EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Rate-constrained source separation for speech enhancement in wireless-communicated binaural hearing aids

David Ayllón*, Roberto Gil-Pita and Manuel Rosa-Zurera

Abstract

A recent trend in hearing aids is the connection of the left and right devices to collaborate between them. Binaural systems can provide natural binaural hearing and support the improvement of speech intelligibility in noise, but they require data transmission between both devices, which increases the power consumption. This paper presents a novel sound source separation algorithm for binaural speech enhancement based on supervised machine learning and time-frequency masking. The system is designed considering the power restrictions in hearing aids, constraining both the computational cost of the algorithm and the transmission bit rate. The transmission schema is optimized using a tailored evolutionary algorithm that assigns a different number of bits to each frequency band. The proposed algorithm requires less than 10% of the available computational resources for signal processing and obtains good separation performance using bit rates lower than 64 kbps.

Keywords: Speech enhancement; Source separation; Hearing aids

1 Introduction

Most people suffering from impaired hearing and wearing hearings aids show a lack of intelligibility when they are in a noisy environment. Modern devices include some mechanisms to increment the hearing comfort of the user, including advanced features such as acoustic feedback cancellation [1,2], automatic environment classification [3,4], or speech enhancement [5,6]. One of the most challenging problems found in the design of hearing aids is the reduction of undesired noise and interference signals to increase speech intelligibility without introducing audible distortions in the target speech. The implementation of signal processing algorithms in hearing aids presents additional challenges: the reduced battery life, which limits the computational capability of the device, the requirement of real-time processing, which limits the processing delay to few milliseconds and reduces the number of frequency bands used for the analysis, and the small size of the device, which limits the number of assembled microphones.

*Correspondence: david.ayllon@uah.es
Department of Signal Theory and Communications, University of Alcala, Alcalá de Henares, Spain

A common approach to remove undesired sound sources is to provide the device with directivity, assuming that the undesired sources and the target source are spatially separated. Directional microphones have been amply included in hearing aids for over 25 years and have proved to significantly increase speech intelligibility in various noisy environments [7]. However, they are usually not applicable to small ear canal devices for reasons of size, the higher internal noise they have compared to omnidirectional microphones, and their fixed directivity pattern which does not allow adapting the directivity to changing acoustic environments [8]. In the last years, microphone arrays composed of omnidirectional microphones have drawn the attention of hearing aid designers [9,10]. The use of multiple channels allows implementing speech enhancement algorithms based on spatial filtering (beamforming) and source separation. Both fixed and adaptive beamforming techniques have been successfully implemented in modern hearing aids [11-14], due to their reduced complexity in comparison to traditional multichannel source separation algorithms based on ICA or clustering. Originated in the computational auditory scene analysis (CASA) [15], the time-frequency masking approach for source separation is a potential solution for

speech enhancement in hearing aids [16], as long as the estimation of the time-frequency mask involves low computational complexity. The ideal binary mask (IBM) is defined in [17] as the one that takes values of zero or one by comparing the local signal-to-noise ratio (SNR) in each time-frequency bin against a threshold, which is typically chosen to be 0 dB. Several studies [18-20] have demonstrated that the application of the IBM to separate speech in noisy conditions entails an improvement in speech intelligibility. Unfortunately, the computation of the IBM needs to have access to the target speech source and noise signals, information that is not available in practice. Hence, the IBM should be estimated somehow from the corrupted signal, obtaining a binary mask that is just an approximation of the IBM.

Many hearing-impaired people have bilateral hearing loss and they are forced to wear two devices. When hearing aids are worn at both ears, these devices usually operate independently. However, there is a new trend of binaural hearing aids that connects both devices in order to exchange information between them. Binaural hearing provides considerable benefits over using a single ear, due to the ability to preserve spatial cues, which are necessary to localize and separate sounds. Unfortunately, the communication between both hearing devices should be implemented with a wireless link, due to aesthetic reasons, which unavoidably increases the power consumption and, consequently, reduces the battery life. This fact opens a new problem: how to reduce the bit rate transmitted between both devices without decreasing the performance of the speech enhancement algorithm.

In the recent years, several works have proposed microphone array-based binaural spatial filtering techniques, using both fixed beamformers [21,22] and adaptive beamformers [23,24]. The work in [25] analyzes the robustness of binaural fixed and adaptive beamformers by means of objective perceptual quality measures. In [26], three different strategies are proposed, two of them are based on the estimation of the short-time spectral amplitude of the original signal, and the third one is based on spatial filtering. The aforementioned proposed solutions have demonstrated their ability to reduce noise and to improve speech quality. However, they assume that the original signals received at the right and left devices are available at both sides, which involves a high bandwidth communication. In practice, the signals are not completely transmitted, and the transmission rate (and the power consumption) depends on the amount of exchanged information. This problem is approached in [27], which evaluates the array gain provided by collaborating hearing aids as a function of the communication rate, using an information theoretic approach. In [28], the authors evaluate the decrement of noise reduction achieved by a binaural multichannel Wiener filter when reducing the bandwidth of the transmission link. The work in [29] proposes two approaches to reduce data transmission: the first approach is to transmit only an estimation of the undesired signal at a determined bit rate and the second approach is to transmit the complete received signal at the determined bit rate. Unfortunately, the performance of the algorithms in [27,29] is notably reduced when the transmission rate decreases (e.g. lower than 16 kbps). An additional problem associated to the use of beamforming techniques for wireless-communicated binaural hearing is the following. The output of the beamformer is obtained by combining a weighted version of the input channels from both devices. If one or several speech signals have been quantized and transmitted to the other device, the beamforming output is directly affected by quantization noise.

The goal of this paper is the design of an energy-efficient speech enhancement algorithm with low computational cost for wireless-communicated binaural hearing aids. The binaural speech enhancement problem is approached from a different perspective, using time-frequency masking rather than spatial filtering. In this context, there are two problems to solve. First, a low-cost speech enhancement algorithm that uses binaural information is designed. The proposed algorithm estimates the IBM, which has been proven to correlate with intelligibility [18-20], using a generalized version of the least-squares linear discriminant analysis (LS-LDA) [30]. The classifier uses a set of features extracted from the short-time Fourier transform (STFT) of the signals received at both ears, assuming that all information has been exchanged between both devices. The second problem to solve is the reduction of the amount of information exchanged between both devices minimizing the effects on the performance obtained by the speech enhancement algorithm. The signal of one of the devices is quantized before being transmitted to the other device, which calculates the binary mask. The quantization of each frequency band can be performed with a different number of bits. An optimization algorithm based on evolutionary computation is proposed to distribute a limited number of bits among the different frequency bands, allowing to assign a value of 0 bits, which avoid transmitting unnecessary information. In the proposed schema, the transmitted signals are only used to estimate the mask, and quantization noise does not directly affect the quality of the output speech signal, although it may affect the mask estimation.

Our previous work in [31] addressed the same problem described in this paper, designing a low-cost speech separation system based on the computation of the time-frequency binary mask that maximizes the W-disjoint orthogonality (WDO) factor and increases the energy efficiency of the wireless-communicated binaural hearing aids. There are 3 main differences of that work with the

one described in this paper. First, the goal of the current design is to obtain a system that estimates the IBM rather than maximizes the WDO. Second, unlike the previous work that only considered the time and level differences between both ears as input features, this work proposes and studies a different combination of features, with the novelty that they are calculated not only from the current time-frequency point but also from the neighbor time-frequency points. And third, the algorithm proposed in this paper optimizes the weights of the classifier and the bit distribution at the same time, and for all frequencies at once.

2 Time-frequency masking source separation

Speech signals are sparse in the time-frequency domain, that is, most of the sample values of a signal are zero or close to zero in this domain. This property is very useful for speech source separation due to the fact that the probability of two or more sources being simultaneously active is low in a sparse representation. Two signals are considered to be W-disjoint orthogonal (WDO) if their STFT representations do not overlap [32]. If this property is strictly met, the original signals can be perfectly demixed by identifying the active time-frequency regions of each source, which leads to a time-frequency binary mask. Usually, speech signals only show an approximate WDO behavior in the sense that the probability of two sources having high energy in the same time-frequency point is low [33]. This fact allows separating sources by time-frequency masking with a good performance.

2.1 The ideal binary mask (IBM)

Let us consider the STFT of a mixture signal $X(k,m) = S(k,m) + N(k,m)$, where $S(k,m)$ is the target speech source, $N(k,m)$ contains all the undesired sources (noise), k represents the frequency index, $k = 1,\cdots,K$, and m the time frame index, $m = 1,\cdots,M$. The 0-dB ideal binary mask (IBM) is then defined as [34]

$$\mathrm{IBM}(k,m) := \begin{cases} 1, & |S(k,m)|^2 > |N(k,m)|^2 \\ 0, & \text{otherwise} \end{cases}, \qquad (1)$$

where the time-frequency bins are associated to the source that has more energy than its interfering sources. It has been demonstrated in [18-20] that the application of the IBM to separate speech from noise entails an improvement in the intelligibility of the target speech signal. Unfortunately, the clean and noise signals are not available in practice, and the IBM must be estimated from the mixtures, which decreases the performance. The study in [19] evaluates the impact of the IBM estimation errors in the intelligibility of the separated signals.

2.2 The W-disjoint orthogonality (WDO) quality factor

It is established in [35] that the performance of a given time-frequency mask depends on two criteria: the amount of preserved target source and the amount of suppressed interfering sources. These two conditions are measured by the preserved-to-signal ratio (PSR) and the signal-to-interference ratio (SIR), respectively. The PSR indicates the amount of the energy of the target source preserved by the mask after separation, and it is calculated as

$$\mathrm{PSR} = \frac{\|M(k,m)\cdot S(k,m)\|^2}{\|S(k,m)\|^2}, \qquad (2)$$

where $M(k,m)$ is the time-frequency mask computed for the separation of the target source $S(k,m)$. If the sources were strictly WDO, the IBM mask defined in (1) would preserve all the energy of the desired signal, obtaining the maximum value PSR = 1.

On the other hand, the SIR measures the amount of energy from the interfering sources suppressed by the mask, and it is given by

$$\mathrm{SIR} = \frac{\|M(k,m)\cdot S(k,m)\|^2}{\|M(k,m)\cdot N(k,m)\|^2}. \qquad (3)$$

If the sources were strictly WDO, the IBM mask in (1) would completely suppress the energy of the interfering signal, and then the SIR would be infinite (SIR = ∞). Both the PSR and the SIR are combined into the WDO factor using the following expression:

$$\begin{aligned} \mathrm{WDO} &= \frac{\|M(k,m)\cdot S(k,m)\|^2 - \|M(k,m)\cdot N(k,m)\|^2}{\|S(k,m)\|^2} \\ &= \mathrm{PSR} - \frac{\mathrm{PSR}}{\mathrm{SIR}}. \end{aligned} \qquad (4)$$

It is clear that WDO sources perfectly separated with the IBM mask defined in (1) have a value of WDO = 1, which is the maximum value. However, this value is only achievable by perfect WDO sources and it obviously decreases (i.e. $WDO \leq 1$) with approximately WDO sources, due to the fact that a small part of the source signals overlap, which implies that the mask is not able neither to preserve all the energy of the desired signal nor to reject all the energy of the interfering signals. Therefore, the WDO factor is a good indicator of the quality of the separation achieved by a time-frequency mask for approximately WDO sources.

3 Proposed binaural speech enhancement system

3.1 System overview

The assumed acoustic scenario and the binaural speech enhancement system proposed in this paper are represented in Figure 1. In this scenario, the hearing aid user

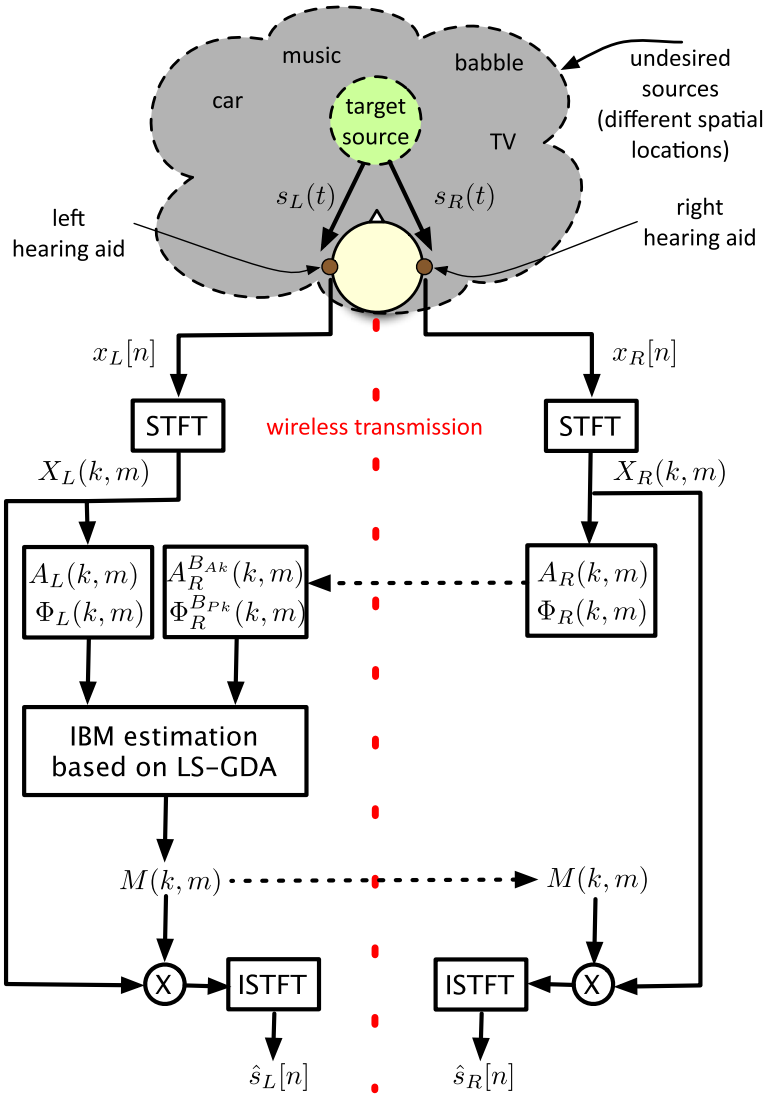

Figure 1 Binaural speech enhancement system. Overview of the wireless-communicated binaural speech enhancement system proposed in this paper.

who wears two hearing aids wants to understand the speech produced by an interlocutor. Assuming that the user is looking at the desired interlocutor, the sound arriving at both devices is a mixture of the desired source coming from the straight-ahead direction (the green circle) and a combination of undesired sound sources coming from other directions (the gray cloud). The origin of the undesired sources may vary: different speakers, babble, music, traffic noise, TV, etc. Hence, the signals entering the left and the right hearing aids, $x_L[n]$ and $x_R[n]$, can be expressed as

$$\begin{aligned} x_L[n] &= s_L[n] + n_L[n] \\ x_R[n] &= s_R[n] + n_R[n], \end{aligned} \tag{5}$$

where $s_L[n]$ and $s_R[n]$ are the target signals arriving at the left and right hearing aids, respectively, and $n_L[n]$ and $n_R[n]$ represent the combination of undesired sources (noises coming from other directions) entering the left and right hearing aids, respectively. The filterbank of each device computes the STFT of each frame, obtaining $X_L(k,m)$ and $X_R(k,m)$ for the left and right ears, respectively. The amplitude (in dB) of the STFT is represented by $A_L(k,m)$ and $A_R(k,m)$ for the left and right hearing aids, respectively, and it is calculated according to

$$\begin{aligned} A_L(k,m) &= 20\log_{10}|X_L(k,m)| \\ A_R(k,m) &= 20\log_{10}|X_R(k,m)|. \end{aligned} \tag{6}$$

We use the logarithmic transformation of the squared amplitude because it provides more meaningful information from the human hearing point of view. The phase of the STFT is represented by $\phi_L(k,m)$ and $\phi_R(k,m)$ for the left and right hearing aids, respectively.

The speech enhancement system is based on the estimation of the IBM defined in (1) from the two binaural mixtures. The IBM is not necessarily the same for the left and the right devices. However, in order to preserve the binaural cues, we assume that the same mask is applied in the right and the left devices. The IBM is calculated here using the energy of the signals of both devices. The mask is calculated only in one of the devices and transmitted to the other one, thus reducing the computational load in one of the devices. In the schema shown in Figure 1, the right device transmits the amplitude and phase of the STFT of its received signal, $A_R(k,m)$ and $\phi_R(k,m)$, to the left device, which calculates the binary mask $M(k,m)$ and transmits it to the right device. Once both devices have the mask, they apply it to the STFT of their received signals and compute the inverse STFT (ISTFT) to obtain a clean version of the original target source, which is directly played in the loudspeaker of the hearing device. The number of bits transmitted can be reduced by transmitting a quantized low-bit version of $A_R(k,m)$ and $\phi_R(k,m)$, instead of their values themselves. The transmitted quantized version of $A_R(k,m)$ and $\phi_R(k,m)$ are labeled as $A_R^{B_{Ak}}(k,m)$ and $\phi_R^{B_{Pk}}(k,m)$, where B_{Ak} and B_{Pk} are the number of bits used to quantize the kth frequency band of the amplitude and phase, respectively. The quantized values from the right device and those directly computed by the left device, $A_L(k,m)$ and $\phi_L(k,m)$, are used by the left device to calculate the binary mask $M(k,m)$. Due to the fact that the binary mask, which is transmitted from the left to the right device, only contains values of 0 and 1, it is coded with only 1 bit, hence K bits are transmitted for each frame. It is worth to mention here that the use of a soft mask may improve the performance of the IBM, but the transmission of continuous values would imply an increment of the transmission rate. The key point of the system proposed in this paper is that the values $A_R(k,m)$ and $\phi_R(k,m)$ of each frequency band are quantized with a different number of bits B_{Ak} and B_{Pk}, limiting the total number of bits transmitted for each frame. The assignation of the number of bits to the different frequency bands is carried out by optimizing the performance of the speech enhancement system, avoiding to transmit unnecessary information.

The proposed transmission schema only makes sense when the latency of the system allows a delay higher than the transmission time plus the processing time. The system can also be implemented symmetrically, for instance, transmitting the information of half of the frequency bands from the left to the right device and the other half from the right to the left device. In this case, each device calculates half of the mask and transmits it to the other device. For the sake of simplicity, the schema in Figure 1 is adopted in this paper, considering that the proposed algorithms are also valid for the symmetric schema. Additionally, it is worth clarifying that the data transmission is not continuous: first, the amplitude and phase information is transmitted from the right to the left device, and after the processing time, the mask is transmitted from the left to the right device. This fact allows transmitting at the maximum bit rate available in the device (around 300 kbps in commercial devices) but only during a part of the processing time of each frame.

Finally, it is worth mentioning that all the design methods described in this paper are carried out offline on a computer. Only when the design has been completed, the optimum solution is then implemented on the digital hearing aid.

3.2 Estimation of the IBM with a least squares generalized discriminant analysis (LS-GDA)

The computational cost associated to the estimation of the IBM must be relatively low, according to the low computational power available in hearing aids. In this work, we propose the use of a low-cost classifier to decide whether a time-frequency point belongs to speech or noise, thus generating a time-frequency binary mask. The classifier uses a set of features extracted from the STFT of the left and right mixtures ($A_L(k,m)$, $\phi_L(k,m)$, $A_R^{B_{Ak}}(k,m)$, and $\phi_R^{B_{Pk}}(k,m)$) and it is trained with the IBM as target output.

The linear discriminant analysis (LDA) [36] is a supervised pattern recognition method that uses a linear combination of a set of input features in order to tackle a classification problem, establishing linear decision boundaries between two or more classes. Let us consider the pattern vector $\mathbf{x_p}$ (i.e. the observations) containing L input features, $\mathbf{x_p} = [x_1, x_2, \ldots, x_L]^T$, which are extracted from the mixture signal in the problem at hand. Each pattern $\mathbf{x_p}$ can be assigned to one of the two possible classes defined in this work, speech or noise. The pattern matrix $\mathbf{P}$ of size $L \times P$ is defined as a matrix that contains the patterns $\mathbf{x_p}$ of a set of P data samples, $\mathbf{P} = [\mathbf{x_1}, \ldots, \mathbf{x_P}]$, and the matrix $\mathbf{Q}$ is defined as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{1} \\ \mathbf{P} \end{bmatrix}, \tag{7}$$

where $\mathbf{1}$ is a row vector of length P. The output of the LDA is obtained as a linear combination of the input features, according to

$$\mathbf{y} = \mathbf{v}^T\mathbf{Q}, \tag{8}$$

where the vector $\mathbf{v} = [v_0, v_1, v_2, \ldots, v_L]^T$ contains the bias v_0 and the weights applied to each of the L input features, and $\mathbf{y}$ is a vector of size $1 \times P$ containing the output of the LDA for the P input patterns. For each of the patterns, the binary mask is generated according to

$$M(k,m) := \begin{cases} 1, & y_p > y_0 \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where y_p is the output of the LDA for the pth pattern and y_0 is a threshold value. The output values of the classifier range from 0 to 1, so the threshold value is set to $y_0 = 0.5$.

The design of the classifier consists in finding the vector $\mathbf{v}$ that minimizes the estimation error. In supervised learning, the true values associated to each data sample are accessible, and they are used to train the classifier. These values are contained in the target vector defined as $\mathbf{t} = [t_1, t_2, \cdots, t_P]^T$, with values of 1 in the case of speech and 0 in the case of noise. In this work, the target values are those corresponding to the IBM defined in (1). The estimation error is defined as the difference between the output values of the LDA (8) and the true values

$$\mathbf{e} = \mathbf{y} - \mathbf{t} = \mathbf{v}^T\mathbf{Q} - \mathbf{t}, \tag{10}$$

and the mean square error (MSE) is computed according to

$$\text{MSE} = \frac{1}{P}\left\|\mathbf{y} - \mathbf{t}\right\|^2 = \frac{1}{P}\left\|\mathbf{v}^T\mathbf{Q} - \mathbf{t}\right\|^2. \tag{11}$$

In the least squares approach (LS-LDA) [30], the weights are adjusted in order to minimize the MSE. The minimization of the MSE is obtained by deriving the expression (11) with respect to every weight of the linear combination, giving raise to the following expression:

$$\mathbf{v} = \mathbf{t}\mathbf{Q}^T\left(\mathbf{Q}\mathbf{Q}^T\right)^{-1}. \tag{12}$$

The LDA is limited to separate both classes linearly. However, it is possible to discriminate classes with more complex decision boundaries by introducing nonlinear transformations of the original input features. In the general case, the matrix $\mathbf{Q}$ can be defined as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{1} \\ f_1(\mathbf{P}) \\ \cdots \\ f_{N_T}(\mathbf{P}) \end{bmatrix}, \tag{13}$$

where $f_1, \ldots, f_{N_T}$ are N_T transformations performed over the original input features contained in $\mathbf{P}$. The weight vector is then defined as $\mathbf{v} = \left[v_0, v_1, \ldots, v_{N_T \cdot L}\right]^T$, and it can also be obtained using expression (12). Henceforth, this is denominated generalized discriminant analysis (GDA), and it is the classification schema used in this paper, which has been labeled as LS-GDA.

The implementation of the proposed classifier is relatively simple, its computational cost being directly related to the number of features included in $\mathbf{Q}$. Considering that the selected data is consecutively stored in memory, and the processor performs the multiply-accumulate (MAC) operation in a single instruction, the number of instructions necessary to process each frequency band by the LS-GDA is approximately $L + 1$, where L is the number of input features (we drop here the constant number of instructions necessary to generate the mask, which is a simple comparison). Hence, limiting the computational cost of the classifier is equivalent to limiting the number of features used for classification. The selection procedure to determine the best set of features to solve the classification problem at hand is included in section 4.

3.3 Evolutionary algorithm to reduce the transmission bit rate

The low-cost classifier proposed in the previous section provides an estimation of the IBM minimizing the MSE. The classifier uses a set of features calculated from the signals received at both ears, which implies that all the information is transmitted from the right device to the left one. Unfortunately, this is not an energy-efficient system. The second step in the design of the binaural speech enhancement system proposed in this paper is the reduction of the transmission bit rate, which implies a reduction in the power consumption, while minimizing the effect that quantization has in the enhanced speech. In this work, we propose to optimize the transmission rate assigning a different number of bits B_{Ak} and B_{Pk} to quantize the values $A_R(k,m)$ and $\phi_R(k,m)$ of each frequency band. The number of bits may also differ between both values of the same frequency. This transmission schema allows assigning more bits to the frequencies and values providing more information to the classifier.

In order to optimize the bit distribution, a tailored evolutionary algorithm is proposed, considering that the number of bits associated to the transmission of the data of each time frame (i.e., the bit rate) is constrained. The algorithm searches the best assignation of bits among frequency bands in order to minimize the MSE obtained by the LS-GDA classifier (the MSE is then the fitness function). The matrix $\mathbf{Q}$ is created including the selected set of features calculated with the values $A_R^{B_{Ak}}(k,m)$ and $\phi_R^{B_{Pk}}(k,m)$ quantized with different number of bits B_{Ak} and B_{Pk}, considering all integer values of bits from 0 to 8. The values $B_{Ak} = 0$ and $B_{Pk} = 0$ mean that no

information from this value in the k-th frequency band is transmitted. Hence, the rows of $\mathbf{Q}$ contain the features quantized with different number of bits. The values $A_R^{B_{Ak}}(k,m)$ and $\phi_R^{B_{Pk}}(k,m)$ received by the left device are simulated by quantizing uniformly the values using $2^{B_{Ak}}$ and $2^{B_{Pk}}$ quantization steps, respectively. The dynamic range has been limited to 90 dB for the amplitude values (A_L and A_R are logarithmic values) and 2π for the phase values.

Each candidate solution is defined by a vector containing the number of bits (between 0 and 8) assigned to the level and phase values ($A_R(k,m)$ and $\phi_R(k,m)$) of each frequency band, a total of $2K$ values (K is the number of frequency bands). The search algorithm selects the quantized features among the rows of the matrix $\mathbf{Q}$ according to the bits of each candidate solution, and then evaluate the classifier using the quantized features. The fitness function is the MSE of the classifier. The complete steps of the search algorithm are as follows:

1. The matrix $\mathbf{Q}$ is created containing the selected set of features calculated using the values $A_R^{B_{Ak}}(k,m)$ and $\phi_R^{B_{Pk}}(k,m)$, quantized with different number of bits, from 0 to 8.
2. An initial population of 100 candidate solutions is generated. Each solution contains $2 \cdot K$ values between 0 and 8 bits, which corresponds with a different number of bits for $A_R(k,m)$ and $\phi_R(k,m)$ for each frequency band.
3. The candidates of the population are validated to fulfill the constraint of the total number of bits. If a candidate solution exceeds by N_D maximum number of bits allowed, the number of bits of a number of N_D random positions of the candidate solution are decreased by one. In case that the number of bits of an element falls below 0, it is set to 0. The procedure iterates until the candidate solution fulfills the requirement.
4. The fitness function (MSE) of the classifier is then evaluated for each candidate solution and frequency band, following the next steps:
   (a) To extract the quantized version of the features from $\mathbf{Q}$, according to the current candidate solution.
   (b) The weight values $\mathbf{v}$ are calculated for each frequency band, using expression (12).
   (c) The MSE of each solution and frequency band is calculated according to expression (11).
   (d) The MSE associated to a candidate solution is the average of the MSE obtained in all frequency bands.
5. A selection process is applied, using the MSE of each solution as ranking. It consists in selecting the best 10% of the solutions of the population, removing the remaining solutions.
6. The remaining 90% solutions of the new generation are then generated by uniform crossover of the best candidates.
7. Mutations are applied in the 1% of the new population, excluding the best obtained solution which is preserved. Mutations consist of increasing or decreasing by one the number of bits of random positions of the mutated candidate solution.
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.

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.

4 Experimental work and results

4.1 Database generation

The suitable database design plays a vital role in any kind of problem based on supervised machine learning. In order to validate the algorithms proposed in this work, a database of binaural speech and noise mixtures has been generated to design and test the classifier. In the case of speech, the TIMIT database described in [37] has been used. It contains a total of 626 speech male/female recordings sampled at 16 kHz with a duration of 4 s. Another 626 noise sources have been selected from an extensive database which contains both stationary and non-stationary noise. Stationary noise refers to monotonous noisy environments, for instance, the aircraft cabin noise. Non-stationary noise to other non-monotonous noises, for example, children shouting in a kindergarten. We have taken into account a variety of noise sources, including those from the following diverse environments: aircraft, bus, cafe, car, kindergarten, living room, nature, school, shops, sports, traffic, train, train station, etc. All the speech and noise signals have been initially normalized with power level of 0 dB.

A number of 1,000 binaural mixtures are generated using the head-related impulse responses (HRIR) included in the CIPIC database [38], which contains recordings of the HRIR with in-the-canal microphones in 43 different human subjects and 2 KEMAR mannequins. The recordings of the database were performed for different spatial directions, splitting the space in 50 elevation

Table 1 Proposed combination of features

SET	*NFtSet*	Features
SET1	3	$A_L, (A_L - A_R)^2, (\phi_R - \phi_L)^2$
SET2	3	$A_L, \text{abs}(A_L - A_R), \text{abs}(\phi_R - \phi_L)$
SET3	4	$A_L, A_L^2, (A_L - A_R)^2, (\phi_R - \phi_L)^2$
SET4	2	$(A_L - A_R)^2, (\phi_R - \phi_L)^2$
SET5	7	$A_L, A_R, A_L^2, A_R^2, A_L \cdot A_R, \text{abs}(A_L - A_R), \text{abs}(\phi_R - \phi_L)$
SET6	6	$A_L, A_L^2, \text{abs}(A_L - A_R), (A_L - A_R)^2, \text{abs}(\phi_R - \phi_L), (\phi_R - \phi_L)^2$

$\boldsymbol{A_L}$ and $\boldsymbol{\phi_L}$ are the amplitude of the STFT of the left signal, respectively; $\boldsymbol{A_R}$ and $\boldsymbol{\phi_R}$ are the amplitude and phase of the STFT of the right signal, respectively.

angles and 25 azimuthal angles, having a total of 1,250 source directions. The mixtures are generated with the following setup: a speech source is placed in the front position (i.e. 0° in azimuth, 0° in elevation), and a different noise source is placed at each side of the head. The speech and noise sources are randomly selected from the TIMIT and noise databases, respectively, and the positions of the noise sources are randomly selected among the positions defined in the CIPIC database, avoiding the front direction. The HRIR used in each mixture is also randomly selected among the HRIR of the different subjects contained in the CIPIC database. The database is generated with SNRs of 0, 3, and 5 dB, considering that the noise power is the addition of the power of both noise sources. Considering that $s[n]$ is the target speech signal and $n_1[n]$ and $n_2[n]$ are the two noise sources, respectively, the binaural mixture is given by

$$\begin{aligned} x_L[n] &= s_L[n] + n_{1L}[n] + n_{2L}[n] \\ x_R[n] &= s_R[n] + n_{1R}[n] + n_{2R}[n], \end{aligned} \tag{14}$$

where the signals at the left and right ear (i.e. $s_{L/R}[n]$, $n_{1L/R}[n]$, and $n_{2L/R}[n]$) are obtained by filtering the original sources with the corresponding HRIR function. For properly designing and testing the speech separation system, the database is split into two different subsets, one for design and another for test. The design set contains the 70% of the 626 signals, and the test set the remaining 30%. It is very important to emphasize that the test sounds are not used in the design process. The sampling rate is 16 kHz and the signals are transformed into the time-frequency domain with a STFT that uses a 128-points Hanning window with 50% of overlap, which corresponds with $K = 64$ frequency bands (DC component is not processed). The target IBM is calculated according to

$$T(k,m) := \begin{cases} 1, & |S_L(k,m)|^2 + |S_R(k,m)|^2 > |N_{1L}(k,m) + N_{2L}(k,m)|^2 + |N_{1R}(k,m) + N_{2R}(k,m)|^2 \\ 0, & \text{otherwise} \end{cases}. \tag{15}$$

Finally, is it worth clarifying that the number of data samples P contained in the matrix $\mathbf{Q}$ is given by $M \times$ *Nmixtures*, where M is the number of time frames of each mixture and *Nmixtures* is the number of mixtures of the database.

4.2 Selection of the input feature space

In this section, we propose different combinations of input features extracted from the STFT of the right and left signals and then we evaluate their performance using the proposed LS-GDA classifier. In order to select the most suitable feature set, we have to find a tradeoff between the speech enhancement obtained by the system and the computational burden associated to the use of each set, bearing in mind the limited computational resources of hearing aids. The study in this section is carried out without considering quantization. Hence, the available information extracted from both mixtures is composed by the values $A_L(k,m)$, $\phi_L(k,m)$, $A_R(k,m)$, and $\phi_R(k,m)$. We propose six different sets of features to compare, which are included in Table 1. Each feature of the set is obtained by applying different simple transformations to the original amplitude and phase values, as defined in expression (13). The transformations f_i includes squared amplitude, amplitude and phase differences, and amplitude product. The different sets contain a variety of number of features, which is labeled as *NFtSet*, and it ranges from 2 to 7. Although quantization is not considered here, we have given more importance to the features extracted from the left signal, which will not be quantized in the final system.

The classification of a time-frequency point into speech or noise can be performed using one of the proposed set of features, where the values are calculated from the STFT of the current time-frequency point. Additionally, we propose to include further information related to the neighbor time-frequency points, calculating also the proposed set of features from the STFT of these points. In this work, we consider to use a T-shaped time-frequency footprint, as it is shown in Figure 2. The value *Nfreqs* represents the number of neighbor frequencies taken in each direction (upper frequencies and lower frequencies), then $2Nfreqs$ is the total number of neighbor frequencies included. The number of previous time frames considered is *Nframes*. Hence, the total number of features L used by the classifier, which depends on the selected set, is given by

$$L = NFtSet(2Nfreqs + Nframes + 1). \tag{16}$$

Note than in the case of a symmetric implementation of the proposed system, an extra number of *Nfreqs* neighbor channels should be transmitted.

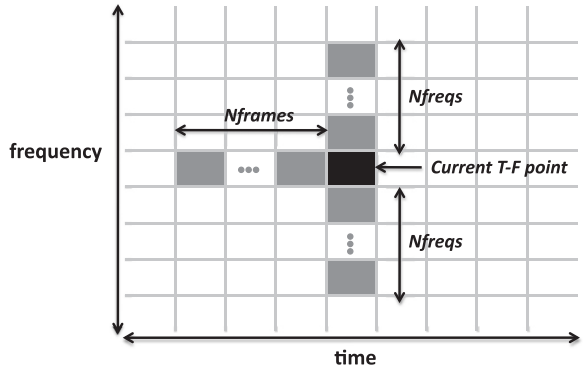

Figure 2 T-shaped time-frequency footprint. Neighbor time-frequency points (gray squares) used to classify the current time-frequency point (black square).

The experiments carried out in this section have two objectives: first, the selection of the best set of features among the six proposed (Table 1) and second, the selection of the optimum time-frequency footprint, finding the best values for *Nfreqs* and *Nframes*. The two problems are solved separately in two different experiments described below.

4.2.1 Selection of the best set of features

The six different sets of features are evaluated using a time-frequency footprint with the same number of neighbor frequencies and time frames, $Nfreqs = Nframes$. The values of *Nfreqs* (and *Nframes*) are varied from 0 to 10, which allows evaluating also the case of using only the information of the current time-frequency point (i.e. $Nfreqs = Nframes = 0$). The comparison is performed in terms of the avarage WDO value obtained by the separation algorithm when the classifier uses each set of features, using the database of speech and noise mixtures previously described, with SNR of 0, 3, and 5 dB . The steps carried out in this experiment are the following:

1. Create the matrix $\mathbf{Q}$ calculating the features corresponding to the evaluated set and time-frequency footprint, using the data from the design set.
2. Calculate the weights of the LS-GDA classifier using Equation (12).
3. Create the matrix $\mathbf{Q}$ calculating the features corresponding to the evaluated set and time-frequency footprint, using now the data from the test set.
4. Generate the binary mask for each mixture of the test database, using the weights calculated in point 2, according to (9).
5. Compute the WDO value for all the mixtures of the test database using the binary mask and the power of the original signals.
6. Repeat steps 1 to 5 for each set of features, time-frequency footprint, and SNR.

The results of the experiment are shown in Figure 3. The represented WDO values have been averaged over all the mixtures in the test database, and they are represented against the total number of features (L), which depends on the feature set and the time-frequency footprint. The different sets of features are represented with lines of different colors, and the different values of *Nfreqs* (and *Nframes*) are represented with squares over the lines. Analyzing the three subfigures, which correspond with different levels of SNR, we deduce that the relative behavior of the different set of features is the same for different SNRs, obtaining higher WDO values with higher SNRs, as it is expected. Additionally, the WDO obtained increases asymptotically with the number of features. It can be easily deduced that the SET2 (red line) represents the best tradeoff in terms of WDO and number of features, for any SNR. In the case of SNR = 0 dB (a), SET2 achieves WDO values around 0.8 with only 50 features. The feature set SET6 achieves the same levels of WDO but using a higher number of features. In the case of the set SET4, which only uses two features, the results are notably worse comparing with the rest of combinations. Adding more features to SET2, as in the cases of SET3, SET5, and SET6, does not bring any improvement. Another important result obtained from this experiment is the noticeable improvement achieved by the introduction of the information of neighbor time-frequency points.

The conclusion of this analysis is that the combination of features labeled as SET2 is the best solution among the evaluated. From here onwards, all the experiments will be carried out with this set of features.

4.2.2 Selection of the best time-frequency footprint

Unlike the previous experiment, which used a time-frequency footprint with the same number of neighbor frequencies and time frames, we consider now that these two values may differ. The objective of the next experiment is to find the optimum values of *Nfreqs* and *Nframes*, using the set of features selected in the previous experiment, SET2. For this purpose, we have evaluated the WDO value obtained by the separation algorithm when the classifier uses different sizes of the time-frequency footprint, varying *Nfreqs* and *Nframes* from 0 to 6 independently, using the features defined in SET2. The steps of the experiment are

1. Create the matrix $\mathbf{Q}$ with the features of SET2 and the time-frequency footprint evaluated, using the data from the design set.
2. Calculate the weights of the LS-GDA classifier using Equation (12).

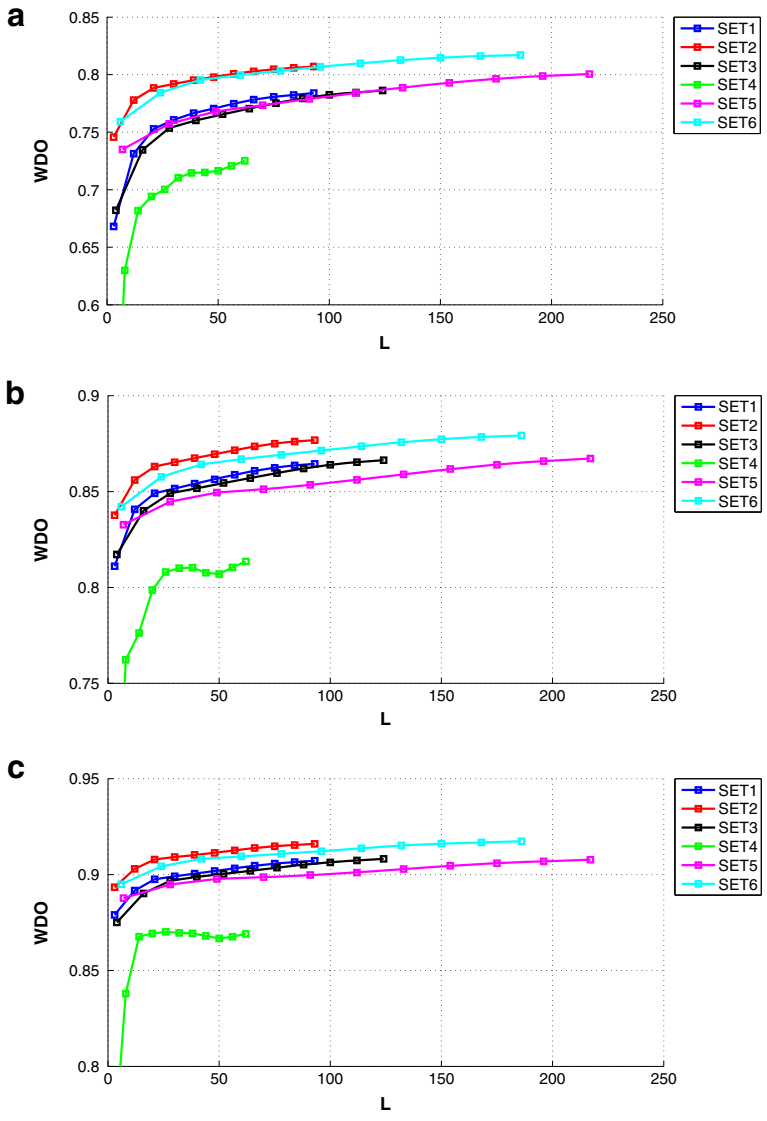

Figure 3 Feature selection. (a) SNR = 0 dB, **(b)** SNR = 3 dB, and **(c)** SNR = 5 dB. Average WDO value obtained by the non-quantified classifier using different combinations of features and different sizes of the time-frequency footprint, with *Nfreqs* = *Nframes*. The different set of features are represented with lines of different colors, and the different values of *Nfreqs* (and *Nframes*) are represented with squares over the lines.

3. Create the matrix **Q** with the features of SET2 and the time-frequency footprint evaluated, using now the data from the test set.
4. Generate the binary mask for each mixture of the test database, using the weights calculated in point 2, according to (9).
5. Compute the WDO value for all the mixtures of the test database using the binary mask and the power of the original signals.
6. Repeat steps 1 to 5 for each value of *Nfreqs* and *Nframes* and each SNR.

Figure 4 shows the results of this experiment. The WDO values have been averaged over all the mixtures in the test database, and they are represented against the total number of features (L), which depends on the values given to *Nfreqs* and *Nframes* (see expression (16)). The different values of *Nframes* are represented with lines of different colors, and the different values of *Nfreqs* with squares over the lines. The plot in (a) corresponds with a SNR of 0 dB, the plot in (b) with a SNR of 3 dB, and the plot in (c) with a SNR of 5 dB. The relative behavior of the different sizes of the time-frequency footprint is the same for the different SNRs. Concerning the number of previous time frames (*Nframes*), the higher WDO values are generally obtained when using only 2 time frames. Regarding the number of neighbor frequencies,

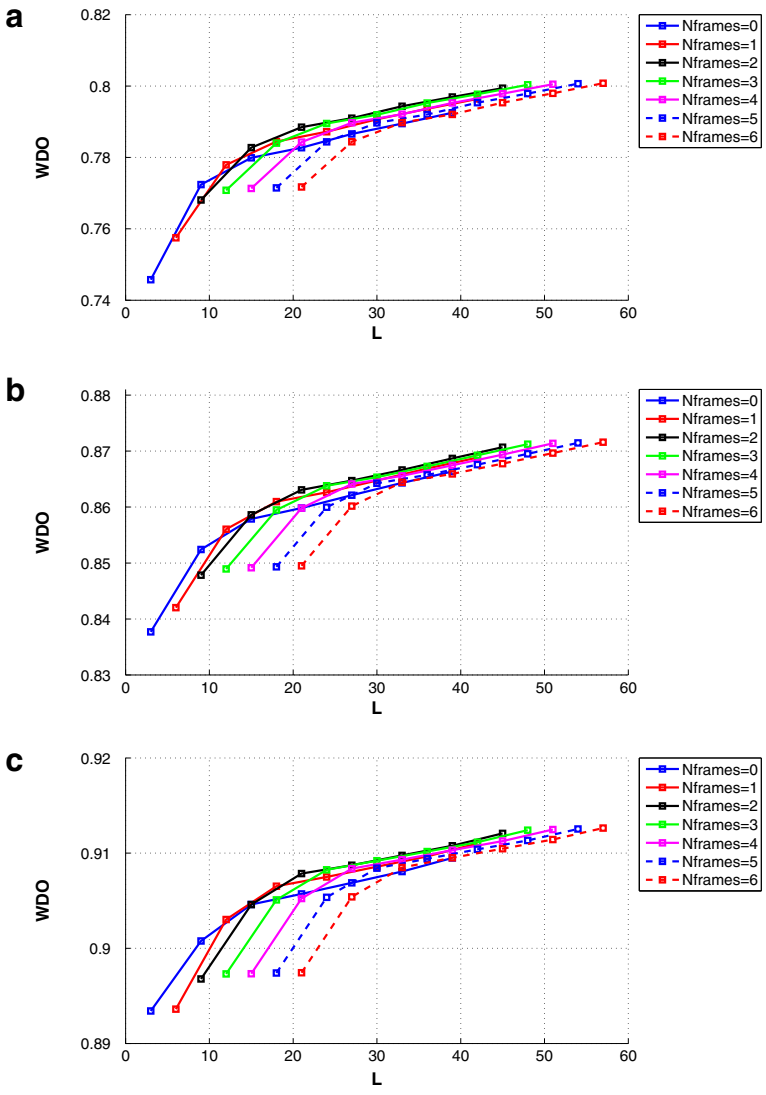

Figure 4 Time-frequency footprint selection. (a) SNR = 0 dB, **(b)** SNR = 3 dB and **(c)** SNR = 5 dB. Description: average WDO value obtained by the non-quantified classifier with the selected combination of features varying the number of neighbor frequencies (*Nfreqs*) and previous time frames (*Nframes*) of the time-frequency footprint. The different values of *Nframes* are represented with lines of different colors and the different values of *Nfreqs* with squares over the lines.

the increment of the WDO values is more noticeable for values up to $Nfreqs = 3$, whereas the amount of increment decreases with higher number of frequencies. Finally, the improvement obtained by the introduction of the information of neighbor time-frequency points is clearly demonstrated.

From the analysis of the results obtained with this experiment, we propose that a time-frequency footprint with $Nfreqs = 3$ and $Nframes = 2$ represent a good tradeoff between speech separation and computational cost. The proposed solution obtains an WDO value of 0.79 for mixtures at 0 dB, using only 27 features to classify each time-frequency point. Finally, it is worth mentioning that a square-shaped time-frequency footprint have been also considered. However, it does not outperform the results of the T-shaped footprint due to the notably higher number of required features.

4.2.3 *Evaluation of the computational cost associated to the proposed solution*

In order to objectively evaluate the computational cost associated to the proposed classifier, we perform a quantitative study of the number of instructions available for speech enhancement in hearing aids, considering the characteristics of a state-of-the-art commercial device. Common DSP's embedded in hearing aids have a

processor with a selective clock speed that usually ranges from 1.28 to 5.12 MHz. They use a Harvard architecture containing a multiplier-accumulator (MAC) with a set of instructions completed in a clock cycle; hence, the number of mega instructions per second (MIPS) is the clock speed value. The sampling rate f_s is usually adjustable but limited by the output frequency range of the loudspeaker, 16 kHz normally being the selected sampling rate. The analysis and synthesis windows have a length of L_{WIN} samples working with 50% of overlap, and the DTF-based frequency analysis contains K frequency bands. According to this, the number of instructions available to process each frequency band (IPF) of a frame is calculated using the next expression:

$$\text{IPF} = \frac{\text{MIPS}}{K}\frac{L_{\text{WIN}}/2}{f_s}. \tag{17}$$

In the special case of a processor with a clock speed of 5.12 MHz (5 MIPS), and working with a sampling rate of 16 kHz, analysis window of 128 samples and 65 frequency bands (i.e., our case), the number of instructions available to process each frequency band of a frame is 308. These instructions are shared between the different signal processing algorithms included in the device: the multi-band compression-expansion algorithm, feedback cancellation, automatic acoustic environment classification, and speech enhancement. Hence, the proposed speech enhancement algorithm should only use a part of the total number of available instructions.

The solution selected after the study carried out in this section uses 27 features. The number of instructions necessary to process each frequency band by the LS-GDA is approximately $L+1$. Therefore, the number of instructions associated to the proposed solution represents less than 9% of the available number of instructions. This result supports the feasibility of implementing the proposed speech enhancement algorithm in real hearing aids.

4.3 Optimizing the transmission rate

The proposed evolutionary algorithm to optimize the bit distribution has been executed different times varying the transmitted bit rate from 0 to 512 kbps. In the case that the bit rate is 512 kbps, all the quantized data is transmitted with the maximum number of bits, $B_{Ak} = 8$ and $B_{Pk} = 8$ (i.e., 16 bits per frequency band, $K = 64$, and 500 frames per second); hence, the optimization is not required. In order to compare the effectiveness of the proposed algorithm, we have also evaluated the performance obtained by an uniform distribution of bits, assigning a constant number of bits to the amplitude and phase values of each frequency band. The values assigned in this case are 1, 2, 4, and 8, which corresponds with transmission rates of 64, 128, 256, and 512 kbps, respectively.

Figure 5 represents the WDO values, averaged over the mixtures of the test set, as a function of the transmission bit rate (kbps). The WDO values obtained by the proposed transmission schema in the case of an optimized distribution of bits are represented by solid lines and in the case of an uniform distribution of bits are represented in dashed lines. The different colors represent different SNR values. Additionally, the limiting WDO value (WDO_{lim}), which is the WDO value obtained by the IBM and represents the upper bound for any possible separation system, is represented by a straight line for each SNR.

In the case of transmitting the quantized values with the maximum number of bits (512 kbps), the WDO values obtained by the proposed algorithm practically match the WDO values in case of non-quantization (i.e., using $A_R(k,m)$ and $\phi_R(k,m)$). The performance is nearly unaffected when the transmission rate is decreased up to 128 kbps, but the decrease begins to be noticeable for lower

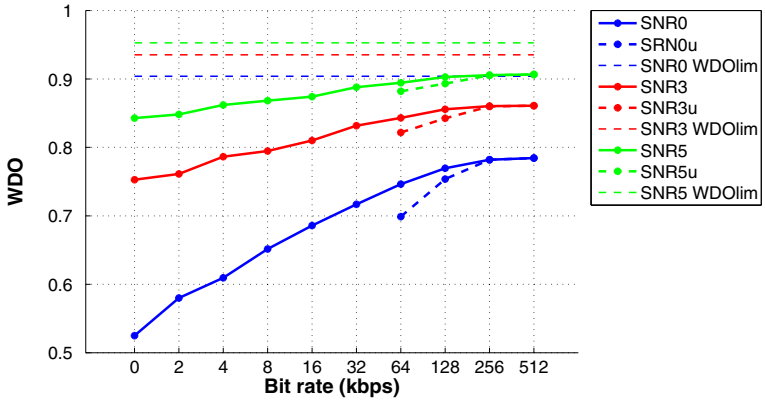

Figure 5 WDO vs. bit rate. WDO values averaged over the test set as a function of the transmission bit rate (kbps) obtained by the proposed transmission schema in the case of an optimized distribution of bits (solid line) and a uniform distribution of bits (dashed line). The different colors represent different SNR values. The limiting WDO value (WDO_{lim}) is represented by a straight line for each SNR.

bit rates. Nevertheless, in the case of SNR = 0 dB (worst case), the performance is only reduced by 4% in the case of transmitting 64 kbps, 12% in the case of transmitting 16 kbps, 17% in the case of transmitting 8 kbps, and 25% in the case of transmitting 2 kbps, which are acceptable transmission rates for hearing aids. Additionally, the figure also shows the case in which no information is transmitted from the right to the left device (0 kbps). In such a case, the features are calculated only using the information available in the left ear (i.e., monaural system), and the performance clearly drops to WDO values around 0.5 for SNR = 0 dB, which supports the use of binaural separation. Moreover, it is noticeable that the results obtained by the optimized distribution outperforms the results obtained by the uniform distribution, the difference increasing when the number of bits decreases. Nevertheless, the use of a uniform distribution does not allow reducing the transmission rate below 64 kbps.

Figure 6 illustrates the bit distribution obtained by the optimization algorithm in the case of a transmission rate of 64 kbps and SNR of 0 dB. The blue bars represent the number of bits assigned to the amplitude values, the red bars represent the number of bits assigned to the phase values, and the dashed black line represents the total number of bits assigned to each frequency band. In the lower frequency bands, the bits are mainly assigned to the phase values, whereas in the higher frequency bands, more bits are assigned to the amplitude values. This behavior is expected due to the fact that interaural time differences predominates in the lower frequencies and interaural level differences predominates in the higher frequencies. The optimization algorithm clearly allows an efficient bit distribution.

5 Conclusions

This paper presents a novel energy-efficient sound separation algorithm with very low computational cost for speech enhancement in wireless-communicated binaural hearing aids. The source separation algorithm is based on supervised machine learning and time-frequency masking, and the design of the system has been carried out considering the power and computational limitations of state-of-the-art hearing aids. First, the computational cost of the algorithm has been constrained, obtaining good separation performance in terms of WDO even for low SNRs when using less than the 10% of the available computational resources for signal processing. The combination of features selected represents a tradeoff between separation performance and computational cost. The improvement associated to the introduction of the information of neighbor time-frequency points in the decision whether a time-frequency point belongs to speech or noise has been proven. Second, the transmission bit rate associated to the information exchange between both devices has been also constrained, optimizing the distribution of number of bits among the different frequency bands with an evolutionary algorithm. The performance of the algorithm in terms of WDO is only reduced by 4% in the case of transmitting 64 kbps, 17% in the case of transmitting 8 kbps, and 25% in the case of transmitting only 2 kbps, which are feasible bit rates for hearing aids. The optimization algorithm allows distributing the bits efficiently. Finally, the advantages of binaural source separation in comparison to the monaural case have been amply demonstrated.

The proposed algorithm has been tested in a scenario where the desired speech source is contaminated with two directional noises, in low SNR conditions. In order to generalize the results for a typical hearing aid application,

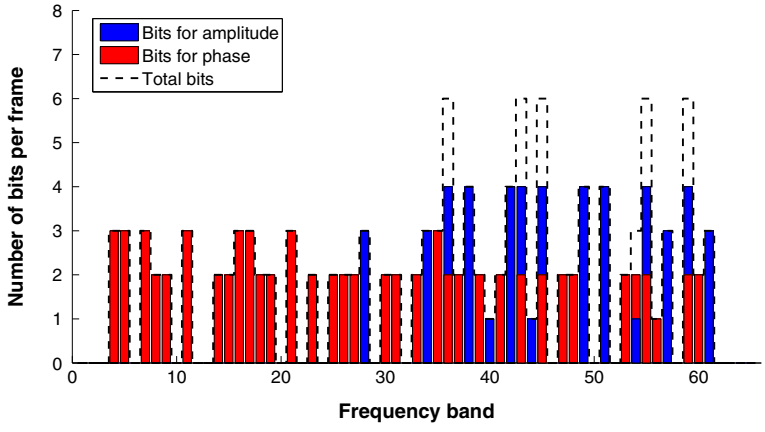

Figure 6 Bit distribution among frequency bands. Distribution of the number of bits per frame among the frequency bands, in the case of a transmission rate of 64 kbps and SNR of 0 dB. The blue bars represent the number of bits assigned to the amplitude values (B_{Ak}), the red bars represent the number of bits assigned to the phase values (B_{Pk}), and the dashed black line the total number of bits assigned to each frequency band ($B_{Ak} + B_{Pk}$).

the proposed algorithm should also be tested with diffuse background noise and reverberations. Additionally, other metrics related to speech quality or intelligibility should be used to evaluate the performance of the algorithm. Finally, it is worth noting that the tradeoff between transmission bit rate and separation performance can be further studied in an information theoretic framework.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work has been funded by the Spanish ministry of economy and competitiveness, under project TEC2012-38142-C04-02 and the scholarship AP2009-3932.

Received: 29 June 2013 Accepted: 5 December 2013
Published: 20 December 2013

References

1. A Spriet, G Rombouts, M Moonen, J Wouters, Adaptive feedback cancellation in hearing aids. J. Franklin Inst. **343**(6), 545–573 (2006)
2. DJ Freed, Adaptive feedback cancellation in hearing aids with clipping in the feedback path. J. Acoust. Soc. Am. **123**, 1618 (2008)
3. P Nordqvist, A Leijon, An efficient robust sound classification algorithm for hearing aids. J. Acoust. Soc. Am. **115**, 3033 (2004)
4. E Alexandre, L Cuadra, L Álvarez, M Rosa-Zurera, F López-Ferreras, Two-layer automatic sound classification system for conversation enhancement in hearing aids. Integr Comput. Aided Eng. **15**, 85–94 (2008)
5. PM Peterson, P Zurek, Multimicrophone adaptive beamforming fo reduction in hearing aids. J. Rehabil. Res. Dev. **24**(4) (1987)
6. V Hamacher, J Chalupper, J Eggers, E Fischer, U Kornagel, H Puder, U Rass, Signal processing in high-end hearing aids: state of the art, challenges, and future trends. EURASIP J. Appl. Signal Process. **2005**, 2915–2929 (2005)
7. DB Hawkins, WS Yacullo, Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation. J. Speech Hear. Disord. **49**(3), 278 (1984)
8. K Chung, Challenges and recent developments in hearing aids. Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms. Trends Amplif. **8**(3), 83–124 (2004)
9. JM Kates, MR Weiss, A comparison of hearing-aid array-processing techniques. J. Acoust. Soc. Am. **99**, 3138 (1996)
10. GH Saunders, JM Kates, Speech intelligibility enhancement using hearing-aid array processing. J. Acoust. Soc. Am. **102**, 1827 (1997)
11. R Stadler, W Rabinowitz, On the potential of fixed arrays for hearing aids. J. Acoust. Soc. Am. **94**, 1332 (1993)
12. M Hoffman, T Trine, K Buckley, D Van Tasell, Robust adaptive microphone array processing for hearing aids: realistic speech enhancement. J. Acoust. Soc. Am. **96**, 759 (1994)
13. JE Greenberg, Modified LMS algorithms for speech processing with an adaptive noise canceller. IEEE Trans. Speech Audio Process. **6**(4), 338–351 (1998)
14. A Spriet, M Moonen, J Wouters, Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications. IEEE Trans. Speech Audio Process. **13**(4), 487–503 (2005)
15. D Wang, GJ Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley Interscience, Hoboken, 2006)
16. D Wang, Time-frequency masking for speech separation and its potential for hearing aid design. Trends Amplif. **12**(4), 332–353 (2008)
17. G Hu, D Wang, Speech segregation based on pitch tracking and amplitude modulation, in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics,* New Paltz, New York, 21–24 October, 2001 (IEEE, Piscataway, 2001), pp. 79–82
18. S Srinivasan, N Roman, D Wang, Binary and ratio time-frequency masks for robust speech recognition. Speech Commun. **48**(11), 1486–1501 (2006)
19. N Li, PC Loizou, Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. J. Acoust. Soc. Am. **123**, 1673 (2008)
20. DS Brungart, PS Chang, BD Simpson, D Wang, Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. **120**, 4007 (2006)
21. JG Desloge, WM Rabinowitz, PM Zurek, Microphone-array hearing aids with binaural output. I. Fixed-processing systems. IEEE Trans. Speech Audio Process. **5**(6), 529–542 (1997)
22. T Lotter, P Vary, Dual-channel speech enhancement by superdirective beamforming. EURASIP J. Appl. Signal Process. **2006**, 175–175 (2006)
23. DP Welker, JE Greenberg, JG Desloge, PM Zurek, Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system. IEEE Trans. Speech Audio Process. **5**(6), 543–551 (1997)
24. N Roman, S Srinivasan, D Wang, Binaural segregation in multisource reverberant environments. J. Acoust. Soc. Am. **120**, 4040 (2006)
25. T Rohdenburg, V Hohmann, B Kollmeier, Robustness analysis of binaural hearing aid beamformer algorithms by means of objective perceptual quality measures, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, New York, 21–24 October, 2007 (IEEE, Piscataway, 2007), pp. 315–318
26. T Wittkop, V Hohmann, Strategy-selective noise reduction for binaural digital hearing aids. Speech Commun. **39**, 111–138 (2003)
27. O Roy, M Vetterli, Rate-constrained beamforming for collaborating hearing aids, in *IEEE International Symposium on Information Theory* Seattle, Washington, 6–12 July, 2006 (IEEE, Piscataway, 2006), pp. 2809–2813
28. S Doclo, M Moonen, T Van den Bogaert, J Wouters, Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids. IEEE Trans. Audio, Speech, Language Process. **17**, 38–51 (2009)
29. S Srinivasan, AC Den Brinker, Rate-constrained beamforming in binaural hearing aids. EURASIP J. Adv. Signal Process. **2009**, 8 (2009)
30. J Ye, Least squares linear discriminant analysis, in *Proceedings of the 24th International Conference on Machine learning* Oregon State University, Corvalis, OR, 20–24 June, 2007 (ACM, New York, 2007), pp. 1087–1093
31. R Gil-Pita, L Cuadra, E Alexandre, D Ayllón, L Alvarez, M Rosa-Zurera, Enhancing the energy efficiency of wireless-communicated binaural hearing aids for speech separation driven by soft-computing algorithms. Appl. Soft Comput. **12**(7), 1939–1949 (2012)
32. A Jourjine, S Rickard, O Yilmaz, Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures, in *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Istanbul, 5–9 June, 2000, vol. 5 (IEEE, Piscataway, 2000), pp. 2985–2988
33. S Rickard, O Yilmaz, On the approximate W-disjoint orthogonality of speech, in *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Renaissance Orlando Resort, Orlando, FL , 13–17 May, 2002, vol. 1 (IEEE, Piscataway, 2002), pp. I–529
34. Y Li, D Wang, On the optimality of ideal binary time–frequency masks. Speech Commun. **51**(3), 230–239 (2009)
35. O Yilmaz, S Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
36. RA Fisher, The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
37. WM Fisher, GR Doddington, KM Goudie-Marshall, The DARPA speech recognition research database: specifications and status, in *DARPA Workshop on Speech Recognition*, (1986), pp. 93–99
38. VR Algazi, RO Duda, DM Thompson, C Avendano, The CIPIC HRTF database, in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* New Paltz, New York, 21–24 October, 2001 (IEEE, Piscataway, 2001), pp. 99–102

doi:10.1186/1687-6180-2013-187
Cite this article as: Ayllón *et al.*: **Rate-constrained source separation for speech enhancement in wireless-communicated binaural hearing aids.** *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:187.