RAIRO-Oper. Res. 54 (2020) 1555–1572
https://doi.org/10.1051/ro/2019106

RAIRO Operations Research
www.rairo-ro.org

MULTIPLE OBJECTIVE OPTIMIZATION APPLIED TO SPEECH ENHANCEMENT PROBLEM

Said Ouznadji[1], Djamal Chaabane[1,*] and Messaoud Thameri[2]

Abstract. Enhancement of speech corrupted by broadband noise is subject of interest in many applications. For several years, the investigation of methods of denoising the vocal signal has yielded very satisfactory results, but certain problems and questions still remain. The term speech quality in speech enhancement is associated with clarity and intelligibility. So, one of these issues is to reach a compromise between noise reduction, signal distortion and musical noise. In this paper, we studied one of the classical techniques based on the spectral subtraction developed by Boll and improved by Berouti where two parameters α and β to control the effects of the distortion and the musical noise are introduced. However, the choice on these parameters (α and β) remains empirical. Our works is to find a compromise between these two parameters to obtain an optimal solution depending on the environment, the unknown noise and its level. Moreover, we propose in this paper, an algorithm based on bi-objective approach precisely Particle Swarm Optimization (PSO) technique in association with speech enhancement technique proposed by Berouti *et al.* Comparative results show that the performance of our proposed method with several types of noise, depending on the environment and on various noise levels, are better than those of spectral subtraction methods of Boll or Berouti.

Mathematics Subject Classification. 90C05, 90C25, 90C29, 65K05.

Received March 9, 2019. Accepted November 10, 2019.

1. Introduction

Speech processing is a fundamental element of digital signal processing, language processing or symbolic data processing, which has grown considerably as a result of the development of communications and telecommunications resources and technology. The particular importance of speech processing is explained by the privileged position of speech as a vector of information in society, certainly due to the extraordinary and fascinating role of the human brain in the production and understanding of speech, as well as in the set of functions that can be implemented in real time.

Modern speech processing techniques [22,32] tend to produce automated systems capable of highlighting the characteristics (i) of the speech signal as it is produced, or sometimes, as it is perceived, (ii) of decoding or of recognition of the information conveyed by the vocal signal, (iii) of artificial speech production when it comes

Keywords. Speech enhancement, spectral subtraction, multiobjective optimization, meta-heuristic, PSO.

[1] USTHB University, Lab. AMCD&RO, DGRSDT, Faculty of Mathematics, Department of Operations Research, Bab-Ezzouar, BP32 El-Alia, 16122 Algiers, Algeria.

[2] École Supérieure Ali CHABATI Reghaia, Algiers, Algeria.

*Corresponding author: chaabanedjamal96@gmail.com, chaabane_dj@yahoo.fr

Article published by EDP Sciences

to synthesis [4,15], to improve the quality in the means of communications, telecommunications and especially in the military and medical systems (hearing aids: prostheses ...) [17].

Similarly, the environment surrounding the speech of a speaker can generate uncorrelated and additive noises, so they have a superposition effect on the speech signal as is often the case with ambient noise. These situations engendered the development of denoising, echo cancellation and source separation algorithms [24,32].

Noise tends to mask the speech signal, which reduces the quality where the speech is strongly affected by the presence of ambient noise. Restricting or reducing this background noise and improving the quality of perception and intelligibility of a speech without disturbing the quality of the speech signal is very difficult to maintain.

The problem of speech enhancement is still a vast area of study and still rich in ideas. The goal is to restore a useful signal from corrupted observations often considered as additive noise. The classical methods, such as spectral subtraction [5,6,27,29] and Wiener filtering [34], or Kalman filtering [21], manage to reduce the additive noise, but introduce in return a residual noise called musical noise, annoying for the human perception. The need to reduce this type of noise while preserving intelligibility and speech quality has led researchers to propose other solutions to this problem such as noise estimation and detection of vocal activity. These first attempts made it possible to improve the conventional spectral subtraction procedure in order to avoid its undesirable effects, thus improving speech intelligibility [11,33].

But later, with the progress of signal processing, new solutions have been proposed, such as the use of wavelets [30], so-called subspace methods [31], and stochastic methods [8,14].

Various variants of the spectral subtraction have been developed to overcome the disadvantage of degradation of the speech signal induced by additive background noise making the listening task very difficult for the listener. They form a family of subtractive algorithms which deal with: amplitude [25,26], power [13], non-linear Spectral Subtraction [20], perceptual masking [1].

In recent years, operational research technique where widely used in this field that has given rise of several approximated algorithms such evolutionary, genetic, heuristic and metaheuristic applied to the speech signal enhancement problem [18,35], sometimes hybrid [19].

A subset of metaheuristics is often called Swarm Intelligence (SI) based algorithms and they have been developed by mimicking the so-called intelligence characteristics of biological swarm agents such as birds, fish, humans and others. In the work of L. Badriasl *et al.* based on the work introduced in [16], they present the application of particle swarm optimization techniques (PSO), in the technical structures or methods of speech enhancement [2,3].

In the spectral subtraction method, the noise spectrum is estimated during speech pauses and is subtracted from the noisy speech spectrum to estimate clean speech. This is achieved by multiplying the noisy speech spectrum with a gain or transfer function and then combining it with the noisy speech phase. The disadvantage of this method is the presence of distortions and musical noise.

In [5], Berouti *et al.* consider the classical spectral subtraction algorithm developed by Boll *et al.* [6], and proposes a method to reduce the effect of musical noise and to mitigate the effect of distortion, by introducing two parameters of control α and β which are fixed experimentally. Here, as in the case of conventional subtraction, not only is the noise spectrum subtracted from the noisy speech spectrum, but this method also subtracts an overestimation of the noise spectrum so that the output does not outperform the background noise.

Our goal is to study speech signal enhancement algorithms based on spectral subtraction and to develop new algorithms by adopting multi-criteria optimization methods, in order to find the optimal solution to improve the quality of the speech signal corrupted by unwanted noise by also using nature-based resolution methods such as particle swarm optimization.

With regards to the work that we are studying, we will limit ourselves to the subtraction techniques of Berouti *et al.* [5] and Boll [6] by developing a bi-objective algorithm based on particle swarm optimization techniques (PSOs).

From a mathematical point of view, we introduce a new model that takes into account the multi-objective aspect of maximizing the signal-to-noise ratio (SNR) as the primary objective and minimizing signal distortion as the second objective. Operational research techniques are used to solve such problems.

Also, in this paper, we propose a new bi-criteria optimization method for speech denoising using an approximate resolution method where the preliminary idea was introduced in extended abstract [23]. In the present paper, however, there are rich explanations and experiments introduced and developed with several types of noise, at different noise levels to validate our approach.

In Section 2, we present some tools needed to understand the concept of speech signal and the theory of multi-objective optimization. In Section 3, our approach is detailed and followed by a formal description of the algorithm. Some experimental results are given in Section 4. Finally, some conclusions and remarks are made in Section 5.

2. Necessary tools

2.1. Speech signal analysis and treatment

In this section, we give some background concepts concerning speech signal enhancement. The speech signal is a non-stationary random process in the long term, but it is considered stationary in time windows analysis for order between 20 and 30 ms. This short-term stationarity property allows gradual analysis and modeling of the speech signal, accompanied of overlapping windows, in order to assure the temporal continuity of signal characteristics.

The human ear has impressive abilities to recognize and distinguish speech from noise. But for the well being of the listener and in order to limit fatigue, seeking to improve the quality of listening through the noise suppression of speech (for applications such as mobile telephony and telephony Hands free).

The systems dedicated to the latter applications are affected by the quality of speech and performance may be poor in the presence of noise especially when there has been no noisy learning environment. Several methods of removing noise from speech have been proposed in the literature to meet the needs of users and applications dedicated to speech. We describe briefly the spectral subtraction technique that will be considered in our work.

Let $x_k(t)$, $s_k(t)$ and $n_k(t)$, $t = 0, 1, \ldots, N-1$, the noisy signal, the clean signal and the noise in the kth frame Respectively. By noting $X_k(\nu)$, $S_k(\nu)$ and $N_k(\nu)$ the Fourier Transforms (FT) of $x_k(t)$, $s_k(t)$ and $n_k(t)$ respectively. One can obtain:

$$X_k(\nu) = S_k(\nu) - N_k(\nu) \tag{2.1}$$

for a classical linear denoising problem where one looks for an estimator $H_k(\nu)$ such as:

$$\widehat{S}_k(\nu) = H_k(\nu) X_k(\nu) \tag{2.2}$$

where $\widehat{S}_k(\nu)$ denotes the estimate of $S_k(\nu)$. The error due to this filtering is given by:

$$e_k(\nu) = \widehat{S}_k(\nu) - S_k(\nu) \tag{2.3}$$

$$= (H_k(\nu) - 1) S_k(\nu) - H_k(\nu) N_k(\nu) \tag{2.4}$$

where $(H_k(\nu) - 1) S_k(\nu)$ represents the distortion of the signal while $H_k(\nu) N_k(\nu)$ designates the residual noise containing the musical noise. In the case where $0 \leq H_k(\nu) \leq 1$, it is generally very difficult to reduce the musical noise without distorting the signal.

The need for a compromise between distortion and musical noise is therefore the best way to increase performance in terms of quality and intelligibility.

Spectral subtraction is very simple to implement, however, it remains limited by the appearance of musical noise, that is why it continues to be the subject of development of methods that use learning techniques such as Deep Learning and multi-objective optimization techniques [28].

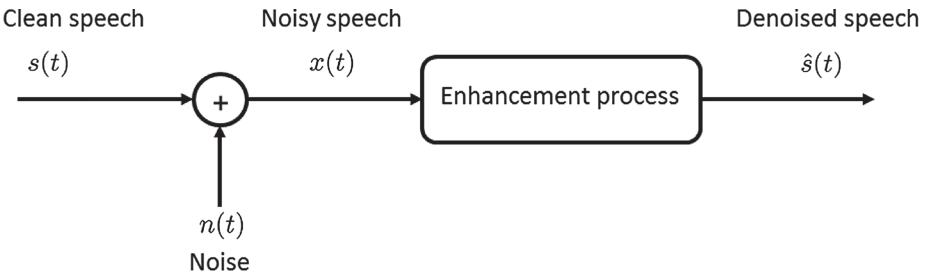

FIGURE 1. Basic diagram of speech enhancement system.

2.2. Spectral subtraction method

Spectral subtraction [5] is one of the popular and conventional methods because of its ease of implementation and its reduced computation in speech processing. It began to be used in 1979 and became an effective algorithm for speech enhancement. The basic spectral subtraction assumes a regular voice signal and the noise is an additive noise. Voice and noise are not related to each other. Figure 1 gives the simple diagram of an enhancement system.

The spectral subtraction method is used to mitigate spectral components of the noise. There are two basic versions of spectral subtraction techniques differing from each other by the use of either the power subtraction or magnitude subtraction. If the magnitude is considered, then this version is referred to as Magnitude Spectral Subtraction (MSS) and the estimated Fourier Transform is given by (where $|\widehat{N}_k(\nu)|$ is the estimated noise spectre)

$$|\widehat{S}_k(\nu)| = |X_k(\nu)| - |\widehat{N}_k(\nu)|. \tag{2.5}$$

If the power is considered then this version is referred to as Power Spectral Subtraction (PSS) and the estimated Fourier Transform is given by

$$|\widehat{S}_k(\nu)|^2 = |X_k(\nu)|^2 - |\widehat{N}_k(\nu)|^2. \tag{2.6}$$

Since the second term of equation (2.6) can be negative, it can be made positive sign or annulated as in equation (2.7). This is part of the first improvements in spectral subtraction:

$$|\widehat{S}_k(\nu)|^2 = \begin{cases} |X_k(\nu)|^2 - |\widehat{N}_k(\nu)|^2 & \text{if} \quad |X_k(\nu)|^2 > |\widehat{N}_k(\nu)|^2 \\ 0 & \text{elsewhere} \end{cases}. \tag{2.7}$$

It is possible to use the silence zones in the frames before the speech to estimate the noise. To find the spectrum of the noise, the first hypothesis consists in saying that it is stationary (Hypothesis which is not always valid) and independent of the frame of analysis. The second hypothesis is that speech varies much faster than noise. On the other hand, it can be said that on short analysis frames (frame duration is less than 30 ms) speech is stationary.

The transition in the time domain is performed by the Inverse Fourier Transform (IFT) keeping the phase ($\arg(X_k(\nu))$) of the noisy signal as shown in equation (2.8). We allow ourselves to do so, firstly, because our ear is not very sensitive to phase changes and, secondly, because an estimate of the phase is a very complicated task.

$$\widehat{s}_k(t) = \text{IFT}[|\widehat{S}_k(\nu)| \cdot \mathrm{e}^{i \times \arg(X_k(\nu))}]. \tag{2.8}$$

The spectral subtraction is widely studied in the literature given its simplicity, but it is limited by the artefacts it generates output, ie the signal distortion and noise music. Attempts to reduce this have led researchers to invest on the proper expression that can be a compromise between the amount of musical noise and signal distortion. Thus, to give more flexibility to the spectral subtraction, several improvements have been made to this technique.

2.2.1. Improvement of classical spectral subtraction proposed by Berouti et al.

Berouti *et al.* found in [6] that after a spectral subtraction, the residual noise contains two types of spectral peaks. The broad peaks perceived as broad band noise and the narrow peaks as tonal. the latter are called musical noise. The authors propose in the spectral subtraction to add to the overestimation of the noise a quantity $\beta\left|\widehat{N}(\nu)\right|^2$ instead of 0 to prevent the noise tolerance threshold from exceeding the noise power as shown in equation (2.9).

$$|\widehat{S}(\nu)|^2 = \begin{cases} |X(\nu)|^2 - \alpha|\widehat{N}(\nu)|^2 & \text{if} \quad |X(\nu)|^2 - \alpha|\hat{N}(\nu)|^2 > \beta|\widehat{N}(\nu)|^2 \\ \beta|\widehat{N}(\nu)|^2 & \text{elsewhere} \end{cases}. \tag{2.9}$$

The introduction of quantity $\beta|\widehat{N}(\nu)|^2$, instead of zero as in equation (2.9), allows to add a broadband noise that will hide neighboring tonal components of the same amplitude or comparable. The parameters α and β are intended to find a compromise between the amount of musical noise and the distortion of the signal. Proper adjustment of these two parameters has a great influence on the quality of the result. The conducted experiments showed that the parameter α depends on the SNR of the corresponding segmental frame, noted $\mathrm{SNR}_{\mathrm{seg}}$:

$$\alpha = \alpha_0 - \frac{3}{2}\mathrm{SNR}_{\mathrm{seg}} \tag{2.10}$$

where α_0 is between 3 and 6. Equation (2.10) is given for a $\mathrm{SNR}_{\mathrm{seg}}$ between $\pm 6\,\mathrm{dB}$, The variable β is very sensitive to noise level. For very high noise levels ($-6\,\mathrm{dB}$), β must be between $0.02 \leq \beta \leq 0.06$. For a low noise ($0\,\mathrm{dB}$ or $5\,\mathrm{dB}$), it is better to choose a β such that $0.005 \leq \beta \leq 0.02$. In the same way, Berouti *et al.* lead to an experimental choice of the parameters α and β.

In the following subsection we present some basic notions of multi-objective optimization theory.

2.3. Multi-objective optimization background

In Multi-Objective Optimization, it is often unclear what constitutes an optimal solution. A solution may be optimal for one objective function, but it is not for another.

Mathematically, the MOO problem using the Pareto method can be written as follows:

$$(P)\begin{cases} \text{“min”}\ z^k = Q^k(x)x & k = 1, \ldots, K \\ \text{s.t.} \qquad\quad x \in S \end{cases} \tag{2.11}$$

where $S = \{x \in \mathbb{R}^n | Ax \leq b; x \geq 0\}$, $x \in \mathbb{R}^{n\times 1}$ and $b \in \mathbb{R}^{m\times 1}$.

$Q^k(x)$, $\forall k \in \{1, \ldots, K\}$ are quadratic or linear convex functions. Consider a convex function ψ defined by:

$$\psi : \mathbb{R}^n \to \mathbb{R}^K \tag{2.12}$$

$S \subset \mathbb{R}^n$ is the admissible region in the decision space variables and $\psi(S) \subset \mathbb{R}^K$ is the admissible region in the criteria space variables.

The optimal values in MOO are achieved when one objective function cannot increases without reducing the other objective function. This condition is called Pareto optimality. The set of optimal solutions in MOO is called Pareto optimal solutions set. A non Pareto optimal solution is called dominated solution (see [10]).

A point $\overline{x} \in S$ is called an efficient solution of Problem (P), if there is no point $x \in S$ such that $Q^k(x) \leq Q^k(\overline{x})$ $\forall k \in \{1, \ldots, K\}$and $\exists \ell \in \{1, \ldots, K\}$ such that $Q^\ell(x) < Q^\ell(\overline{x})$. Otherwise, $\overline{x}$ is not efficient and the vector $\left(Q^k(\overline{x})\right)_{k\in\{1,\ldots,K\}}$ is dominated.

A common approach for the solution of MOO's is to transform the original multicriteria problem into a series of scalarized, single criterion subproblems which are then solved using classical methods from constrained or unconstrained programming.

Let Λ be a $(K-1)-$simplex defined by

$$\Lambda = \left\{ \lambda \in \mathbb{R}^K \mid \sum_{i=1}^{K} \lambda_i = 1,\ \lambda_i \geq \epsilon\ (i = 1, 2, \ldots, K) \right\}$$

where ϵ is a positive number such that $K\epsilon < 1$.

It can be shown that for sufficiently small ϵ, a point $\overline{x} \in S$ is efficient if and only if there exists $\lambda \in \Lambda$ such that $\lambda C\overline{x} \equiv \max\{\lambda Cx \mid x \in S\}$, (see [7, 12]).

3. The Proposed method

3.1. Multi objective problem model

In our case, we are faced to an optimization problem that is not mathematically well written but realistic. We have two contradictory non linear objective functions: speech distortion and noise reduction (noise reduction inevitably involves the distortion of the enhanced signal). The relative signal has to be segmented (between 20 ms and 30 ms frames) in order to obtain stationarity that permits an accurate analysis using relative mathematical tools. To ensure temporal continuity of the characteristics of the analysis and the model, the considered frames have to be little bit overlapped.

Two decision variables are considered:

- α: Parameters of over-subtraction in a frame.
- β: Spectral flatness parameters in a frame.

The decision space variables is constrained by the inequalities $1 \leq \alpha \leq 6$; $\alpha \in \mathbb{R}$ and $0 \leq \beta < 1$; $\beta \in \mathbb{R}$.

The first non linear objective function that allows us to reduce the amount of speech distortion is the following

$$f_1(\alpha, \beta) = \left(|\widehat{S}(v)|^2 - |S(v)|^2 \right) = \begin{cases} \left(|X(v)|^2 - |S(v)|^2 - \alpha \cdot |\widehat{N}(v)|^2 \right)^2 & \text{if} \quad |X(v)|^2 \geq (\alpha + \beta) \cdot |\widehat{N}(v)|^2 \\ \left(\beta \cdot |\widehat{N}(v)|^2 - |S(v)|^2 \right)^2 & \text{elsewhere} \end{cases}.$$

The second linear objective function, which allows us to reduce the noise, is defined as the ratio of the signal to the noise:

$$f_2(\alpha, \beta) = \frac{|\widehat{S}(v)|^2}{(|\widehat{N}(v)|^2} = \begin{cases} \alpha - \dfrac{|X(v)|^2}{|\widehat{N}(v)|^2} & \text{si } (\alpha + \beta) < \dfrac{X|(v)|^2}{|\widehat{N}(v)|^2} \\ -\beta & \text{otherwise} \end{cases}.$$

The multi objective problem model is given by

$$(P) \begin{cases} \text{“min”} f_1(\alpha, \beta) \\ \text{“max”} f_2(\alpha, \beta) \\ (\alpha, \beta) \in S \end{cases} \tag{3.1}$$

where $S = \{(\alpha, \beta) \in \mathbb{R}^2 | 1 \leq \alpha \leq 6 \text{ and } 0 \leq \beta < 1\}$.

Both objectives in the equation (3.1) are contradictory. In fact, one can observe, when minimizing the distortion $f_1(\alpha, \beta)$ the noise is being augmented. This causes a deterioration in the second objective value $f_2(\alpha, \beta)$ (since $f_2(\alpha, \beta)$ represents the ratio of signal to the noise (SNR)).

Our methodology consists of considering two approaches: one is the lexicographic technique where $f_1(\alpha, \beta)$ is optimized first then $f_2(\alpha, \beta)$ is evaluated and vice versa. The second approach is a convex aggregation of $f_1(\alpha, \beta)$ and $f_2(\alpha, \beta)$ which satisfies the conditions of the efficiency. Theorem of Geoffrion 1968 justify the use of convex aggregation.

In addition, as we consider many frames in order to achieve the stationarity of the signal, the bi-objective model should be solved for each frame. Other complexity in our modelization is the fact that the problem is

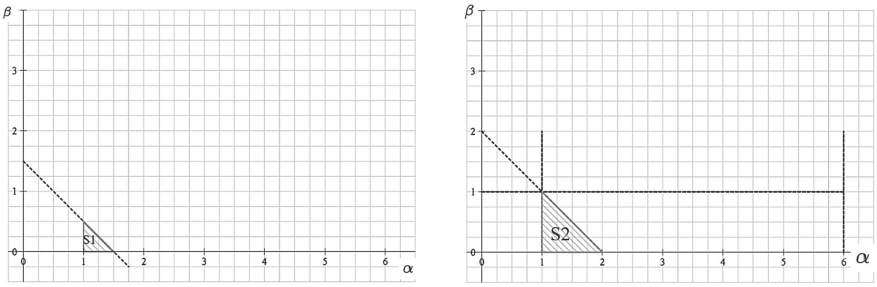

FIGURE 2. Case: $\frac{|X(v)|^2}{|\widehat{N}(v)|^2} < 2$ and $\frac{|X(v)|^2}{|\widehat{N}(v)|^2} = 2$.

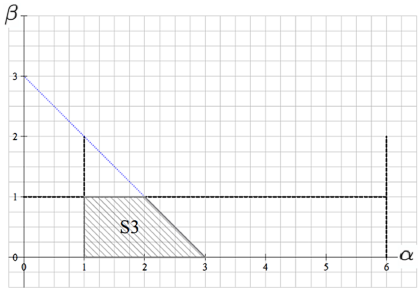

FIGURE 3. Case: $\frac{|X(v)|^2}{|\widehat{N}(v)|^2} > 2$.

not linear. In our study, however, we do need to find all non dominated solutions, but, only one efficient couple (α, β) that has the highest SNR per frame. Once the relative output signal is obtained, the whole signal is composed from these outputs. In our strategy, we focus on only one efficient couple (α, β) per frame to retrieve the denoised respective stationary signal.

For a given frame, we can have one of the domains shown in the Figures 2 and 3.

3.2. The resolution

The problem has a non linear bi-objective nature within many frames, its resolution is very difficult and for each frame there is an infinite number of efficient solutions. In our case, we use weighted sum method and lexicographic method associated to an approximated method.

Particle Swarm Optimization (PSO) algorithm (for more details see [9]) is used to develop our meta-heuristic. The latter is based on population of solutions, it starts with a random initialization of the swarm in the search space, the size of the swarm, N, is fixed initially and each particle represents a potential solution in the search space.

The size of the search space is $D = 2$, so the particle i of the swarm is modelled by a position vector $\overrightarrow{x}_i = (\alpha, \beta) = (x_{i,1}, x_{i,2})$ and by its velocity vector $\overrightarrow{v}_i = (v_{i,1}, v_{i,2})$. This particle keeps in memory the best position by which it has already passed, that we note $P_{\text{best}_i} = (p_{\text{best}_{i,1}}, p_{\text{best}_{i_2}})$. The best position reached by all the particles of the swarm is $\overrightarrow{G}_{\text{best}} = (g_{\text{best}_1}, g_{\text{best}_2})$. Hence each particle will be influenced by its best known position and by the best known position in the search space, thus the swarm is guided to the best solutions. The new position of a particle at the iteration $t+1$ is determined by the equation (3.2).

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{3.2}$$

where for $j \in \{1, 2\}$

$$v_{i,j}^{t+1} = v_{i,j}^{t} + c_1 r_1 (P_{\text{best}_{i,j}}^{t} - x_{i,j}^{t}) + c_2 r_2 (g_{\text{best}_j}^{t} - x_{i,j}^{t}) \tag{3.3}$$

with

- r_1 and r_2 are two random numbers uniformly distributed in the interval [1,0];
- c_1 and c_2 are two parameters representing the confidence of the particle in itself (cognition) and in the swarm (social behavior), respectively. These two parameters are among the most important parameters of the algorithm in that they control the balance between exploration and exploration trends. A relatively high value of c_1 will encourage particles to move to their best local experiences, while higher values of c_2 will result in faster convergence to the best overall position.

For the case of a particle that leaves the search domain, it will be returned to the search space by applying one of the following strategies:

- The particle is left outside the search space, but its objective function is not evaluated. Thus, it will not be able to attract the other particles outside the search space;
- The particle is stopped at the boundary and the components associated with its speed are cancelled;
- The particle bounces on the border. The particle is stopped on the boundary, but the corresponding components of the velocity are multiplied by a coefficient drawn randomly in the interval $[-1, 0]$.

We chose the second case because it is simple to implement. Several authors have reported that the use of a constricting factor usually gives a better rate of convergence, without having to set maximum speed $V_{\max}$. For this reason, we will use the variation that has been demonstrated by Clerc and Kennedy and that it consists in adding factor of construction of χ which aims to control the speed of the particles to avoid the problem of swarm divergence that causes the premature convergence of the algorithm and allows the system to converge. The implemented meta-heuristic is technically presented in the Algorithm 1. In the Algorithm 1, we proposed to evaluate the particle's position with f_1 then we consider f_2 to evaluate the particle's position. This version of algorithm is referred to as "SSBPSO$_{12}$". Also, an other version of the algorithm can be considered where we start with f_2 to evaluate the particle's position then f_1 evaluate the particle's position. This version of algorithm is referred to as "SSBPSO$_{21}$". The last version of algorithm consider the aggregation of the two objective functions f_1 and f_2. It is referred to as Spectral Subtraction based PSO with Aggregation denoted "SSBPSOaggr". The performance of the proposed algorithms as compared to the benchmark algorithms proposed by Berouti *et al.* and Boll are shown in the following section.

4. Results

In this section, we present a fair comparison between our proposed methods and the benchmark algorithms proposed by Berouti *et al.* [5] and Boll [6]. This comparison is assessed by considering three type of criteria:

(1) Subjective criteria: The evaluation of speech and noise is summarized in Table 1. The considered criteria in our work are:
   - SIG: it allows ua to evaluate the speech distortion and considers five scores of speech signal distortion as given in the Table 1.
   - BAK: it allows us to evaluate the background noise and considers five scores as given in the Table 1.
   - OVRL: it allows us to evaluate the global quality of speech and considers five scores as given in the Table 1.

(2) Objective criteria: to assess the speech quality using the objective criteria, one compute the distance between the clean signal (without noise) and the processed signal (after denoising). The considered criteria are:

Algorithm 1: THE SSBPSO_{12} algorithm.

Input
$\downarrow N$: the size of swarm particles;
$\downarrow x_i$, v_i: random position and speed respectively for each particle;
Output
$\uparrow G_{Best}$
Initialization

- $\overrightarrow{P}_{Best_i} \leftarrow x_i$ For the ith particle;
- Compute $\overrightarrow{G}_{Best}$ as the solution of
  $g_{Best_1} = \arg\left\{ \min_{\overrightarrow{P}_{Best_i}} f_1(\overrightarrow{P}_{Best_i}(1)),\ 1 \le i \le N \right\}$;

while *(Number of iteration is achieved)* **do**

- ► For each particle Compute the particle's movement for $j \in \{1, 2\}$ as

$$\begin{cases} v_{i,j}^{t+1} = \chi \left(v_{i,j}^{t} + c_1 r_{1_{i,j}}^{t} \left[P_{Best_{i,j}}^{t} - x_{i,j}^{t} \right] \right. \\ \qquad \left. + \ c_2 r_{2_{i,j}}^{t} \left[g_{Best_j}^{t} - x_{i,j}^{t} \right] \right) \\ x_{i,j}^{t+1} = x_{i,j}^{t} + v_{i,j}^{t+1} \end{cases}$$

- ► Evaluate the positions of particles by the first objective function f_1;
- ► Update $\overrightarrow{P}_{Best_i}$ and $\overrightarrow{G}_{Best}$ as follows:

$$\begin{cases} P_{Best_i} \leftarrow P_{Best_i}(t+1) & \text{if } f_1(x_i(t+1)) \ge P_{Best_i} \\ P_{Best_i} \leftarrow x_i(t+1) & \text{otherwise} \end{cases}$$

$g^* \leftarrow G_{Best}$: The best achieved position;
Evaluate the particle's position with f_2;
while *(Number of iteration is achieved)* **do**

- ► For each particle Compute the particle's movement for $j \in \{1, 2\}$ as

$$\begin{cases} v_{i,j}^{t+1} = \chi \left(v_{i,j}^{t} + c_1 r_{1_{i,j}}^{t} \left[p_{Best} t_{i,j}^{t} - x_{i,j}^{t} \right] \right. \\ \qquad \left. + \ c_2 r_{2_{i,j}}^{t} \left[g_{Best} t_{i,j}^{t} - x_{i,j}^{t} \right] \right) \\ x_{i,j}^{t+1} = x_{i,j}^{t} + v_{i,j}^{t+1} \end{cases}$$

- ► Evaluate the positions of particles by the second objective function f_2;

if $f_2(x_i) < f_2(P_{Best_i})$ *and* $f_1(x_i) = g^*$ **then** $P_{Best} \leftarrow x_i$ **else**
  if $f_2(P_{Best_i}) < f_2(G_{Best})$ *and* $f_1(x_i) = g^*$ **then** $G_{Best} \leftarrow P_{Best_i}$

Return $(\alpha, \beta) \leftarrow G_{Best}$

- Segmental Signal to Noise Ration (SNR_{seg}): It is considered as temporal domain criterion. It is obtained by computing the SNR of each frame of the speech signal (the latter is segmented into several frames.) as following:

$$\mathbf{SNR}_{\text{seg}} = \frac{10}{M} \sum_{m=0}^{M-1} \log \frac{\sum_{n=mN}^{mN+N-1} x^2(n)}{\sum_{n=mN}^{mN+N-1} (s(n) - (\hat{s})(n))^2} \tag{4.1}$$

where N is frame size.

TABLE 1. Speech and noise evaluation with respect to subjective criteria: SIG, BAK and OVRL.

	SIG	BAK	OVRL
Score	The speech signal is	The background noise is	The voice sequence is
5	Without distortion	Imperceptible	Excellent
4	Slightly distorted	Slightly imperceptible	Good
3	Somewhat distorted	Perceptible but not annoying	Fair
2	Fairly distorted	Somewhat annoying	Poor
1	Very distorted	Very annoying	Bad

- Log-Likelihood Ratio (LLR): It is considered as spectral domain criterion. It is based on autoregrissive model of the speech and it is given by

$$\mathbf{LLR} = \log\left(\frac{a_y \mathbf{R}_x a_y}{a_x \mathbf{R}_x a_x}\right) \tag{4.2}$$

  where $\mathbf{R}_x$ is the autocorrelation matrix a_x and a_y are LP coefficient vectors. when the criterion tends to zero, the resemblance between the clean signal and the processed one becomes stronger.
- Weighted-Slope Spectral distance (WSS): It is considered as spectral domain criterion too. The global degradation WSS is obtained by averaging the degradations of the frames. It is given by

$$\mathbf{WSS} = \frac{1}{M}\sum_{m=0}^{M-1} \frac{\sum_{k=1}^{K} W(k,m)\left[S(\nu_k,m) - \hat{S}(\nu_j,m)\right]^2}{\sum_{k=1}^{K} W(k,m)} \tag{4.3}$$

  where $W(k,m)$ are the computed weights, $S(\nu_k,m)$ and $\hat{S}(\nu_j,m)$ are the spectral slopes for jth frequency band and mth frame of clean and processed signals respectively.
- PESQ (Perceptual Evaluation of Speech Quality): It is considered as perceptual criterion. It allows to evaluate the quality of listening under many conditions of degradation, resulting in a correlation close to the subjective notes.

(3) Composite measurement criteria: To perform the evaluation of the speech quality, one can define composite measurements criteria. These latter are obtained by linear combination of the above criteria. Herein, one define the following composite measurement criteria:
- C_{sig}: it allows to evaluate the signal distortion and it is obtained by linear combination of LLR, PESQ and WSS. The coefficients of this combination are determined by minimizing the mean square error between the objective criterion C_{sig} and the SIG criterion. C_{sig} is given by:

$$\mathbf{C}_{\text{sig}} = 3.093 - 1.029\ \mathbf{LLR} + 0.603\ \mathbf{PESQ} - 0.009\ \mathbf{WSS}. \tag{4.4}$$

- C_{bak}: it allows to evaluate the residual noise and it is obtained by linear combination of PESQ, WSS and SNR_{sig}. It is given by

$$\mathbf{C}_{\text{bak}} = 1.634 + 0.478\ \mathbf{PESQ} - 0.007\ \mathbf{WSS} + 0.063\ \mathbf{SNR}_{\text{sig}}. \tag{4.5}$$

- C_{ovl}: it allows to evaluate the global quality of the denoised frame. It is given by:

$$\mathbf{C}_{\text{ovl}} = 1.594 + 0.805\ \mathbf{PESQ} - 0.512\ \mathbf{WSS} - 0.512\ \mathbf{LLR}. \tag{4.6}$$

Our proposed algorithms are compared to those proposed by Berouti *et al.* [5] and Boll [6]. To assess this comparison, one considers typical speech signal from noisex-92 database where the signal is recorded in specific

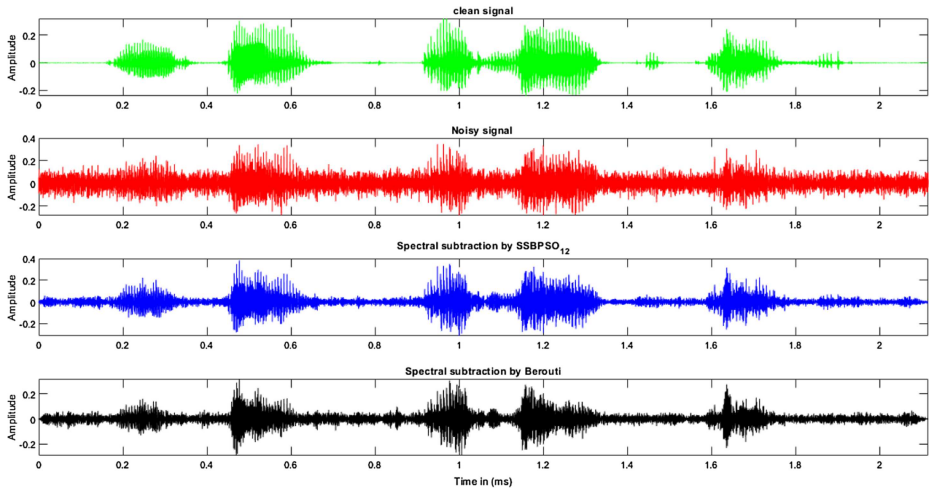

FIGURE 4. Temporal representation of speech enhancement for noisy signal at 0 dB for car environment.

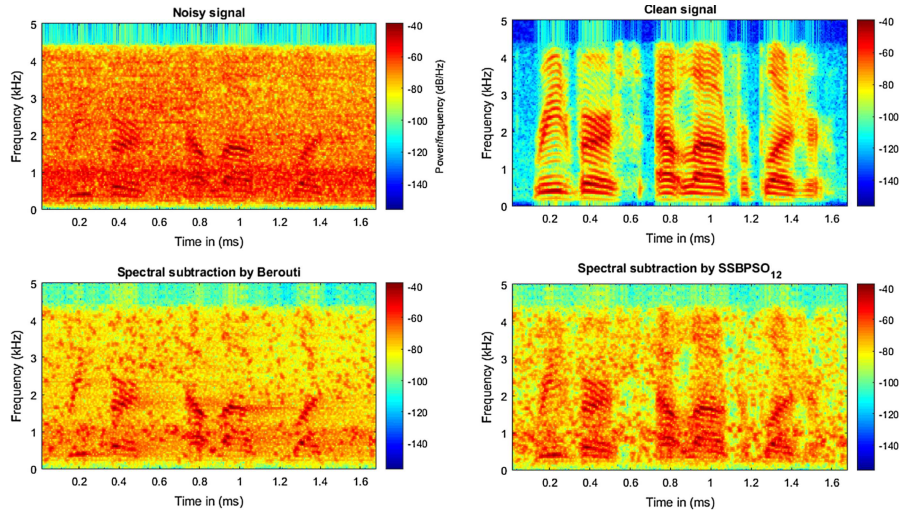

FIGURE 5. Spectral representation of speech enhancement for noisy signal at 0 dB for car environment.

condition (Airport, Car and Babel environments) and for different SNR levels (0 dB, 5 dB, 10 dB and 15 dB). Du to the space limitation, we limit ourselves to corrupted speech signals in Airport environment.

Our experimentations are realized by considering the following parameters: The initial silence (noise only) length is equal to 0.25 s and the sampling frequency is 8000 Hz. The signals are segmented into 210 frames. For Berouti's algorithm, β is fixed to 0.03 and the size of the random population is equal to 15 for PSO algorithm.

First, we show the time and spectral representations in Figures 4 and 5 respectively for very hostile environment (SNR = 0 dB). From these latter, one can observe that the proposed algorithms performs well and the obtained representations are close to the representations of clean signal. However, one needs to compute the previous criteria in order to have a fair comparison of our algorithms.

TABLE 2. LLR Criteria for airport environment.

SNR /Algorithm	SSBerouti	SSBol	$SSBPSO_{12}$	$SSBPSO_{21}$	SSBPSOaggr
0 dB	0,97759068	1.0167736	0.609637772	0.698367733	1.082450529
5 dB	0,82329599	0.82019281	0.467750479	0.525550612	0.967402554
10 dB	0,6074029	0.74160592	0.392434165	0.423415228	0.813120916
15 dB	0,48494509	0.57965148	0.312802158	0.306082303	0.752071654

TABLE 3. SNR_{seg} Criteria for airport environment.

SNR /Algorithm	SSBerouti	SSBol	$SSBPSO_{12}$	$SSBPSO_{21}$	SSBPSOaggr
0 dB	−3.659279142	−1.468198313	0.163661332	−1.945483943	0.983470618
5 dB	−2.250728397	0.695718728	1.166108518	−0.456095405	1.875493744
10 dB	−0.227545583	2.065264016	2.479453917	1.065405289	3.109460636
15 dB	0.894974392	3.04215772	3.38212537	3.083587536	4.567646897

TABLE 4. WSS Criteria for airport environment.

SNR /Algorithm	SSBerouti	SSBol	$SSBPSO_{12}$	$SSBPSO_{21}$	SSBPSOaggr
0 dB	72.4292605	86.42219	56.01946508	70.0228485	56.83657247
5 dB	62.9661795	72.0758978	47.78114659	54.40656155	49.94542114
10 dB	49.4873484	58.8066567	35.91735251	43.63332492	38.03858355
15 dB	41.9559725	52.940283	32.03124618	31.27676582	30.99677776

Note that, for the following obtained results, we considered only the airport environment. The similar results are obtained for car environment and Babel environment from noisex-92 database.

The numerical values of different criteria for different algorithms are given in Tables 2–8 and their graphical representation are given in Figures 6–12. From these figures, one can observe that our proposed algorithms outperforms the Berouti's algorithm. For example, in Figures 6, 8 and 9, one shows the evolution of the most appropriate three parameters to evaluate the speech signal distortion (LLR, WSS and PESQ). Herein, our proposed algorithms produce very low distortion as compared to Berouti's algorithm. We improved the performance of denoising the speech signal up to 37.63% for very hostile environment (SNR = 0 dB) based on LLR criterion (see Fig. 6 and Tab. 2). If considering WWS criteria, one improved the performance up to 22.65% in the same conditions (see Fig. 8 and Tab. 4). Based on the perceptual evaluation (PESQ), one improved the denoising performance up to 34.69% (see Fig. 9 and Tab. 5). Finally, our algorithms greatly reduce the residual noise batter than Berouti's algorithm where the obtained SNR_{seg} by our algorithm is higher than the obtained one by Berouti's algorithm with 3.82 dB which means that we have a very low residual noise (see Fig. 7 and Tab. 3).

In order to compare the obtained values of the parameter α by considering Berouti's algorithm and our proposed algorithms, we plotted the obtained values *versus* frame index in Figures 13–16. From the latter, one can observe that the obtained values of α for PSO_{12} and PSO_{21} are lower than those obtained by Berouti' algorithm especially in the case of high level of noise (SNR = 0 dB in Fig. 13). Also, PSO_{21} gives more stable values of α over frames with mean value about 2.5.

Remark 4.1. We carried out several tests with different initial conditions and we observed that the proposed algorithm using the operational research outperforms the traditional algorithms but not necessary for the same numerical values of (α, β).

TABLE 5. PESQ Criteria for airport environment.

SNR /Algorithm	SSBerouti	SSBol	$SSBPSO_{12}$	$SSBPSO_{21}$	SSBPSOaggr
0 dB	1.88059498	1.9433524	2.533140807	2.2623221	1.987207604
5 dB	2.31598388	2.2893901	2.837409664	2.740571146	2.292170455
10 dB	2.57744388	2.48688878	3.259132231	3.056980739	2.62030116
15 dB	2.86049336	2.88917132	3.329843981	3.309807378	2.574267998

TABLE 6. C_{sig} Criteria for airport environment.

SNR /Algorithm	SSBerouti	SSBol	$SSBPSO_{12}$	$SSBPSO_{21}$	SSBPSOaggr
0 dB	2.56919462	2.44078175	3.488991453	3.108354193	2.665915439
5 dB	3.07567109	2.98084075	3.892612465	3.715113767	3.030212766
10 dB	3.57679494	3.30022154	4.331185807	4.107965192	3.493992925
15 dB	3.94126525	3.76224639	4.490741284	4.492364267	3.59243087

TABLE 7. C_{bak} Criteria for airport environment.

SNR /Algorithm	SSBerouti	SSBol	$SSBPSO_{12}$	$SSBPSO_{21}$	SSBPSOaggr
0 dB	1.79538499	1.86547063	2.463015714	2.102664536	2.247987876
5 dB	2.15848115	2.26762746	2.72927863	2.534413066	2.498195636
10 dB	2.50527136	2.54119787	3.096649336	2.856924052	2.81612989
15 dB	2.7640074	2.83609785	3.214520598	3.191416581	2.935284413

TABLE 8. C_{ovrl} Criteria for airport environment.

SNR /Algorithm	SSBerouti	SSBol	$SSBPSO_{12}$	$SSBPSO_{21}$	SSBPSOaggr
0 dB	2.10034771	2.03285527	2.928907554	2.567445072	2.241631443
5 dB	2.59607622	2.51248903	3.304158508	3.150231928	2.594269161
10 dB	3.0114406	2.80459664	3.765253686	3.532647624	3.02075444
15 dB	3.35471346	3.25241938	3.890150976	3.88274344	3.064247607

5. CONCLUSION

The improvement of the quality and intelligibility of the speech signal is not simple because of the non-stationary of signal and of the noise. In addition, the noise presents a random nature depending on the hostile environment and the technology adopted such as mobile telephony or hearing aids. Nevertheless, it is clear from this paper that it is always possible to contribute to an improvement of the speech signal by developing appropriate methods and algorithms.

In this paper, we have introduced multi-objective optimization in the treatment of the problem of speech signal enhancement, which is usually handled by various methods such as Wiener, Kalman filtering or that among the most classic, the spectral subtraction proposed by Berouti *et al.* that we have addressed in this work.

We have formulated the appropriate mathematical model with the aim of improving the quality and intelligibility of the speech signal through a compromise between musical noise and distortion.

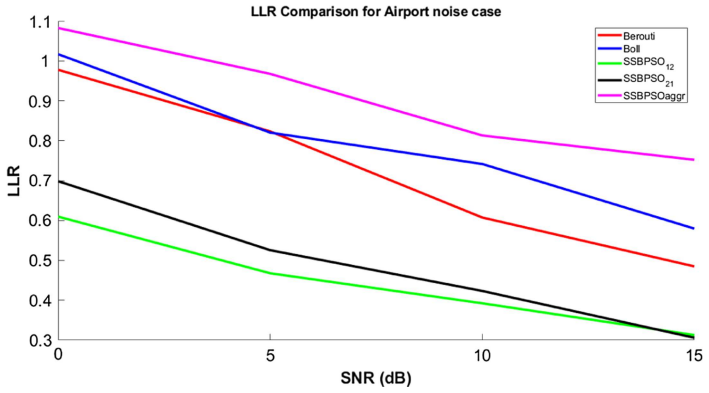

FIGURE 6. LLR comparison for airport environment.

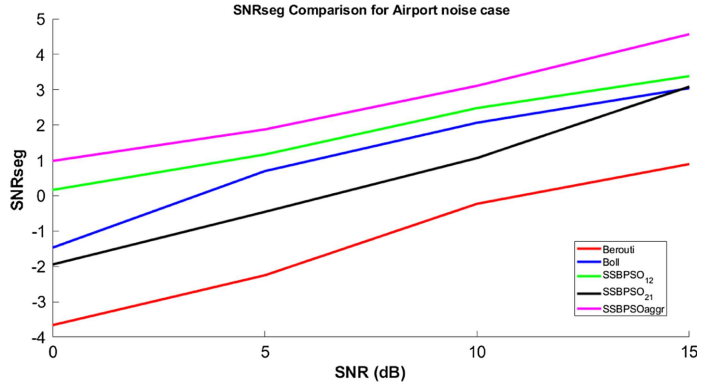

FIGURE 7. $\mathrm{SNR}_{\mathrm{seg}}$ comparison for airport environment.

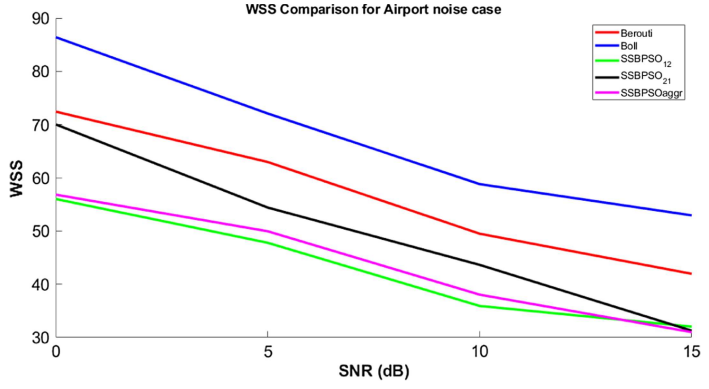

FIGURE 8. WSS comparison for airport environment.

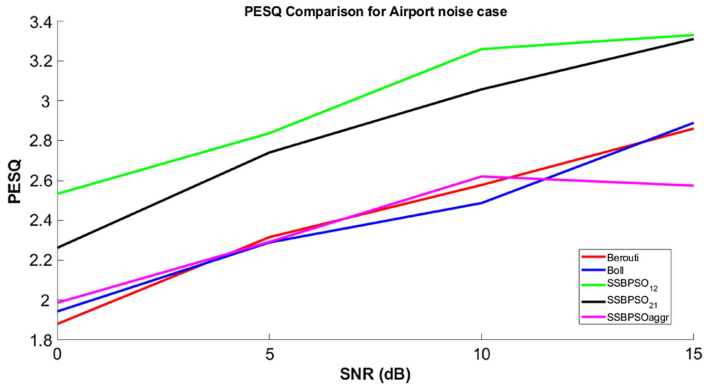

FIGURE 9. PESQ comparison for airport environment.

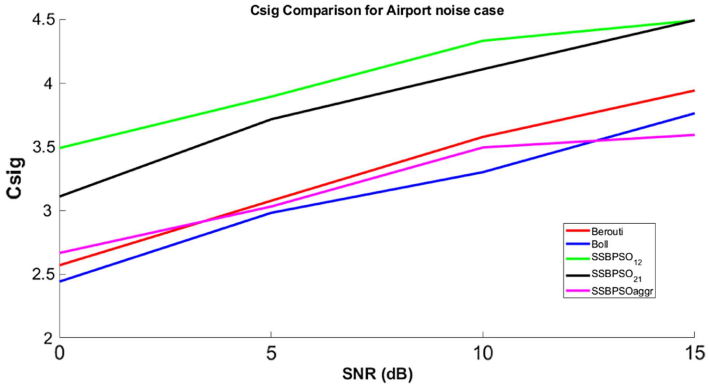

FIGURE 10. C_{sig} comparison for airport environment.

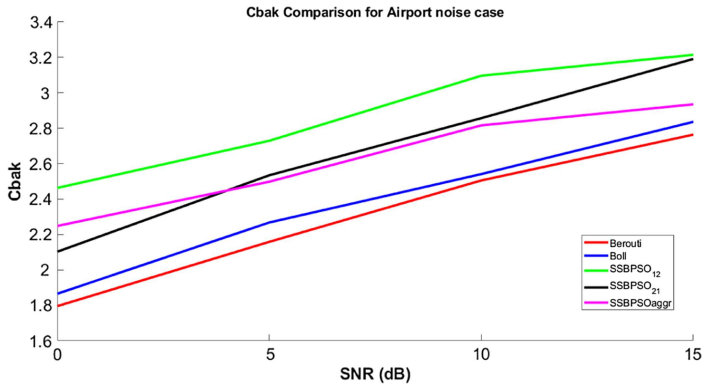

FIGURE 11. C_{bak} comparison for airport environment.

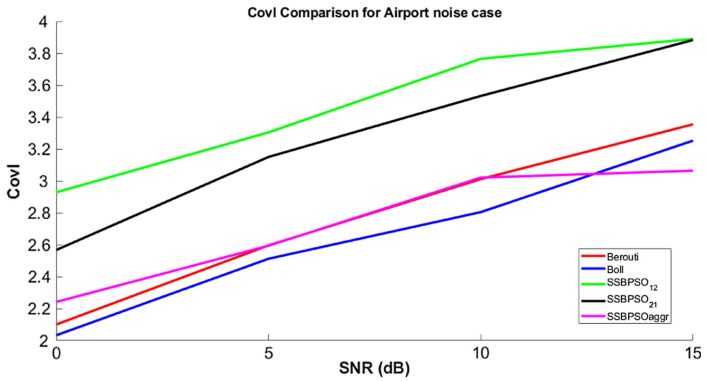

FIGURE 12. C_{ovrl} Comparison for airport environment.

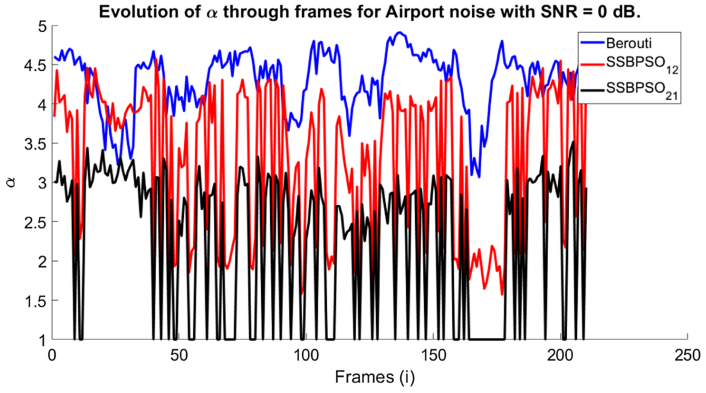

FIGURE 13. Values of α *versus* frame index for SNR = 0 dB in airport environment.

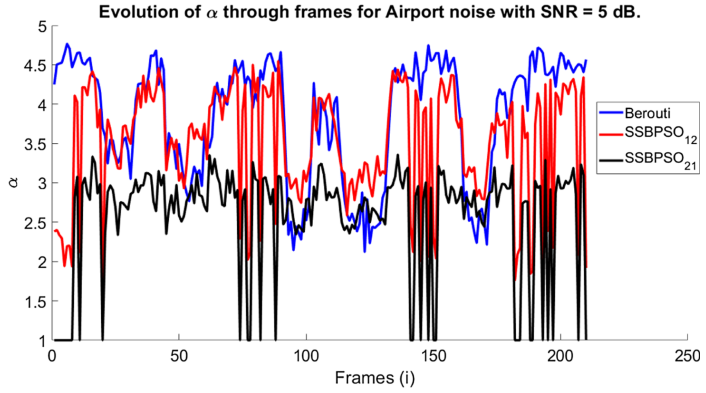

FIGURE 14. Values of α *versus* frame index for SNR = 5 dB in airport environment.

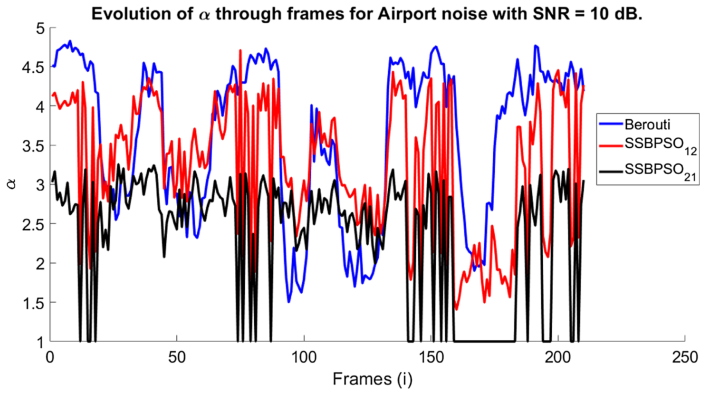

FIGURE 15. Values of α *versus* frame index for SNR = 10 dB in airport environment.

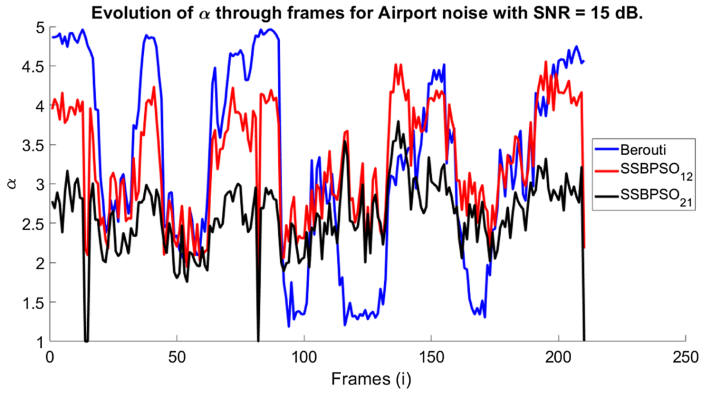

FIGURE 16. Values of α *versus* frame index for SNR = 15 dB in airport environment.

To solve the optimization problem, we developed a meta-heuristics based on the particle swarm optimization technique. Through measurements based on objective and subjective evaluation criteria (SNR, PESQ), the experimental results show that the performance is better compared to voice quality and intelligibility. The proposed method surpasses the algorithm developed by Berouti *et al.*

Acknowledgements. We thank all the anonymous reviewers for their comments that greatly improved the manuscript and the DGRSDT of the Algerian Ministry of Higher Education and Research for the financial support received.

REFERENCES

[1] A. Amehraye, D. Pastor, A. Tamtaoui and D. Aboutajdine, From maskee to audible noise, in perceptual speech enhancement. *Int. J. Signal Process.* **5** (2008) 93–96.

[2] L.B. Asl and V.M. Nezhad, Improved particle swarm optimization for dual channel speech enhancement. In: *International Conference on Signal Acquisition and Processing. IEEE* (2010).

[3] L.B. Asl and V.M. Nezhad, Speech enhancement using particle swarm optimization techniques. In: *International Conference on Measuring Technology and Mechatronics Automation. IEEE* (2010) 441–444.

[4] D.W. Beeks and R. Collins, Speech Recognition and Synthesis. CRC Press (2001).

[5] M. Berouti, B. Beranek, N.R. Schwartz and J. Makhoul, Enhancement of speech corrupted by acoustic noise. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-79* (1979) 208–211.

[6] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **27** (1979) 113–120.

[7] D. Chaabane, *Contribution à l'Optimisation Multicritère en Variables Discrètes*. Ph.D. thesis, UMONS, Polytechnique Faculty of Mons, Belgium (2007).

[8] V. Chapke and H. Kaur, Review of speech enhancement techniques using statistical approach. In Vol 5 of *International Journal of Electronics Communication and Computer Engineering. Technovision-2014* (2014).

[9] R.C. Eberhart and J.Kennady, A new optimizer using particles swarm theory. In: *Proceedings of the Sixth International Symposium on Micro Machine and Human Science. IEEE Press, Piscataway, NJ, Nagoya, Japan* (1995) 39–43.

[10] M. Ehrgott, Multicriteria Optimization, 2nd edition. Springer Berlin Heidelberg (2005).

[11] V.M. Gaikwad and S.S. Vasekar, Survey on quality and intelligibility offered by speech enhancement algorithms. In: *International Conference on Computing Communication Control and Automation. IEEE* (2015).

[12] A.M. Geoffrion, Proper efficiency and the theory of vector maximization. *J. Math. Anal. App.* **22** (1968) 618–630.

[13] P. Handel, Power spectral density error analysis of spectral subtraction type of speech enhancement methods. *EURASIP J. Adv. Signal Process.* **2007** (2006) 096384.

[14] R.C. Hendriks, R. Heusdens and J. Jensen, An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model. *IEEE, Trans. Audio Speech Lang. Process.* **15** (2007) 406–415.

[15] I. Isewon, J. Oyelade and O. Oladipupo, Design and implementation of text to speech conversion for visually impaired people. In: Vol. 7 of *International Journal of Applied Information Systems (IJAIS)*. Foundation of Computer Science FCS, New York, USA (2014).

[16] J. Kennedy, R.C. Eberhart and Y. Shi, Y., Swarm Intelligence. Morgan Kaufmann Publ, San Francisco (2001).

[17] B. Kollmeier, Psychoacoustic, Speech and Hearing Aids. World Scientific Publishing Co. Pvt. Ltd. (1995).

[18] P. Kunche and K.V.V.S. Reddy, Metaheuristic Applications to Speech Enhancement. *Springer Briefs in Electrical and Computer Engineering, Speech Technology*. Springer Series Editor, Amy Neustein, Fort, NJ, USA Lee (2016).

[19] P. Kunche, K.V.V.S. Reddy, G.S.B. Rao and R.U. Maheswari, A new approch to dual channel speech enhanced based on hybrid PSOGSA. *Int. J. Speech Technol.* **18** (2015) 45–56.

[20] S. Li, J.Q. Wang and X.J. Jing, The application of nonlinear spectral subtraction method on millimeter wave conducted speech enhancement. *Math. Probl. Eng.* **2010** (2010) 371782.

[21] Y. Liu, Y. Shi and X. Ma, Single Channel Speech Enhancement Using Complex Kalman Filter in Noisy Reverberant Environments. Mobimedia, Qingdao, China (2018).

[22] P.C. Loizou, Speech Enhancement: Theory and Practice, 2nd edition. CRC, Press (2017).

[23] S. Ouznadji, D. Chaabane and M. Thameri, Multiple objective optimization applied to speech enhancement problem. In: *2017 International Conference on Mathematics and Information Technology (ICMIT)* (2017) 24–28.

[24] M. Pariente, *Machine learning applied to audio source separation for audio prosthesis*. Master, Université Paris Descartes, École Normale Supérieure, École des Hautes (2017).

[25] E. Plourde, B. Champagne, Auditory based spectral amplitude estimators for speech enhancement. In: Vol. 16 of *IEEE Transactions on Audio, Speech and Language Processing* (2008) 1614–1623.

[26] E. Plourde, *Baysien Short-time Spectral Amplitude Estimators for single-Channel Speech Enhancement*. Thèse de doctorat, Université McGill (2009) 168.

[27] P. Rajalakshmi and P. Kopperundevi, A novel approach to speech enhancement using modified spectral subtraction. *IJAREEIE* **6** (2017) 291–296.

[28] R. Ram and M.N. Mohanty, The use of deep learning in speech enhancement. In: Vol. 14 of *Proceedings of the First International Conference on Information Technology and Knowledge Management*. ACSIS (2018) 107–111.

[29] V. Sailaja and P.Sunitha, Performance analysis of spectral subtraction method for speech enhancement. In: Special Issue from 2nd National Conference on Computing, Electrical, Electronics and Sustainable Energy Systems, Rajahmundry, India (2017).

[30] N.A. Sheela Selvakumari and V. Radha, A hybrid approach for noise reduction using wiener filter and wavelet transform. *Int. J. Pure Appl. Math.* **119** (2018) 731–743.

[31] S. Surendran and T. Kishore Kumar, Perceptual subspace speech enhancement with variance normalization. *Proc. Comput. Sci.* **54** (2015) 818–828.

[32] S.V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, 4th edition. John Wiley & Sons, Ltd. (2008).

[33] S. Verma and A. Garg, Various methods for speech intelligibility enhancement: a brief survey. *IJEEE* **8** (2016).

[34] Y. Yang and C. Bao, DNN-Based AR-Wiener Filtering for Speech Enhancement. In: *ICASSP*. IEEE (2018) 2901–2905.

[35] Z. Ye, M. Zhu and J. Wang, On Modification and Application of the artificial bee colony algorithm. *J. Inf. Process. Syst.* **14** (2018) 448–454.