comparison to DeepMMSE, MBSS-GA performed the best for fwSegSNR and LLR for babble noise at 0 dB SNR. In comparison to NVIDIA’s method, MBSS-GA outperformed 9 out of 21 cases for babble noise and 10 out of 21 cases for street noise.
TABLE 8. Experimental results for proposed self-adaptive framework implementation MBSS-GA for different optimized metrics, compared to MBSS, averaged over 13 sentences.

(a) Babble noise

| Input SNR (dB) | Method  | Optimized metric | Exe. time (s) | PESQ  | NCM |
|---------------|---------|------------------|---------------|-------|-----|
| 0             | MBSS    | n/a              | 0.03          | 1.47  | 0.57|
| 0             | MBSS-GA | SNR              | 1.74          | 1.46  | 0.60|
| 0             | MBSS-GA | SegSNR           | 1.72          | 1.42  | 0.52|
| 0             | MBSS-GA |fwSegSNR          | 11.55         | 1.50  | 0.57|
| 0             | MBSS-GA | SRMR             | 13.28         | 1.39  | 0.46|
| 10            | MBSS    | n/a              | 0.04          | 2.08  | 0.86|
| 10            | MBSS-GA | SNR              | 1.59          | 2.03  | 0.86|
| 10            | MBSS-GA |SegSNR            | 1.80          | 1.91  | 0.82|
| 10            | MBSS-GA |fwSegSNR          | 11.21         | 2.13  | 0.85|
| 10            | MBSS-GA |SRMR              | 13.08         | 1.86  | 0.81|

(b) Pink noise

| Input SNR (dB) | Method  | Optimized metric | Exe. time (s) | PESQ  | NCM |
|---------------|---------|------------------|---------------|-------|-----|
| 0             | MBSS    | n/a              | 0.05          | 1.48  | 0.66|
| 0             | MBSS-GA | SNR              | 1.68          | 1.52  | 0.67|
| 0             | MBSS-GA |SegSNR            | 1.62          | 1.42  | 0.64|
| 0             | MBSS-GA |fwSegSNR          | 11.81         | 1.60  | 0.66|
| 0             | MBSS-GA |SRMR              | 13.04         | 1.45  | 0.48|
| 10            | MBSS    | n/a              | 0.03          | 2.34  | 0.94|
| 10            | MBSS-GA | SNR              | 1.61          | 2.23  | 0.93|
| 10            | MBSS-GA |SegSNR            | 1.72          | 2.01  | 0.90|
| 10            | MBSS-GA |fwSegSNR          | 10.87         | 2.44  | 0.92|
| 10            | MBSS-GA |SRMR              | 12.85         | 2.01  | 0.87|

TABLE 9. Experimental results for proposed self-adaptive framework implementation MBSS-GA, compared to state of the are neural network-based speech enhancement methods, averaged over 13 sentences with babble noise.

| Babble Noise | Objective measure | No process | DeepMMSE | NVIDIA RTX-Voice | MBSS-GA | MBSS-GA (blind) |
|--------------|-------------------|------------|----------|------------------|---------|----------------|
| 0            | fwSegSNR          | 3.36       | 4.52     | 3.79             | 5.22    | 3.47           |
| 5            | fwSegSNR          | 4.29       | 5.77     | 5.02             | 6.10    | 4.56           |
| 10           | fwSegSNR          | 5.44       | 6.34     | 5.34             | 7.02    | 5.60           |
| 0            | LLR               | 1.33       | 1.22     | 1.64             | 1.15    | 1.47           |
| 5            | LLR               | 1.20       | 1.06     | 1.38             | 1.16    | 1.47           |
| 10           | LLR               | 1.06       | 0.96     | 1.16             | 1.17    | 1.48           |
| 0            | PESQ              | 1.36       | 1.62     | 1.61             | 1.49    | 1.43           |
| 5            | PESQ              | 1.52       | 2.22     | 1.94             | 1.76    | 1.64           |
| 10           | PESQ              | 1.78       | 2.66     | 2.29             | 2.11    | 1.92           |
| 0            | NCM               | 0.49       | 0.64     | 0.58             | 0.53    | 0.51           |
| 5            | NCM               | 0.67       | 0.81     | 0.77             | 0.70    | 0.67           |
| 10           | NCM               | 0.80       | 0.88     | 0.80             | 0.80    | 0.79           |
| 0            | SIG               | 1.90       | 2.24     | 1.27             | 2.01    | 1.47           |
| 5            | SIG               | 2.29       | 2.89     | 2.16             | 2.27    | 1.70           |
| 10           | SIG               | 2.73       | 3.23     | 2.74             | 2.92    | 1.91           |
| 0            | BAK               | 1.46       | 1.92     | 1.71             | 1.65    | 1.50           |
| 5            | BAK               | 1.76       | 2.33     | 2.09             | 1.98    | 1.81           |
| 10           | BAK               | 2.06       | 2.54     | 2.31             | 2.28    | 2.09           |
| 0            | OVL               | 1.60       | 1.96     | 1.37             | 1.71    | 1.37           |
| 5            | OVL               | 1.95       | 2.61     | 2.11             | 2.04    | 1.65           |
| 10           | OVL               | 2.35       | 2.96     | 2.57             | 2.56    | 1.94           |

21 cases for street noise, outperforming fwSegSNR and LLR almost consistently.

Despite using the less sophisticated MBSS algorithm, the proposed evolutionary approach performed comparably well to the state-of-the-art DNN methods for the chosen optimized speech metrics, fwSegSNR and LLR. In addition, the evolutionary method performed comparably the best at 0 dB SNR, indicating that low SNR conditions are less optimal for the neural network methods. This shows that a training-independent method has the potential to perform comparably well with training-dependent methods in low SNR conditions. The occasional better performance of the evolutionary approach shows that there is room for collaboration.

It was noted that the speech corpus used for these tests was very popular and may have been used in the training of the neural network-based methods. This could have given an advantage to these methods. In future work, testing with a newly generated speech corpus would be of interest to evaluate the effectiveness of the proposed self-adaptive framework.

E. NEUROEVOLUTIONARY APPROACH

A comparison of speech enhancement performance for the tested configurations of the two-stage approach, as well as DeepMMSE and MBSS-GA, is shown in Table 11. The algorithms were only tested in the best performing conditions for MBSS-GA identified in the previous
TABLE 10. Experimental results for proposed self-adaptive framework implementation MBSS-GA, compared to state of the are neural network-based speech enhancement methods, averaged over 13 sentences with street noise.

| Input SNR (dB) | Objective measure | No process MBSS-GA | NVIDIA RTX-Voice | MBSS-GA | MBSS-GA (blind) |
|---------------|-------------------|--------------------|-------------------|---------|-----------------|
| 0             | fwSegSNR          | 2.93               | 5.54              | 4.39    | 5.24            | 3.07 |
| 5             |                   | 3.90               | 6.12              | 5.38    | 6.41            | 4.09 |
| 10            |                   | 4.95               | 6.66              | 5.94    | 7.41            | 5.21 |
| 0             | LLR               | 1.34               | 1.01              | 1.43    | 1.11            | 1.47 |
| 5             |                   | 1.20               | 0.91              | 1.24    | 1.08            | 1.45 |
| 10            |                   | 1.07               | 0.82              | 1.07    | 1.14            | 1.44 |
| 0             | PESQ              | 1.43               | 1.89              | 1.76    | 1.61            | 1.49 |
| 5             |                   | 1.63               | 2.38              | 2.10    | 1.92            | 1.71 |
| 10            |                   | 1.91               | 2.93              | 2.54    | 2.28            | 2.05 |
| 0             | NCM               | 0.65               | 0.75              | 0.78    | 0.64            | 0.63 |
| 5             |                   | 0.78               | 0.85              | 0.78    | 0.77            | 0.76 |
| 10            |                   | 0.87               | 0.90              | 0.79    | 0.84            | 0.84 |
| 0             | SIG               | 2.02               | 2.80              | 1.81    | 2.27            | 1.62 |
| 5             |                   | 2.47               | 3.21              | 2.49    | 2.57            | 1.89 |
| 10            |                   | 2.88               | 3.56              | 2.99    | 2.71            | 2.15 |
| 0             | BAK               | 1.57               | 2.21              | 1.88    | 1.84            | 1.61 |
| 5             |                   | 1.90               | 2.49              | 2.20    | 2.14            | 1.93 |
| 10            |                   | 2.19               | 2.73              | 2.41    | 2.41            | 2.23 |
| 0             | OVL               | 1.74               | 2.43              | 1.82    | 1.95            | 1.51 |
| 5             |                   | 2.14               | 2.86              | 2.37    | 2.29            | 1.84 |
| 10            |                   | 2.52               | 3.24              | 2.78    | 2.53            | 2.16 |

TABLE 11. Experimental results for novel two-stage hybrid of MBSS-GA and DeepXi speech enhancement methods, averaged over 13 sentences with babble noise.

| Input SNR (dB) | Objective measure | No process MBSS-GA | DeepMMSE | two-stage baseline config. 1 MBSS-GA + DeepMMSE) | two-stage hybrid config. 1 MBSS-GA + DeepMMSE) | two-stage hybrid config. 2 MBSS-GA + DeepMMSE) |
|---------------|-------------------|--------------------|----------|-------------------------------------------------|------------------------------------------------|------------------------------------------------|
| 0             | fwSegSNR          | 3.36               | 4.52     | 5.22                                           | 7.03                                           | 5.47                                           | 7.42                                           |
| 0             | LLR               | 1.33               | 1.22     | 1.15                                           | 0.87                                           | 1.22                                           | 0.79                                           |
| 0             | PESQ              | 1.36               | 1.62     | 1.49                                           | 1.59                                           | 1.42                                           | 1.73                                           |
| 0             | NCM               | 0.49               | 0.64     | 0.53                                           | 0.70                                           | 0.65                                           | 0.71                                           |
| 0             | SIG               | 1.90               | 2.24     | 2.01                                           | 2.91                                           | 2.02                                           | 3.10                                           |
| 0             | BAK               | 1.46               | 1.92     | 1.65                                           | 2.39                                           | 2.07                                           | 2.38                                           |
| 0             | OVL               | 1.60               | 1.96     | 1.71                                           | 2.36                                           | 1.75                                           | 2.53                                           |

subsections, namely, for babble noise at 0 dB SNR. The best observed performance was with a first neural method stage, followed by an evolutionary-based stage (i.e., configuration 2). This configuration performed significantly better than the baseline of two subsequent neural stages in all measures except BAK and NCM. The proposed two-stage method outperformed the state-of-the-art DeepMMSE method by an average 32% improvement over all tested objective speech measures. The baseline configuration, however, only outperformed the DeepMMSE method by an average 25% improvement. The second configuration outperformed the baseline configuration by an average of 5% improvement.

The configuration with the evolutionary method as a first stage performed significantly worse than the configuration with a neural stage, followed by an evolutionary method as a second stage. This was expected since the DeepMMSE method would not have been trained for speech processed by the evolutionary approach. The increase in
TABLE 12. Experimental results for novel two-stage hybrid of a first stage composed of either DCCRN or DeepMMSE speech enhancement methods, with a second stage composed of the MBSS-GA method, averaged over 13 sentences with babble noise (a), street noise (b), white noise (c), pink noise (d), car noise (e), and factory noise (f).

| Input SNR (dB) | Method | PESQ | NCM | SIG | BAK | OVL |
|---------------|--------|------|-----|-----|-----|-----|
| -5            | DeepMMSE | 1.30 | 0.45 | 2.17 | 1.81 | 1.68 |
| -5            | -2stage baseline | 1.19 | 0.44 | 1.84 | 1.71 | 1.37 |
| -5            | -2stage -MBSS-GA | 1.29 | 0.47 | 2.11 | 1.82 | 1.68 |
| -5            | DCCRN | 1.39 | 0.56 | 1.86 | 1.48 | 1.62 |
| -5            | -2stage baseline | 1.28 | 0.57 | 1.27 | 1.29 | 1.22 |
| -5            | -2stage -MBSS-GA | 1.42 | 0.59 | 1.96 | 1.85 | 1.69 |

-5 DeepMMSE | 0.16 | 0.70 | 3.03 | 2.42 | 2.46 |
-5 -2stage baseline | 0.15 | 0.70 | 2.91 | 2.39 | 2.36 |
-5 -2stage -MBSS-GA | 0.17 | 0.71 | 3.10 | 2.38 | 2.53 |
-5 DCCRN | 1.82 | 0.79 | 2.84 | 2.13 | 2.44 |
-5 -2stage baseline | 1.66 | 0.79 | 2.11 | 1.98 | 1.94 |
-5 -2stage -MBSS-GA | 1.89 | 0.79 | 2.89 | 2.55 | 2.50 |

-5 DeepMMSE | 2.29 | 0.88 | 3.77 | 2.99 | 3.16 |
-5 -2stage baseline | 2.27 | 0.88 | 3.69 | 2.97 | 3.11 |
-5 -2stage -MBSS-GA | 2.23 | 0.87 | 3.69 | 2.74 | 3.09 |
-5 DCCRN | 2.48 | 0.92 | 3.65 | 2.84 | 3.17 |
-5 -2stage baseline | 2.28 | 0.92 | 3.05 | 2.72 | 2.76 |
-5 -2stage -MBSS-GA | 2.42 | 0.91 | 3.63 | 2.92 | 3.13 |

-5 DeepMMSE | 1.90 | 0.78 | 3.12 | 2.67 | 2.66 |
-5 -2stage baseline | 1.94 | 0.77 | 3.10 | 2.70 | 2.67 |
-5 -2stage -MBSS-GA | 2.01 | 0.78 | 3.18 | 2.60 | 2.74 |
-5 DCCRN | 1.37 | 0.64 | 1.83 | 1.68 | 1.61 |
-5 -2stage baseline | 1.26 | 0.66 | 1.25 | 1.36 | 1.17 |
-5 -2stage -MBSS-GA | 1.62 | 0.68 | 2.45 | 2.34 | 2.14 |

-5 DeepMMSE | 2.33 | 0.87 | 3.56 | 3.04 | 3.08 |
-5 -2stage baseline | 2.38 | 0.87 | 3.53 | 3.06 | 3.08 |
-5 -2stage -MBSS-GA | 2.28 | 0.87 | 3.53 | 2.78 | 3.04 |
-5 DCCRN | 1.76 | 0.82 | 2.68 | 2.23 | 2.35 |
-5 -2stage baseline | 1.62 | 0.82 | 2.02 | 2.00 | 1.89 |
-5 -2stage -MBSS-GA | 2.02 | 0.85 | 2.99 | 2.74 | 2.65 |

-5 DeepMMSE | 2.74 | 0.94 | 3.97 | 3.37 | 3.44 |
-5 -2stage baseline | 2.76 | 0.94 | 3.93 | 3.36 | 3.42 |
-5 -2stage -MBSS-GA | 2.54 | 0.93 | 3.85 | 2.93 | 3.31 |
-5 DCCRN | 2.22 | 0.91 | 3.32 | 2.75 | 2.90 |
-5 -2stage baseline | 2.11 | 0.92 | 2.68 | 2.62 | 2.50 |
-5 -2stage -MBSS-GA | 2.44 | 0.92 | 3.55 | 2.96 | 3.11 |

-5 DeepMMSE | 1.50 | 0.65 | 2.72 | 2.13 | 2.18 |
-5 -2stage baseline | 1.38 | 0.65 | 2.35 | 2.00 | 1.85 |
-5 -2stage -MBSS-GA | 1.56 | 0.66 | 2.76 | 2.11 | 2.22 |
-5 DCCRN | 1.58 | 0.68 | 2.50 | 1.70 | 2.09 |
-5 -2stage baseline | 1.44 | 0.69 | 1.79 | 1.54 | 1.59 |
-5 -2stage -MBSS-GA | 1.55 | 0.71 | 2.53 | 2.04 | 2.09 |

-5 DeepMMSE | 1.93 | 0.81 | 3.46 | 2.62 | 2.84 |
-5 -2stage baseline | 1.85 | 0.81 | 3.23 | 2.55 | 2.67 |
-5 -2stage -MBSS-GA | 1.96 | 0.81 | 3.43 | 2.50 | 2.83 |
-5 DCCRN | 2.20 | 0.87 | 3.38 | 2.38 | 2.91 |
-5 -2stage baseline | 2.02 | 0.87 | 2.75 | 2.29 | 2.49 |
-5 -2stage -MBSS-GA | 2.12 | 0.87 | 3.31 | 2.72 | 2.64 |

-5 DeepMMSE | 2.43 | 0.92 | 4.00 | 3.06 | 3.34 |
-5 -2stage baseline | 2.40 | 0.92 | 3.82 | 3.00 | 3.23 |
-5 -2stage -MBSS-GA | 2.36 | 0.91 | 3.87 | 2.80 | 3.25 |
-5 DCCRN | 2.80 | 0.95 | 3.94 | 2.96 | 3.43 |
-5 -2stage baseline | 2.57 | 0.95 | 3.42 | 2.86 | 3.07 |
-5 -2stage -MBSS-GA | 2.54 | 0.94 | 3.81 | 2.98 | 3.28 |

performance was attributed to the training independence of the evolutionary approach and its capability of adapting to its input. The results showed that a training-independent evolutionary method complemented a training-dependent deep learning method for speech enhancement when used as a postprocess. These experiments demonstrated the advantages of building a cooperative approach using evolutionary and deep learning-based techniques that are capable of achieving robust speech enhancement in adverse conditions. This collaborative approach shows promise as an adaptive approach for DNN methods of speech enhancement.

The success of the baseline configuration also shows the validity of a multipass approach for DNN methods to increase...
TABLE 13. Experimental results for novel two-stage hybrid of a first stage composed of either DCCRN or DeepMMSE speech enhancement methods, with a second stage composed of the MBSS-GA method, averaged over 13 sentences with car noise (a), and factory noise (b).

(a) Car noise

| Input SNR (dB) | Method                | PESQ | NCM   | SIG  | BAK  | OVL  |
|---------------|-----------------------|------|-------|------|------|------|
| -5            | DeepMMSE             | 1.93 | 0.86  | 3.59 | 2.56 | 2.90 |
| -5            | -2stage baseline     | 1.86 | 0.85  | 3.35 | 2.54 | 2.74 |
| -5            | -2stage -MBSS-GA     | 1.94 | 0.86  | 3.51 | 2.48 | 2.86 |
| -5            | DCCRN                | 2.02 | 0.88  | 3.33 | 2.07 | 2.79 |
| -5            | -2stage baseline     | 1.83 | 0.88  | 2.81 | 2.02 | 2.43 |
| -5            | -2stage -MBSS-GA     | 2.00 | 0.88  | 3.34 | 2.44 | 2.79 |
| 0             | DeepMMSE             | 2.34 | 0.93  | 4.01 | 2.90 | 3.31 |
| 0             | -2stage baseline     | 2.33 | 0.93  | 3.83 | 2.88 | 3.21 |
| 0             | -2stage -MBSS-GA     | 2.25 | 0.92  | 3.83 | 2.71 | 3.18 |
| 0             | DCCRN                | 2.58 | 0.95  | 3.90 | 2.65 | 3.34 |
| 0             | -2stage baseline     | 2.40 | 0.95  | 3.45 | 2.59 | 3.03 |
| 0             | -2stage -MBSS-GA     | 2.44 | 0.95  | 3.81 | 2.94 | 3.24 |
| 5             | DeepMMSE             | 2.85 | 0.97  | 4.42 | 3.26 | 3.71 |
| 5             | -2stage baseline     | 2.85 | 0.97  | 4.27 | 3.24 | 3.63 |
| 5             | -2stage -MBSS-GA     | 2.65 | 0.97  | 4.21 | 2.94 | 3.55 |
| 5             | DCCRN                | 3.17 | 0.99  | 4.32 | 3.20 | 3.77 |
| 5             | -2stage baseline     | 2.97 | 0.99  | 3.88 | 3.10 | 3.46 |
| 5             | -2stage -MBSS-GA     | 2.82 | 0.98  | 4.17 | 3.13 | 3.57 |

(b) Factory noise

| Input SNR (dB) | Method                | PESQ | NCM   | SIG  | BAK  | OVL  |
|---------------|-----------------------|------|-------|------|------|------|
| -5            | DeepMMSE             | 1.41 | 0.54  | 2.47 | 1.94 | 1.95 |
| -5            | -2stage baseline     | 1.33 | 0.52  | 2.29 | 1.90 | 1.76 |
| -5            | -2stage -MBSS-GA     | 1.39 | 0.55  | 2.54 | 1.99 | 1.96 |
| -5            | DCCRN                | 1.34 | 0.59  | 1.73 | 1.33 | 1.48 |
| -5            | -2stage baseline     | 1.25 | 0.59  | 1.10 | 1.11 | 1.09 |
| -5            | -2stage -MBSS-GA     | 1.38 | 0.61  | 2.02 | 1.77 | 1.68 |
| 0             | DeepMMSE             | 1.84 | 0.76  | 3.29 | 2.51 | 2.71 |
| 0             | -2stage baseline     | 1.74 | 0.73  | 3.14 | 2.46 | 2.56 |
| 0             | -2stage -MBSS-GA     | 1.84 | 0.76  | 3.27 | 2.46 | 2.69 |
| 0             | DCCRN                | 1.85 | 0.83  | 2.87 | 2.03 | 2.44 |
| 0             | -2stage baseline     | 1.73 | 0.83  | 2.23 | 1.86 | 2.01 |
| 0             | -2stage -MBSS-GA     | 1.91 | 0.83  | 3.07 | 2.47 | 2.59 |
| 5             | DeepMMSE             | 2.42 | 0.91  | 3.89 | 3.00 | 3.28 |
| 5             | -2stage baseline     | 2.39 | 0.89  | 3.81 | 2.99 | 3.22 |
| 5             | -2stage -MBSS-GA     | 2.38 | 0.90  | 3.83 | 2.81 | 3.23 |
| 5             | DCCRN                | 2.51 | 0.95  | 3.67 | 2.69 | 3.18 |
| 5             | -2stage baseline     | 2.58 | 0.95  | 3.06 | 2.58 | 2.79 |
| 5             | -2stage -MBSS-GA     | 2.46 | 0.94  | 3.72 | 2.90 | 3.20 |

performance. These results, however, could vary for different noise conditions.

The NCM varied the least between the tested two-stage configurations. This indicates that the proposed approach is better for increasing quality, rather than intelligibility. This lack of intelligibility increase, however, is a common limitation seen in single-channel speech enhancement. Intelligibility was not greatly affected by these methods. In future work, auditory tests would need to be performed to gain better insight into the impact of the proposed approaches.

A comparison of speech enhancement performance for the tested configurations of the two-stage approach, as well as baselines from the DeepMMSE and DCCRN, is shown in Table 12 and Table 13. The two-stage neuroevolutionary configuration improved performances from their standalone counterparts (DCCRN and DeepMMSE) mainly for noise conditions at −5 dB, at 0 dB and on a few occasions at 5 dB. For tests with DeepMMSE, the neuroevolutionary configuration outperformed or performed equally well to DeepMMSE at a −5 dB input SNR for all noise types except car noise and at a 0 dB input SNR for pink noise, street noise, babble noise, and factory noise. For tests with DCCRN, the neuroevolutionary configuration outperformed or performed equally well to DCCRN at −5 and 0 dB input SNR for all six tested noise conditions. In addition, tests for white noise, pink noise, and factory noise were successful at a 5 dB input SNR. In general, at a −5 and a 0 dB input SNR, the neuroevolutionary configuration improved the DCCRN performance by an average of 13% and 6%, respectively, over the six tested noise conditions.

The neuroevolutionary configuration increased the performance of DCCRN more frequently than DeepMMSE. This is attributed to the DeepMMSE training set having a larger variety of noise conditions, which included colored noise recordings. DCCRN, which did not have colored noise recordings in its training set, saw a performance increase when used in the neuroevolutionary configuration when used for these conditions. This is made more evident when seeing that the neuroevolutionary configuration failed to significantly improve the performance of DCCRN for noise conditions similar to its training material, such as street noise and car noise. In addition, the neuroevolutionary configuration significantly increased performance for DCCRN at −5 dB input SNR, which were unknown noise conditions. This shows that neuroevolutionary collaboration can also be effective for noise conditions that differ from those used in the training material of the DNN method.

The two-stage baseline configuration with two consecutive DNN passes performed generally poorer than the standalone methods themselves (i.e., DeepMMSE and DCCRN), with very few exceptions. It was also noted that the NCM varied little in any of the tested conditions for the different configurations; thus, the two-stage configurations had little effect on the intelligibility.

The results showed that a training-independent evolutionary method complemented a training-dependent deep learning method for speech enhancement when used as a
postprocess. These experiments demonstrated the advantages of building a cooperative approach using evolutionary and deep learning-based techniques that are capable of achieving robust speech enhancement in adverse conditions. This collaborative approach shows promise as an adaptive approach for DNN methods of speech enhancement. The tests effectively showed that a collaborative approach to speech enhancement can be used to improve the performance for difficult conditions (i.e., at low SNR, for unknown noise conditions or for highly nonstationary noises). The increase in performance was attributed to the training independence of the evolutionary approach and its capability of adapting to its input.

X. CONCLUSION
This work investigated a two-stage deep neuroevolutionary approach to speech enhancement. The approach coupled an existing deep learning method and an evolutionary self-adaptive approach for speech enhancement. The self-adaptive tuning approach was achieved using evolutionary optimization to adaptively tune the parameters of selected speech enhancement algorithms based on chosen objective speech metrics. It was found that this training-independent evolutionary method complemented a deep learning method for speech enhancement when used as a postprocess in difficult noise conditions.

The proposed two-stage method outperformed state-of-the-art methods for speech enhancement and thus demonstrated the effectiveness of cooperative approaches to deal with unknown noise conditions, highly nonstationary adverse conditions, and thus unpredictable environments. The proposed neuroevolutionary framework, by using a self-tuning optimization of key parameters of speech enhancement, succeeded in dealing with the unpredictability and versatility of noisy speech.

In future work, further optimization of the proposed self-adaptive framework could see lower execution time and be tested for real-time embedded applications. The compatibility of single-ended objective measures in the framework for specific postprocess applications such as speech recognition would also be explored. Further testing would be done to evaluate the proposed approaches for more noise conditions (e.g., stationary noise) and more speech segments from different corpora (e.g., a larger number of speakers). In addition, more recent and cutting edge EOAs would be explored to maximize the potential benefit of the proposed self-adaptive tuning control framework, such as those shown in [17], [18]. An approach using signal characteristics such as the one used in [10] instead of objective measures could also be evaluated for real-time applications or embedded applications.

REFERENCES
[1] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” Speech Commun., vol. 49, nos. 7–8, pp. 588–601, Jul. 2007.
[2] P. C. Loizou, Speech Enhancement: Theory Practice. Boca Raton, FL, USA: CRC Press, 2013.
[3] M. A. B. Messaoud, A. Bouzid, and N. Ellouze, “Speech enhancement based on wavelet packet of an improved principal component analysis,” Comput. Speech Lang., vol. 35, pp. 58–72, May 2016.
[4] Y. Xian, Y. Sun, W. Wang, and S. M. Naqvi, “Multi-scale residual convolutional encoder decoder with bidirectional long short-term memory for single channel speech enhancement,” in Proc. 28th Eur. Signal Process. Conf. (EUSIPCO), Jan. 2021, pp. 431–435.
[5] P. P. Ingale and S. L. Naibalwar, “Deep neural network based speech enhancement using mono channel mask,” Int. J. Speech Technol., vol. 22, no. 3, pp. 841–850, Sep. 2019, doi: 10.1007/s10772-019-09627-4.
[6] R. Liang, F. Kong, Y. Xie, G. Tang, and J. Cheng, “Real-time speech enhancement algorithm based on attention LSTM,” IEEE Access, vol. 8, pp. 48464–48476, 2020.
[7] S. Abdullah, M. Zamanli, and A. Demesthenous, “Towards more efficient DNN-based speech enhancement using quantization correlation measure,” IEEE Access, vol. 9, pp. 24350–24362, 2021.
[8] R. LeBlanc and S. A. Selouani, “Self-adaptive tuning for speech enhancement algorithm based on evolutionary approach,” in Proc. IEEE Int. Conf. Cogn. Mach. Intell. (CogMI), 2019, pp. 16–22.
[9] D. Giacobello, J. Atkins, J. Wung, and R. Prabhu, “Results on automated tuning of a voice quality enhancement system using objective quality measures,” in Proc. 135th Audio Eng. Soc. Conv., 2013.
[10] M. A. Abd El-Fattah, M. I. Dessouky, A. M. Abbas, S. M. Diab, E.-S.-M. El-Rabaie, W. Al-Naaimy, S. A. Alshebeili, and F. E. Abd El-samie, “Speech enhancement with an adaptive Wiener filter,” Int. J. Speech Technol., vol. 17, no. 1, pp. 53–64, Mar. 2014, doi: 10.1007/s10772-013-9195-5.
[11] P. Kunche and K. V. S. Reddy, “Application of meta-heuristics to speech enhancement,” in Metaheuristic Applications to Speech Enhancement (SpringerBriefs in Electrical and Computer Engineering), P. Kunche and K. Reddy, Eds. Cham, Switzerland: Springer, 2016, pp. 25–37.
[12] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Mar. 1979, pp. 206–211.
[13] I. Tashiev, A. Lovitt, and A. Acero, “Unified framework for single channel speech enhancement,” in Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process., Aug. 2009, pp. 883–888.
[14] D. Giacobello, J. Wung, R. Pichevar, and J. Atkins, “Tuning methodology for speech enhancement algorithms using a simulated conversational database and perceptual objective measures,” in Proc. 4th Joint Workshop Hands-Free Speech Commun. Microphone Arrays (HSCMA), May 2014, pp. 62–66.
[15] C.-C. Lee, Y.-C. Lin, H.-T. Lin, H.-M. Wang, and Y. Tsao, “SERIL: Noise adaptive speech enhancement using regularization-based incremental learning,” in Proc. Interspeech, Oct. 2020, pp. 2432–2436.
[16] U. Mahbub, C. Shahnaz, and S. Anowarul Fattah, “An adaptive noise cancellation scheme using particle swarm optimization algorithm,” in Proc. Int. Conf. Commun. Control Comput. Technol., Oct. 2010, pp. 683–686.
[17] K. Prajna, K. V. S. Reddy, G. S. B. Rao, and R. U. Maheswari, “A comparative study of BA, APSO, GSA, hybrid PSOGSA and SPSO in dual channel speech enhancement,” Int. J. Speech Technol., vol. 18, no. 4, pp. 663–671, Dec. 2015, doi: 10.1007/s10772-015-9308-2.
[18] A. Aggarwal, T. K. Rawat, and D. K. Upadhyay, “Design of optimal digital FIR filters using evolutionary and swarm optimization techniques,” in Proc. Int. Conf. Commun., vol. 70, no. 4, pp. 373–385, Apr. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1434841115300133
[19] T. M. F. Taha, S. K. Wajid, and A. Hussain, “Speech enhancement based on adaptive noise cancellation and particle swarm optimization,” J. Comput. Sci., vol. 15, no. 5, pp. 691–701, May 2019. [Online]. Available: https://thesispub.com/abstract/jcssp.2019.691.701
[20] B. Widrow, J. Glover, and J. McCool, “Adaptive noise cancelling: Principles and applications,” Proc. IEEE, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.
[21] S. Kamath and P. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in Proc. ICASSP, vol. 4, 2002, p. 44164.
[22] Y. Hu and P. C. Loizou, “A generalized subspace approach for enhancing speech corrupted by colored noise,” IEEE Trans. Speech Audio Process., vol. 11, no. 4, pp. 334–341, Jul. 2003.
