Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

An evolutionary algorithm to optimize the microphone array configuration for speech acquisition in vehicles

David Ayllón*, Roberto Gil-Pita, Manuel Utrilla-Manso, Manuel Rosa-Zurera

Department of Signal Theory and Communications, University of Alcala, Alcalá de Henares, Madrid 28508, Spain

ARTICLE INFO

Article history:
Received 16 May 2013
Received in revised form
21 January 2014
Accepted 8 May 2014

Keywords:
Microphone array
Speech enhancement
Speech acquisition
Intelligent vehicles

ABSTRACT

Speech acquisition using microphone arrays is included in a variety of trending applications. Multi-channel speech enhancement based on spatial filtering aims at improving the quality of the acquired speech. The optimization of the filter coefficients has been the primary focus in beamformer design. However, the array configuration plays an important role in the quality of the speech acquisition system and it should also be optimized. In some applications, the possibilities for microphone placement are very large, and the search of the optimum solution, which involves exploring all possible microphone configurations, is an unfeasible task. This work presents a novel search algorithm based on evolutionary computation to approximate the optimum array configuration. A realistic car noise model based on real measurements is proposed and used in the design. The obtained results support the suitability of the method, notably improving the results obtained by linear arrays with the same number of elements, which are the typical arrays currently assembled in vehicles.

1. Introduction

Nowadays, microphone arrays are useful for a variety of applications where speech acquisition is essential: multi-conference systems (Elko, 1996), monitoring and room control (Busso et al., 2005), hearing aids (Hamacher et al., 2005) or vehicles (Cho and Krishnamurthy, 2003). The common objective of using a microphone array is to enhance the quality of the acquired speech, although the final application of the enhanced speech may differ: speech enhancement in hearing aids, speech recognition in control systems, or voice coding for mobile communications in vehicles. The success of these speech-based applications relies on the quality of the speech acquired by the microphones, which is usually contaminated by different types of noise and interferences. In the last few years, vehicles have become smarter, offering advanced speech-based applications like hands-free telephony, teleconferencing, speech dialog and recognition, allowing the driver to focus on driving instead of handling an electronic device, which increments the driving safety. These advanced features are also demanded in the military environment to enhance vocal communications among military vehicles and the command post. The requirements in military applications are even higher than in civil applications and the environment conditions are more adverse, hence, the robustness of the speech acquisition system is absolutely essential. Some different applications of microphone arrays in the military environment are found in Okandan et al. (2007) and Goldman (2009).

The use of directional microphones pointing towards the desired speech source is an alternative for low-noise speech acquisition, but directional microphones are normally larger and more expensive than omnidirectional microphones, and they do not allow adapting the directivity to the movements of the desired sound source. The use of microphone arrays built of cheap omnidirectional microphones combined with spatial filtering solve these limitations. Beamforming techniques combine the signals collected at the M input channels of the array in such a way that the signals coming from the desired direction are coherently combined and the signals coming from different directions are incoherently combined. There are different strategies to compute the filter coefficients, for instance, the well-known superdirective beamformer (Hansen and Woodyard, 1938), which optimizes the filter coefficients to maximize the array gain. A comprehensive review of beamforming techniques for speech enhancement can be found in Benesty et al. (2008). Nevertheless, the design of the spatial filter is not the only factor affecting the quality of the acquired speech: the position, geometry and number of microphones of the array also have a strong influence in the performance of the system (Feng et al., 2012). Consequently, the selection of the array configuration should be carefully studied in the design. In some applications, the area available to place the microphones is very large, e.g. within a vehicle or a room. This fact implies a large range of possible array configurations, which makes unfeasible an exhaustive search to obtain the best solution.

* Corresponding author. Tel.: +34 918856662
E-mail address: david.ayllon@uah.es (D. Ayllón).

http://dx.doi.org/10.1016/j.engappai.2014.05.006
0952-1976/© 2014 Elsevier Ltd. All rights reserved.

In this paper, a tailored search algorithm based on evolutionary computation is proposed to approximate the optimum microphone array configuration, in terms of array gain, for speech acquisition in a vehicle. In spite of the fact that different array configurations for vehicles have been proposed during the last few years, for instance, in Ayllón et al. (2012), Grenier (1992), Martin et al. (2001), and Oh et al. (1992), none of them corresponds with an optimized solution but just rough approximations that obtain good results. Evolutionary algorithms have been largely used in engineering to solve optimization and search problems in a wide range of applications, for instance, automatic speech/music discrimination (Ruiz-Reyes et al., 2010), adaptation of non-native speech in a speech recognition system (Selouani and Alotaibi, 2013), antenna array design in Chabuk et al. (2012), or mobile robot localization (Kwok et al., 2006). Additionally, we propose a novel vehicle noise model based on noise recordings under normal driving conditions. The speech inside a vehicle is mainly contaminated by two types of interferences: background noise due to normal driving conditions, which has been usually modeled as a diffuse noise field (Bitzer et al., 1999), and directional signals coming from the loudspeakers. The proposed model is used by the search algorithm to approximate the optimum array.

The remainder of the paper is organized as follows. In Section 2 the proposed car model is described, analyzing the signals recorded in a real car under normal driving conditions. Section 3 describes the signal model and the beamforming technique used in this work. Section 4 describes the proposed search algorithm, and Section 5 contains a description of the experiments carried out in this paper and the obtained results. Finally, Section 6 ends with the conclusions derived from the results.

2. Proposed car model

The background noise due to normal driving conditions inside a car is commonly modeled as a diffuse noise field (Bitzer et al., 1999), which assumes an infinite number of spatially uncorrelated isotropic noise sources. A measure for describing the noise environment is the complex noise field coherence between two signals $x_i(t)$ and $x_j(t)$, with discrete time index t. In the frequency domain, it is defined as (White and Boashash, 1990)

$$\Gamma_{ij}(k) = \frac{\Phi_{x_i x_j}(k)}{\sqrt{\Phi_{x_i}(k)\Phi_{x_j}(k)}}, \tag{1}$$

where $\Phi_{x_i}(k)$ and $\Phi_{x_j}(j)$ represent the auto-power spectral density (PSD) of x_i and x_j, respectively, $\Phi_{x_i x_j}$ represents their cross-PSD, and k represents frequency, $k = 0, \ldots, K-1$. In a diffuse noise field, all microphones receive equal amplitude and random phase noise signals from all directions. In such a case, the coherence between the noise signals acquired by two microphones depends only on the distance between them, according to

$$\Gamma(k) = \text{sinc}\left(\frac{2\pi f d_{nm}}{c}\right), \tag{2}$$

where d_{nm} is the distance between the n-th and the m-th microphone and c is the speed of sound.

In order to validate the assumption of diffuse noise, we have carried out several recordings of the background noise inside a car under normal driving conditions, using an uniform linear array composed of eight omnidirectional microphones with a microphone distance of 35 mm. The analysis of the real noise acquired inside the car reveals that the coherence of that noise differs from the coherence of an ideal diffuse noise field. Fig. 1 shows the spatial coherence of an ideal diffuse noise field with a microphone distance of 35 mm (red line), and the average spatial coherence of the real noise acquired in the car (black line). In the latter case, the coherence has been calculated by averaging the coherence between all possible pairs of consecutive microphones. Although the lower frequencies are highly correlated and the coherence starts to decrease with frequency, for frequencies above 500 Hz the coherence is much higher than the expected for an ideal diffuse noise field. This fact motivates the idea that using a model with a finite number of noise sources is more adequate.

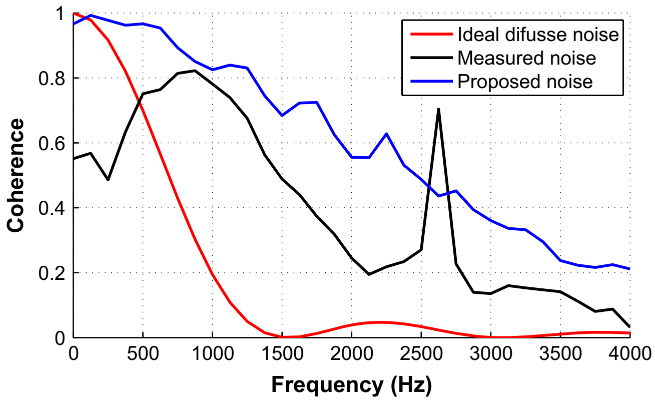

Fig. 1. Spatial coherence of an ideal diffuse noise field (red line), of the real noise acquired in a car (black line), and the proposed noise model (blue line). The coherence has been calculated using two microphones with a distance of 35 mm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

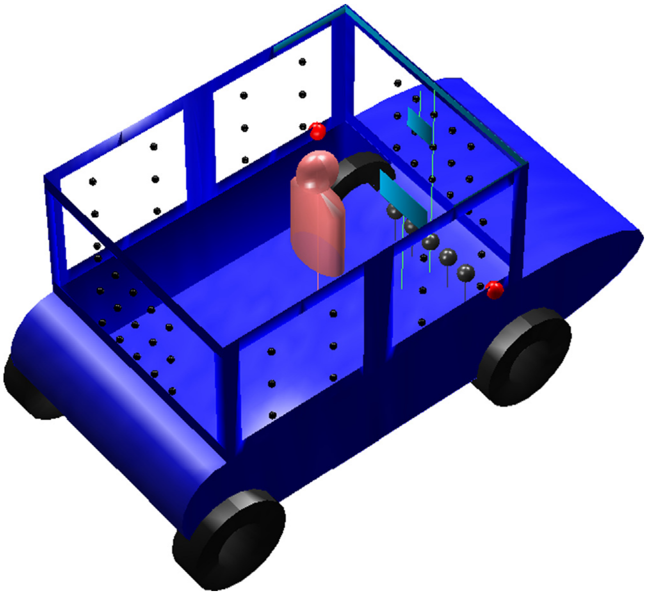

Fig. 2. Car noise model. Black spheres represent the positions of the 63 noise sources, and the two red spheres the front loudspeakers. The target source is located in the head of the dummy driver, and the light blue rectangles represent the areas where the microphone can be placed: the area in the rear-view mirror (A1), the area located in the center of the dashboard (A2), and the upper front edge between the windshield and the roof and the edge between front lateral windows and the roof (A3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

The car model we propose have the following characteristics: (a) the volume of the car is approximated by a cuboid; (b) most of the noise energy inside the car comes from limited areas along its contour, such as windows, wheels and engine; (c) there is a large but finite number of spatially uncorrelated white noise sources distributed along the different areas of the vehicle where noise sources are. Fig. 2 shows the different noise sources considered in this model, represented by black spheres: 15 for the front windscreen and 15 for the rear windscreen, 6 for each lateral window, 1 for each wheel and 5 for the engine, totaling 63 noise sources. The appropriate number of noise sources has been obtained experimentally, trying to fit the spatial coherence of the real noise in the speech frequency band. The coherence calculated between two microphones separated 35 mm using the proposed noise model is also represented in Fig. 1 (blue line). It is clear that the coherence of the proposed model approximates the coherence of the real noise better than the ideal diffuse noise does.

In addition to the background noise sources, the signals coming from the loudspeakers are a different type of interfering source. In the proposed model only the front loudspeakers are considered, which are represented by red spheres in Fig. 2. The target source is assumed to be in the head of the driver, and the light blue rectangles represent the search areas where the microphones can be placed: the area in the rear-view mirror (labeled A1), the area located in the center of the dashboard (labeled A2), and the area that comprises the upper front edge between the windshield and the roof and the edge between both front lateral windows and the roof (labeled A3).

Finally, it is worth clarifying that the proposed model does not intend to be very accurate, but just a simple approximation to demonstrate that a noise field composed of a finite number of noise sources distributed non-uniformly is more realistic than an ideal diffuse noise field as well as it is employed to aid the optimization of the microphone array configuration.

3. Signal model and beamforming

Let us define $\mathbf{C} = [\mathbf{r}_1, \ldots, \mathbf{r}_m, \ldots, \mathbf{r}_M]$ as a microphone array composed of M elements, where $\mathbf{r}_m$ is a vector that represents the position of the m-th microphone with respect to the origin of coordinates.

3.1. Convolutive mixing model

Let us consider a set of M microphones that receive the signals coming from N different sources, $s_n(t)$, $n \in \{1, \ldots, N\}$, to generate M mixtures, $x_m(t)$, $m \in \{1, \ldots, M\}$, where t represents discrete time. The general expression for the additive mixing model is given by

$$x_m(t) = \sum_{n=1}^{N} s_n(t) * h_{mn}(t), \quad m = 1, \ldots, M, \tag{3}$$

where $h_{mn}(t)$ is the impulse response of a linear time-invariant (LTI) filter that describes the acoustic channel between the n-th source and the m-th microphone, and the operator $*$ represents linear convolution. The type of the mixing model depends on the assumptions made on the acoustic filter. In this work we assume an echoic mixing model to consider the reflections produced inside the vehicle, that is, the microphones receive several delayed and attenuated versions of the same source signal. The process is described by

$$x_m(t) = \sum_{n=1}^{N} \sum_{p=1}^{N_p} a_{mnp} \cdot s_n(t - \delta_{mnp}), \quad m = 1, \ldots, M, \tag{4}$$

where N_p is the number of different paths that the signals take from the sources to the microphones, and a_{mnp} and δ_{mnp} are the attenuations and delays introduced in the p-th path.

In this work, the echoic microphone responses are simulated using a room impulse response generator (RIRG). The acoustic impulse response between two points inside a defined room is calculated by using the simple image method described in Allen and Berkley (1979). This method considers the dimensions of the room (i.e. vehicle interior in this case), the reflection coefficient and the number of virtual sources to calculate the impulse response of the acoustic channel associated to each pair of source and microphone, both located in any point inside the vehicle. The original method has been completed to consider also the attenuation due to distance.

3.2. Directivity pattern and array gain

A wave source may be considered to come from the near-field of the array if

$$|r| < \frac{2L^2}{\lambda} \tag{5}$$

where $|r|$ is the distance between the source and the array, L is the effective length of the array, and λ is the wavelength. Considering the distances inside a vehicle and expression (5), we opt to assume near-field sources in the problem approached in this paper.

The directivity pattern or frequency response of an array $\mathbf{C}$ composed of M elements is given by

$$D(k, \mathbf{s}, \mathbf{C}) = \sum_{m=1}^{M} W_m(k) A_m(k, \mathbf{s}, \mathbf{C}), \tag{6}$$

where $\mathbf{s}$ is a vector that represents the position of the source with respect to the origin of coordinates, $W_m(k)$ are the complex weights applied to the m-th element of the array, and $A_m(k, \mathbf{s}, \mathbf{C})$ is the frequency response of the m-th element of the array $\mathbf{C}$ with respect to the source described by $\mathbf{s}$. For the sake of simplicity, the directivity pattern can be formulated in matrix notation according to

$$D(k, \mathbf{s}, \mathbf{C}) = \mathbf{w}_k^H \cdot \mathbf{a}_{k\mathbf{s}\mathbf{C}}, \tag{7}$$

where $\mathbf{w}_k = [W_1(k), \ldots, W_M(k)]^T$ is the array weight vector, $\mathbf{a}_{k\mathbf{s}\mathbf{C}}$ is a column vector containing the M microphone responses of the array $\mathbf{C}$, $\mathbf{a}_{k\mathbf{s}\mathbf{C}} = [A_1(k, \mathbf{s}, \mathbf{C}), \ldots, A_M(k, \mathbf{s}, \mathbf{C})]^T$, and $(\cdot)^H$ denotes the Hermitian transpose.

The array gain is defined as the improvement in signal-to-noise ratio (SNR) between a reference sensor and the array output, and it can be expressed as $G = G_d / G_n$, where G_d is the gain to the desired signal (i.e. the power of the directivity pattern for the steering direction) and G_n is the average gain to all noise sources, which depends on the nature of the noise field. Considering that the target source is located at $\mathbf{s}_0$, and that there are N finite noise sources with the same power interfering with the target source, the array gain is given by

$$G(k, \mathbf{s}_0, \mathbf{C}) = \frac{|D(k, \mathbf{s}_0, \mathbf{C})|^2}{\frac{1}{N} \sum_{n=1}^{N} |D(k, \mathbf{s}_n, \mathbf{C})|^2} = \frac{|\mathbf{w}_k^H \cdot \mathbf{a}_{k\mathbf{s}_0\mathbf{C}}|^2}{\mathbf{w}_k^H \cdot \mathbf{H}_\mathbf{C} \cdot \mathbf{w}_k}, \tag{8}$$

where $\mathbf{s}_n$ represents the position of the n-th noise source and $\mathbf{H}_\mathbf{C}$ is the noise cross-spectral matrix.

3.3. Robust superdirective beamformer

The minimum variance distortionless response (MVDR) filter due to Capon (1969) is perhaps the most widely used superdirective beamformer. The basic idea is to maximize the array gain (i.e. the output SNR) by finding the filter coefficients that minimize the output power with the constraint that the desired signal is not affected. This is equivalent to minimize the denominator in expression (8) with the constraint that the numerator has a constant value, which also guarantees a constant power at the output of the beamformer for the steering direction. The optimization problem is solved independently for each frequency band, and it is expressed according to

$$\min_{\mathbf{w}_k} \quad \{\mathbf{w}_k^H \cdot \mathbf{H}_\mathbf{C} \cdot \mathbf{w}_k\} \quad \text{subject to} \quad \mathbf{w}_k^H \cdot \mathbf{a}_{k\mathbf{s}_0\mathbf{C}} = 1 \tag{9}$$

The optimization problem is solved using Lagrange multipliers, resulting in

$$\mathbf{w}_k = \frac{\mathbf{H}_\mathbf{C}^{-1}\mathbf{a}_{k\mathbf{s}_0\mathbf{C}}}{\mathbf{a}_{k\mathbf{s}_0\mathbf{C}}^H\mathbf{H}_\mathbf{C}^{-1}\mathbf{a}_{k\mathbf{s}_0\mathbf{C}}}. \tag{10}$$

The previous solution may lead to undesirable amplification of incoherent noise due to electrical sensor noise, channel mismatch or errors in the microphone placement. In order to control the amplification of incoherent noise, an additional constraint is usually placed on the white noise gain. In the case of uncorrelated white noise, the noise cross-spectral matrix is equivalent to the identity matrix, i.e. $\mathbf{H}_\mathbf{C} = \mathbf{I}$, hence the white noise gain is given by

$$G_W(k,\mathbf{C}) = \frac{|\mathbf{w}_k^H \cdot \mathbf{a}_{k\mathbf{s}_0\mathbf{C}}|^2}{\mathbf{w}_k^H \cdot \mathbf{w}_k}. \tag{11}$$

The inverse of the white noise gain is called susceptibility, $K(k,\mathbf{C}) = 1/G_W(k,\mathbf{C})$, and is used as a measure of the sensitivity of the array with respect to uncorrelated errors. To keep the white noise gain above a lower limit is equivalent to maintain the susceptibility below an upper limit. Considering this additional constraint, the optimization problem in now defined as

$$\begin{aligned} &\min_{\mathbf{w}_k} \quad \{\mathbf{w}_k^H \cdot \mathbf{H}_\mathbf{C} \cdot \mathbf{w}_k\} \\ &\quad \text{subject to} \quad \mathbf{w}_k^H \cdot \mathbf{a}_{k\mathbf{s}_0\mathbf{C}} = 1 \\ &\quad\quad \text{and } K_{k\mathbf{C}} < K_{LIMIT}. \end{aligned} \tag{12}$$

Cox et al. (1987) propose that the optimum way of solving this problem is the addition of a small amount to each diagonal matrix element prior to inversion. The solution is given in Cox et al. (1987) as

$$\mathbf{w}_k = \frac{(\mathbf{H}_\mathbf{C}+\epsilon\mathbf{I})^{-1}\mathbf{a}_{k\mathbf{s}_0\mathbf{C}}}{\mathbf{a}_{k\mathbf{s}_0\mathbf{C}}^H(\mathbf{H}_\mathbf{C}+\epsilon\mathbf{I})^{-1}\mathbf{a}_{k\mathbf{s}_0\mathbf{C}}}, \tag{13}$$

where ϵ is a Lagrange multiplier that is iteratively adjusted until the susceptibility constraint is satisfied.

The solution implemented in this work is the next. Since the white noise gain monotonically increases with the ϵ value, the desired value of ϵ is found with an iterative technique employing an interval search algorithm between a minimum and a maximum value of ϵ. The center point of the interval is the ϵ value used to calculate the filter coefficients and to evaluate the susceptibility $K(k,\mathbf{C})$. If $K(k,\mathbf{C})$ is higher than the limit, the lower bound of the interval is set to the current ϵ. In the opposite case, the higher bound of the interval is the one set to the current ϵ. This procedure is repeated a fixed number of iterations. The value of K_{LIMIT} used in this paper is 1, which guarantees that white noise is never amplified.

Finally, the array output $Y(k,\mathbf{C})$ is expressed as the combination of the weighted input channels, according to

$$Y(k,\mathbf{C}) = \frac{1}{M}\sum_{m=1}^{M} W_m(k)X_m(k,\mathbf{C}), \tag{14}$$

where $X_m(k,\mathbf{C})$ is the discrete Fourier transform (DFT) of the input signal received by the m-th microphone of the array $\mathbf{C}$.

4. Proposed search algorithm

Once the beamforming technique has been defined, the second part in the design of the speech acquisition system is the selection of the microphone array configuration, which is the main objective in this paper. Since we are using a superdirective beamformer, it is assumed that the optimum microphone array is the one that maximizes the array gain. In this context, the objective is twofold: firstly, to find the best array geometry (i.e. microphone positions) and, secondly, to find a tradeoff between number of microphones and robustness in the speech acquisition.

Ideally, the microphones of the array can be placed in any point of the vehicle interior, which represents a large area in comparison to the microphone dimensions. Consequently, there exists a huge range of possible array configurations, which makes impossible to perform an exhaustive search. Heuristics optimization methods are useful to obtain a good approximation to the global optimum of a given function. Evolutionary algorithms (EAs), simulated annealing (SA), and tabu search (TS) are three popular iterative algorithms for heuristic optimization (Youssef et al., 2001). The three optimization methods have several similarities, and the selection of the most suitable technique depends on each specific problem. Since the optimization problem approached in this paper has a large search space, we opt for EAs because they perform a parallel search of the state space using a set of candidate solutions, against the point-by-point search performed by SA and TS techniques.

EAs are inspired in natural evolution laws, such as selection, mutation and crossover, to iteratively search for the optimum solution from the solutions obtained in previous iterations (Haupt and Haupt, 2004). The three main parts of an evolutionary algorithm are the generation of the candidate solutions of the population, the evaluation of a fitness function, and the evolution of the population (Alexandre et al., 2007). The candidate solutions are defined for each specific problem, and they are composed of a set of elements that may be binary or continuous values. The fitness function is defined as the cost function to optimize, and it also depends on the specific problem to solve. The evolution is based on the crossover and mutation operators whose characteristics can also be adapted to each specific problem.

In order to solve the search problem approached in this paper we have developed a tailored evolutionary algorithm, which purpose is to approximate the optimum array configuration in terms of a fitness function. This fitness function is defined as a speech-weighted array gain obtained by the robust superdirective beamformer described in Section 3.3. Each candidate solution of the population represents a different microphone array $\mathbf{C}$. The details of the proposed search algorithm are described in the remainder of this section.

4.1. Fitness function

The fitness function is defined as a frequency-weighted array gain, using higher weights in the frequency bands more important for speech understanding. The spectral weights are based on those used to calculate the Articulation Index (AI) (American National Standards Institute, 1997) for sentences. Let us recall the expression of the array gain in (8). The fitness function $G_A(\mathbf{C})$ depends on the array and is obtained according to

$$G_A(\mathbf{C}) = \frac{\sum_{k=1}^{K} W_{SP}(k)G(k,\mathbf{s}_0,\mathbf{C})}{\sum_{k=1}^{K} W_{SP}(k)}, \tag{15}$$

where $W_{SP}(k)$ are the speech weights calculated from the original weights defined in American National Standards Institute (1997) for the center frequencies of the critical bands interpolated to the center frequencies of the DFT.

4.2. Proposed evolutionary algorithm with multiple populations

The EA proposed in this work uses the fitness function defined in (15). Each candidate solution of the population represents a different microphone array $\mathbf{C}$, which is defined by M microphone positions. The size of the population is a crucial issue for the EA performance. On the one hand, a large population could cause more genetic diversity (and thus, a higher search space) and consequently suffer from slower convergence. On the other hand, with a very small population, only a reduced part of the search space could be explored, thus increasing the risk of prematurely converging to a local extreme. In this specific case we have found that a population of 100 candidate solutions is a good tradeoff between design time and performance. The initial population is randomly generated, and the successive populations are generated by crossover and mutations from the best candidates of the previous iteration. The number of iterations has been set to 100, which has been found to be a enough number to reach convergence.

With the aim to ensure convergence, the complete EA is executed 10 different times, using a different initial population in each case. Once the 10 runs are completed, the algorithm is executed an extra time, introducing the best 10 solutions of the previous runs in the initial population.

Unfortunately, the implementation of the proposed EA is computationally unaffordable, and we propose to discretize the search space in order to decrease the computation time. The impulse response between the microphones of each candidate solution and the 66 different sources (i.e. 63 noise sources, 2 loudspeakers and the target source) has to be computed with the RIRG in each iteration of the EA, which makes computationally unfeasible the execution of the proposed algorithm. In order to evaluate this problem, we have calculated the time required for the computation of the microphone responses of a single microphone with the 66 sources, using 100 different microphone positions. The average time is 0.1 s for a reflection coefficient of $r=0$ (best case), and 23.1 s for a reflection coefficient of $r=0.7$ (worst case of the evaluated), using a powerful computer composed of two 6-core processors at 2.93 GHz and 32 GB of RAM. Considering that each candidate solution contains an array composed of $N=8$ microphones (maximum evaluated), the population contains 100 candidate solutions, the number of iterations is 100, and the EA algorithm is executed a total of 11 times, the total number of microphones to evaluate is 880,000. The computation of the impulse responses for this number of microphones in the worst case ($r=0.7$) needs more than 235 days, which makes unfeasible the execution of the algorithm. In order to reduce the computational cost, we propose to discretize the search areas considering that the microphone positions can be only in a grid of 5 mm of resolution. The impulse responses between each point of the grid and each source can be pre-computed, hence they do not have to be computed again in each iteration, which notably decreases the computational cost of the algorithm. The dimension of the grid has been chosen to find a good tradeoff between the grid resolution and the computational cost and memory necessary to pre-calculate the microphone responses. With a grid of 5 mm, there are 1271 possible microphone positions in the area A1, 3321 in the area A2, and 5533 in the area A3, which means that a total of 10,125 microphone responses need to be calculated, which is only the 1% of the total number of microphone responses required in the case of not using the grid. Once the microphone responses are pre-calculated, the computation time of the remaining steps of the algorithm is relatively low.

The complete steps of the EA implemented in this work are the next:

1. The microphone responses between each point of the grid and each source for the different search areas considered are pre-calculated, using the RIRG.
2. An initial population of 100 candidate solutions is generated. Each solution selects M random positions from the grid.
3. The candidates of the population are validated to fulfill two restrictions:
   (a) The microphones of the candidate solutions must be in one of the positions defined in the grid. In the case that a microphone is outside the limited area, it is moved back to the edge, and, in the case that the microphone is within the limited area but it does not correspond with any of the positions of the grid, it is moved to the closer position in the grid. Note that only the initial population fulfill necessarily this condition.
   (b) The microphones of each candidate solution must keep a minimum distance of 35 mm between them. When two microphones are closer than this distance, they are randomly mutated along the two different directions until they fulfill the requirements.
4. The array gain obtained by the proposed beamformer is then computed for each candidate solution, and the fitness function $G_A(\mathbf{C})$ in expression (15) is calculated and used as ranking in order to determine the best solutions from the population.
5. After evaluating each candidate solution of the population, a selection process is then applied. It consists in selecting a subpopulation of 10% candidate individuals that best fit the fitness function. These elite individuals are those that will survive to the next generation, and the remaining solutions are removed.
6. Breed the new generation by recombining the parents by using a crossover operator. A 90% novel candidates for the next generation are generated by random crossover of the previously selected 10% best candidates. The crossover operator implemented in this EA is a uniform crossover (UX) operator with a crossover probability of 0.5: the offspring has approximately half of the elements from the first parent and the other half from the second parent, and these elements are randomly selected.
7. To mutate or randomly change the offspring. The main purpose of a mutation operator is to maintain diversity within the population and inhibit premature convergence to local extreme. Mutations are applied to the whole new population, and the mutation operator consists of adding a random Gaussian value to the position of each microphone. Only the best obtained solution is not mutated. The standard deviation of the random mutation factor is updated iteration by iteration using the performance of the EA. When the highest gain obtained in the current iteration outperforms the highest gain obtained in the previous iteration, the standard deviation is increased by 20%, otherwise, it is decreased by 10%, never allowing a value lower than 0.125 mm.
8. The process is repeated from steps 3 to 7 until 100 generations are evaluated. Since the best solution of each iteration is not modified, the best solution obtained in the last iteration is considered the best solution.
9. The complete EA from steps 2 to 8 is executed 10 different runs, each one evaluating a different initial population of candidate solutions.
10. Finally, the algorithm is evaluated an extra time, introducing the 10 best solutions obtained previously in the initial population. This last step increases the chances of convergence of the optimization algorithm.

The candidate solution that achieves the highest array gain in the last execution of the algorithm (step 10) is selected as the best approximation of the optimum solution.

Finally, the values of the parameters of the evolutionary algorithm (population size, crossover rate, mutation scheme and number of generations) have been found to obtain a quite good tradeoff between design time and performance for the experiments carried out in this paper.

5. Experimental work

5.1. Description of the experiments

The search algorithm proposed in this paper has been evaluated varying the number of microphones from 4 to 8, and limiting the position of the microphones in four different areas (see Fig. 2): A1 is a rectangle with dimensions $20\times15\ \text{cm}^2$ located in the rear-view mirror; A2 is a rectangle with dimensions $40\times20\ \text{cm}^2$ located in the central dashboard; A3 is defined by three rectangles, the first with dimensions $150\times5\ \text{cm}^2$ located in the upper front edge between the windshield and the roof, and the other two of dimensions $50\times5\ \text{cm}^2$ located at the edge between both front lateral windows and the roof (corners); finally, the fourth area contains all the previous areas, i.e. A1+A2+A3. Furthermore, the search algorithm has been executed varying the reflection coefficient of the car model from 0.0 to 0.7 in steps of 0.1. Consequently, a total of $5\times4\times8$ optimized solutions have been computed, regarding to the five different number of microphones, four different search areas and eight different reflection coefficients. All the experiments have been carried out with a sampling rate of 8 kHz using a 64-point DFT.

In order to assess the validity of the proposed search algorithm, the performance of the optimized microphones arrays is compared to the one obtained by linear arrays placed in the same areas and composed of the same number of microphones. In the case of A1, A2 and A3, we have used uniform linear arrays of M elements placed along the horizontal axis, centered in the corresponding areas, with a space between microphones of 2.5 cm in the first case, and 3.5 cm in the other two cases (3.5 cm are not used in the first case because only 6 microphones would fit in that area). In the case of A4, the microphones are split in the three different areas: 2 microphones separated 3.5 cm in the center of A3, 1 microphone in the center of A1 and another microphone in the center of A2, for $M=4$; a microphone is included in A2 separated 3.5 cm in the horizontal line from the previous one, for $M=5$; another microphone is included in A1 separated 3.5 cm in the horizontal line from the existing one, for $M=6$; a microphone is added in A3 separated 3.5 cm in the horizontal line from the two previous one, for $M=7$; and a final microphone is included in A2 separated 3.5 cm in the horizontal line from the previous two, for $M=8$.

The performance of the optimized solutions and the linear arrays are compared in terms of the frequency-weighted array gain G_A defined in expression (15), which represents the array gain in the speech frequency band, and the PESQ score (ITU-T, 2001), which is a speech quality measure correlated with intelligibility (Hu and Loizou, 2008). To calculate the PESQ score obtained by the different arrays, we have simulated a realistic scenario with the next setup: a speech source is placed in the position of the head of the driver (target source); a different speech source with a power of 3 dB lower than the target source is located in the position of both front loudspeakers; and spatially uncorrelated white noise is located in each of the 63 noise sources with a power in such way that the addition of all of the noise sources results in a SNR of 10 dB with respect to the target source (if we assume that the target source has a power of 60 dB, which is a normal level for conversation, the background noise would have a power of 50 dB). The PESQ value is evaluated 100 times for each array, using different speech sources in both the target and the loudspeakers. The speech sources are randomly selected from the TIMIT database (Fisher et al., 1986).

The results obtained are analyzed in the next subsection.

5.2. Results

Table 1 contains the frequency-weighted array gain (G_A), in dB, and the PESQ score obtained by the optimized arrays (OPT) and by the linear arrays (LIN) in the area A1, for the different number of microphones (M from 4 to 8) and for reflection coefficient values of $r=0.0$, $r=0.3$, and $r=0.7$. The same results are contained in Table 2 for the area A2, in Table 3 for the area A3, and in Table 4 for the combination of the three areas A1+A2+A3. Analyzing the values of the array gain, a nearly linear and positive dependence with the number of microphones M is deduced, for all cases. As the number of microphones increases, G_A always increases. On the contrary, there is a nearly linear but negative dependence of G_A with the reflection coefficient r: G_A always decreases when r increases. Although there is not a linear relationship between G_A and the PESQ score, the increment of the first implies the increment of the second. Hence, the behavior of the PESQ score is the same than as the behavior of G_A, regarding M, r and the search areas.

Table 2
Frequency-weighted array gain (G_A), in dB, and PESQ score obtained by the optimized arrays (OPT) and by the linear arrays (LIN) in the area A2, for the different number of microphones and for reflection coefficient values of 0.0, 0.3, and 0.7.

M	**A2**											
	OPT						LIN					
	$r=0.0$		$r=0.3$		$r=0.7$		$r=0.0$		$r=0.3$		$r=0.7$	
	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ
4	11.5	2.92	9.1	2.78	6.1	2.55	8.4	2.65	5.9	2.47	2.7	2.25
5	12.4	3.00	10.0	2.79	6.9	2.63	8.8	2.69	6.4	2.51	3.2	2.30
6	13.1	3.04	10.7	2.89	7.6	2.65	9.1	2.75	6.8	2.56	3.9	2.34
7	13.5	3.08	11.1	2.89	8.1	2.71	9.5	2.78	7.2	2.60	4.4	2.37
8	13.8	3.10	11.5	2.95	8.6	2.74	9.9	2.80	7.6	2.62	4.8	2.40

Table 1
Frequency-weighted array gain (G_A), in dB, and PESQ score obtained by the optimized arrays (OPT) and by the linear arrays (LIN) in the area A1, for the different number of microphones and for reflection coefficient values of 0.0, 0.3, and 0.7.

M	**A1**											
	OPT						LIN					
	$r=0.0$		$r=0.3$		$r=0.7$		$r=0.0$		$r=0.3$		$r=0.7$	
	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ
4	9.6	2.77	7.2	2.61	4.5	2.46	5.6	2.57	3.5	2.42	2.1	2.31
5	10.8	2.90	8.1	2.70	5.6	2.51	6.1	2.60	4.0	2.46	2.7	2.36
6	11.5	2.93	9.0	2.75	6.4	2.56	6.5	2.63	4.4	2.49	3.1	2.38
7	12.0	2.97	9.5	2.79	7.1	2.58	6.8	2.66	4.7	2.52	3.4	2.40
8	12.7	3.01	10.1	2.82	7.6	2.66	7.1	2.68	5.0	2.54	3.8	2.42

Table 3
Frequency-weighted array gain (G_A), in dB, and PESQ score obtained by the optimized arrays (OPT) and by the linear arrays (LIN) in the area A3, for the different number of microphones and for reflection coefficient values of 0.0, 0.3, and 0.7.

M	**A3**											
	OPT						LIN					
	$r=0.0$		$r=0.3$		$r=0.7$		$r=0.0$		$r=0.3$		$r=0.7$	
	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ
4	11.8	2.99	10.3	2.92	7.3	2.70	2.5	2.39	3.2	2.44	3.3	2.41
5	12.5	2.99	11.3	3.00	8.2	2.75	3.1	2.43	3.9	2.49	4.0	2.46
6	13.9	3.03	11.9	2.99	9.1	2.79	3.5	2.46	4.4	2.52	4.5	2.49
7	14.4	3.12	12.5	3.06	9.6	2.84	3.9	2.49	4.9	2.55	5.0	2.52
8	15.0	3.19	13.1	3.09	10.4	2.88	4.3	2.51	5.4	2.58	5.5	2.56

Table 4
Frequency-weighted array gain (G_A), in dB, and PESQ score obtained by the optimized arrays (OPT) and by the linear arrays (LIN) in the area A1+A2+A3, for the different number of microphones and for reflection coefficient values of 0.0, 0.3, and 0.7.

M	**A1+A2+A3**											
	OPT						LIN					
	r=0.0		r=0.3		r=0.7		r=0.0		r=0.3		r=0.7	
	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ	G_A	PESQ
4	12.6	3.03	9.8	2.87	7.3	2.68	8.2	2.71	6.4	2.61	4.2	2.43
5	12.9	3.04	11.0	2.94	8.3	2.73	9.5	2.81	7.5	2.70	5.2	2.52
6	13.6	3.11	11.6	2.96	9.2	2.79	10.3	2.86	8.3	2.75	6.0	2.58
7	15.1	3.17	12.5	3.04	9.9	2.86	10.5	2.88	8.7	2.78	6.7	2.64
8	15.7	3.22	13.7	3.15	10.8	2.93	11.0	2.90	9.0	2.81	7.0	2.67

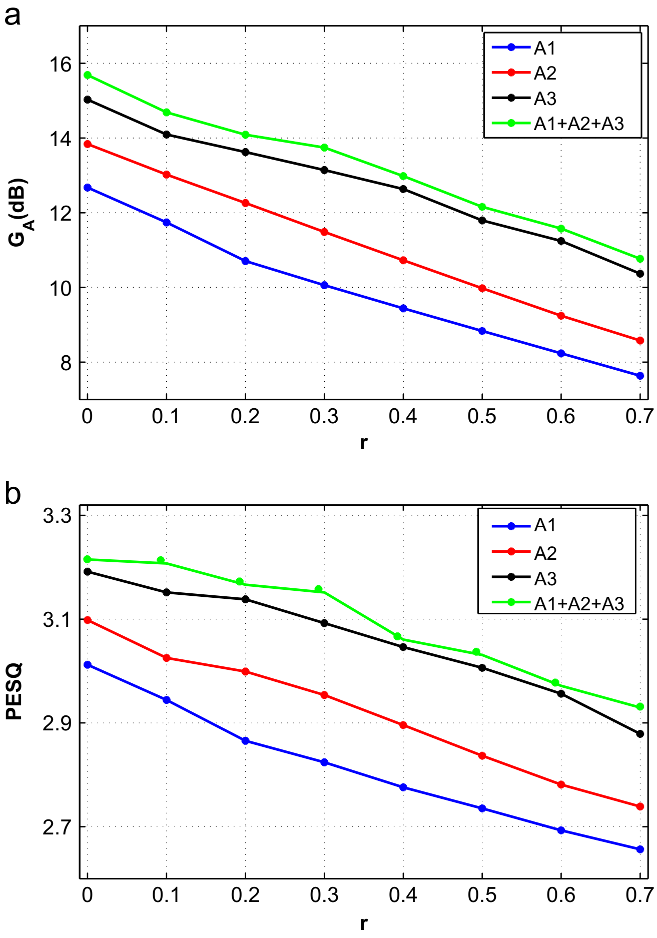

Fig. 3. Frequency-weighted array gain G_A (a) and PESQ score (b) as a function of the reflection coefficient r, for the different search areas (A1, A2, A3 and A1+A2+A3) and in the case of 8 microphones. The reflection coefficient r has been varied from 0 to 0.7 in steps of 0.1.

Comparing the three different areas, the optimized arrays in area A3 obtained the highest G_A and PESQ scores, followed by the optimized arrays in area A2, which perform better than those in area A1. Nevertheless, the results are improved when the arrays are optimized using the combination of the three areas, A1+A2+A3 (Table 4). Regarding the linear arrays, both the G_A and PESQ scores obtained by the optimized solutions are notably higher than the ones obtained by the linear arrays, for any number of microphones, reflection coefficient and search area.

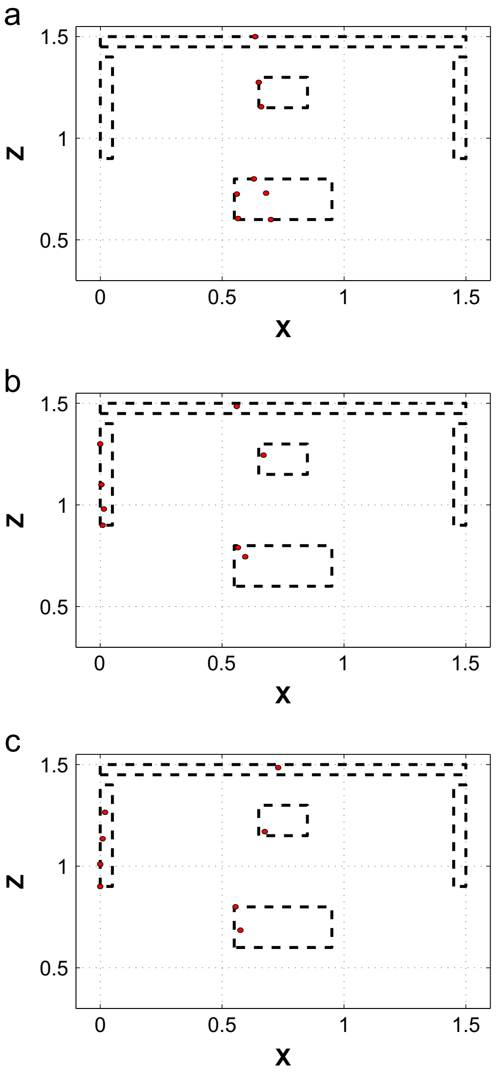

Fig. 4. Optimized microphone arrays in area A1+A2+3 for the 8-microphones case with reflection coefficient values of r=0 in (a), r=0.3 in (b) and r=0.7 in (c). The two vertical rectangles represented at both sides are those placed along the '*Y*'-axis (edge between lateral windows and the roof).

In order to evaluate the impact of reverberation, Fig. 3 represents the values of G_A in (a) and the PESQ scores in (b) as a function of the reflection coefficient r, for the different search areas (A1, A2, A3 and A1+A2+A3) and in the case of 8 microphones. The reflection coefficient r has been varied from 0 to 0.7 in steps of 0.1. Both the array gain G_A and the PESQ score clearly decrease when the reflection coefficient increases, with a relationship almost linear. Although reverberation clearly affects the performance of the four different search areas, the area A3 is less affected than area A1 and A2, and the combination of the three areas, A1+A2+A3, is a little less affected than the area A3.

Finally, and by way of illustration, Fig. 4 represents the positions of the microphones of the optimized arrays obtained by the proposed algorithm in the best case, which is the 8-microphones case in the area A1+A2+A3. The two vertical rectangles represented at both sides are those placed along the '*Y*'-axis (edge between lateral windows and the roof). The optimized arrays are shown for reflection coefficients of r=0.0, r=0.3 and r=0.7, which have obtained array gains of G_A=15.7, G_A=13.7 and G_A=10.8 respectively. From this figure we can deduce that the optimized arrays contain microphones among the three different areas as well as that microphones placed in the edge between the left lateral window and the roof have significative importance removing reverberations.

Finally, after analyzing different optimized arrays we can deduce that the solutions have irregular configurations, as those shown in Fig. 4. This fact has the next explanation. The use of symmetric arrays involves that the sidelobes are located at fixed and symmetric positions, so they are more appropriate to reject directional noise rather than isotropic noise. However, the use of asymmetric arrays originates an isometric distribution of the sidelobes, resulting in a more uniform attenuation for all directions. Actually, the use of logarithmic spiral arrays have been used in several applications of acoustic arrays to obtain high directivity, for instance, in Kitagawa and Thompson (2006) and Humphreys et al. (1998).

6. Conclusions

Speech acquisition systems typically include spatial filtering to enhance the quality of the acquired speech. The design of the spatial filter is not the only important factor affecting the performance, but the microphone array configuration also plays an important role. In this paper we have proposed a novel search algorithm based on evolutionary computation to approximate the optimum microphone configuration for speech acquisition inside a vehicle. The optimization criterion is the maximization of a speech-weighted array gain obtained by a robust superdirective beamformer. The results obtained reveal that the optimized arrays obtain high array gains and output speech quality, notably outperforming the results obtained by linear arrays with the same number of elements, which supports the validity of the proposed algorithm.

Although the feasibility of the algorithm has been demonstrated in the case of speech acquisition inside a car, the method is easily extensible for any other application. However, in order to generalize the results, the optimized microphone arrays should be tested with real measurements inside a vehicle. In such a case, the decrement in the performance due to movements of the head of the speaker could also be evaluated. Finally, the simple car model proposed in this paper can be improved to be more realistic using different reflection coefficients for different materials and introducing additional objects inside the car, such as seats.

Acknowledgments

This work has been funded by the Spanish Ministry of Economy and Competitiveness, under Project TEC2012-38142-C04-02 and the Scholarship AP2009-3932.

References

Alexandre, E., Cuadra, L., Rosa, M., Lopez-Ferreras, F., 2007. Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms. IEEE Trans. Audio Speech Lang. Process. 15 (8), 2249–2256.

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65 (4), 943–950.

American National Standards Institute, 1997. American National Standard: Methods for Calculation of the Speech Intelligibility Index. Acoustical Society of America. S3.5. New York.

Ayllón, D., Benito-Olivares, V., Llerena-Aguilar, C., Gil-Pita, R., Rosa-Zurera, M., 2012. Three-dimensional microphone array for speech enhancement in hands-free systems for cars. In: Proceedings of the 45th International Conference on Audio Engineering Society.

Benesty, J., Chen, J., Huang, Y., 2008. Microphone Array Signal Processing, first ed. Springer, Berlin.

Bitzer, J., Kammeyer, K.D., Simmer, K., 1999. An alternative implementation of the superdirective beamformer. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 7–10.

Busso, C., Hernanz, S., Chu, C.W., Kwon, S., Lee, S., Georgiou, P.G., Cohen, I., Narayanan, S., 2005. Smart room: participant and speaker localization and identification. In: Proceedings of the IEEE International Conference Acoustics, Speech, and Signal Processing, vol. 2, pp. 1117–1120.

Capon, J., 1969. High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE 57 (8), 1408–1418.

Chabuk, T., Reggia, J.A., Lohn, J., Linden, D., 2012. Causally-guided evolutionary optimization and its application to antenna array design. Integr. Comput. Aided Eng. 19 (2), 111–124.

Cho, J., Krishnamurthy, A., 2003. Speech enhancement using microphone array in moving vehicle environment. In: Proceedings of the IEEE Intelligent Vehicles Symposium, pp. 366–371.

Cox, H., Zeskind, R., Owen, M., 1987. Robust adaptive beamforming. IEEE Trans. Acoust. Speech. Signal Process. 35, 1365–1376.

Elko, W.G., 1996. Microphone array systems for hands-free telecommunication. Speech Commun. 20 (3–4), 229–240.

Feng, Z.G., Yiu, K.F.C., Nordholm, S.E., 2012. Placement design of microphone arrays in near-field broadband beamformers. IEEE Trans. Signal Process. 60 (3), 1195–1204.

Fisher, W., Doddington, G., Marshall, K., 1986. The DARPA speech recognition research database: specification and status. In: Proceedings of the DARPA Speech Recognition Workshop, pp. 93–99.

Goldman, G., 2009. Computationally Efficient Algorithms for Estimating the Angle of Arrival of Helicopters Using Acoustic Arrays. U.S. Army research laboratory, Maryland.

Grenier, Y., 1992. A microphone array for car environments. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 305–308.

Hamacher, V., Chalupper, J., Eggers, J., Fischer, E., Kornagel, U., Puder, H., Rass, U., 2005. Signal processing in high-end hearing aids: state of the art, challenges, and future trends. EURASIP J. Appl. Signal Process. 2005 (18), 2915–2929.

Hansen, W.W., Woodyard, J.R., 1938. A new principle in directional antenna design. Proc. Inst. Radio Eng. 26 (3), 333–345.

Haupt, R.L., Haupt, S.E., 2004. Practical Genetic Algorithms, second ed. John Wiley and Sons, New Jersey.

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 1 (1), 229–238.

Humphreys, W.M., Hunter, W.W., Meadows, K.R., Brooks, T.F., 1998. Design and use of microphone directional arrays for aeroacoustic measurements. In: The 36th Aerospace Sciences Meeting & Exhibition, 98-0471.

Recommendation P ITU-T, 2001. 862-perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Int. Telecommun. Union—Telecommun. Stand. Sec. (ITU-T).

Kitagawa, T., Thompson, D.J., 2006. Comparison of wheel/rail noise radiation on Japanese railways using the TWINS model and microphone array measurements. J. Sound Vib. 293 (3), 496–509.

Kwok, N.M., Liu, D.K., Dissanayake, G., 2006. Evolutionary computing based mobile robot localization. Eng. Appl. Artif. Intell. 19 (8), 857–868.

Martin, R., Petrovsky, A., Lotter, T., 2001. Planar superdirective microphone arrays for speech acquisition in the car. In: The Seventh European Conference on Speech Communication and Technology.

Oh, S., Viswanathan, V., Papamichalis, P., 1992. Hands-free voice communication in an automobile with a microphone array. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 281–284.

Okandan, M., Parker, E.P., Hall, N.A., Peterson, K., Resick, P., Serkland, D., 2007. Ultrasensitive Directional Microphone Arrays for Military Operations in Urban Terrain. Sandia National Laboratories, New Mexico.

Ruiz-Reyes, N., Vera-Candeas, P., García-Galán, S., Munoz, J.E., 2010. Two-stage cascaded classification approach based on genetic fuzzy learning for speech/music discrimination. Eng. Appl. Artif. Intell. 23 (2), 151–159.

Selouani, S.A., Alotaibi, Y.A., 2013. Adaptation to non-native speech using evolutionary-based discriminative linear transforms. Eng. Appl. Artif. Intell. 26 (2), 899–904.

White, L.B., Boashash, B., 1990. Cross spectral analysis of non-stationary processes. IEEE Trans. Inf. Theory. 36 (4), 830–835.

Youssef, H., Sait, M.S., Adiche, H., 2001. Evolutionary algorithms, simulated annealing and tabu search: a comparative study. Eng. Appl. Artif. Intell. 14 (2), 167–181.