Applied Acoustics 204 (2023) 109238

Contents lists available at ScienceDirect

Applied Acoustics

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

Improved Aures tonality metric for complex sounds

Juan Estreder [a], Gema Piñero [a,*], Maria de Diego [a], Jussi Rämö [b], Vesa Välimäki [b]

[a] Inst.Telecommunication and Multimedia Applications (iTEAM), Universitat Politècnica de València, Valencia, Spain
[b] Acoustics Lab, Department of Information and Communications Engineering, Espoo, Finland

ARTICLE INFO

Article history:
Received 13 September 2022
Received in revised form 23 December 2022
Accepted 20 January 2023

Keywords:
Tonality
Frequency masking
Complex sounds
Subjective test
Perceptual equalization

ABSTRACT

This paper proposes improvements to the Aures' tonality metric, which can be used for estimating the frequency masking of complex sounds. The perception of tonality has been extensively studied in simple sounds, such as pure tones and narrowband noise signals, but there are no solid conclusions in the case of complex sounds. Previously, Aures' method has been mostly used in the psychoacoustic analysis of noise signals. The modifications presented here are a better spectral resolution, a lowered tonal threshold, and a different exponent in one of their weighting functions. These may appear to be minor changes from the original Aures (OA) method, but they have proven to be significant for perception. The improved Aures (IA) method has been validated by a subjective test using three different multitone signals in the presence of a narrowband noise, whose result is the subjectively perceived masking thresholds. Results show that the IA method presents an average error of 0.8 dB when predicting the subjective masking thresholds provided by the test, while the average errors of the OA and a baseline spectral flatness method exceed 5 dB. In addition, a second subjective test has been carried out to assess the perceptual equalization of a music signal using the proposed IA and spectral flatness methods. The second test confirms that the IA method is preferred. Therefore, the improved Aures method is proposed as a reliable tonality metric for complex sounds, such as multi-tone signals and music.

1. Introduction

The way we perceive sounds has been studied from a systematic point of view for several decades [1], giving rise to the field of psychoacoustics [2]. The knowledge provided by psychoacoustic studies has been used in the field of digital audio processing, and there are numerous algorithms designed to process sounds from a perceptual point of view [3]. A perceptual effect that has been extensively studied and modeled in the audio processing field is the frequency masking between two sounds, also known as simultaneous masking. According to [2], when two different sounds are perceived by the auditory system at the same time, the target sound (or masked sound) suffers a change in its perception due to the other sound (or masking sound). This effect can be analyzed using the frequency masking threshold model (MTM), which quantifies the minimum sound pressure level of the masked sound to be audible in the presence of the masking sound.

The most popular applications where the MTM is applied are audio coding [4–6], adaptive audio equalization [7,8], speech enhancement regarding environmental noise [9,10] and evaluation of noise annoyance in general, but especially in cars [11,12]. Generally speaking, the MTM is used to boost the speech or music above the level of the added noise in order to improve its intelligibility, audibility or any other desired characteristic.

Different approaches have been proposed to estimate the frequency masking threshold of the masker sound, but the model originally proposed in [5,6] is commonly used in the audio processing field. This model applies a "masking pattern" at every critical band mimicking the way our hearing system processes the sounds. The shape of the masking patterns have been thoroughly studied over the years as shown in [13], where an extensive analysis of different masking patterns, also called *spreading functions*, can be found.

Although these *spreading functions* represent the masking pattern for narrowband signals (understood as those signals whose frequency content is within a single critical band), they have also been used for the analysis of broadband and multi-tone signals, which are referred here as complex sounds. The masking of broadband signals was first studied by Green [14], whose experiments showed that the energy contributions of the different maskers should be summed up to generate the overall masking curve of the broadband signal. This phenomenon is commonly known as "additivity of masking" [2,15].

* Corresponding author.
E-mail address: gpinyero@iteam.upv.es (G. Piñero).

https://doi.org/10.1016/j.apacoust.2023.109238
0003-682X/

The experiments described in [16,17] show that the narrowband noise and pure tone signals do not mask equally even though they present the same SPL values and are located at the same critical bands. Therefore, the MTM not only considers the masking patterns, but also the asymmetry of the masking. This asymmetry is modeled through the tonality offset or tonality metric, which estimates the shape of the energy distribution of the masker sound, ranging from a flat (noise-like) shape to a peak (tone-like) spectrum [16]. Its effect on the masking threshold was studied by Zwicker through empirical tests [2], but several models and methods have been proposed afterwards to estimate it for its inclusion in the MTM. The tonality metric most commonly used at present for audio signals was proposed by Johnston [5]. Similar approaches to this method have been used in perceptual audio coding [6,13,18] and music equalization in presence of noise [7,19]. Although the approach of Johnston has been widely used, it can be inaccurate for certain complex sounds such as speech, multi-tone and high frequency noise signals [6,20].

In [21], a second set of methods to estimate the tonality metric can be found. However, none of these methods have managed to accurately describe the tonality of complex sounds [21,22]. A last tonality metric is the method proposed by Aures [23], which was studied in depth by Hastings [22,24]. Aures' tonality has proven to be very relevant in the psychoacoustic modelling of machinery tonal noises. In [25], it is shown how the subjective annoyance almost linearly depends on the "Tonality exceeded 5% of the time" factor for aircraft noises with similar loudness level. Recently, [26–28] have proposed the inclusion of the tonality into the psychoacoustic annoyance model created by Zwicker [2] to evaluate tonal noises produced by high-voltage transformers, wind turbines and aircraft engines, respectively. Torija et al. [29] also use Aures' tonality to predict the subjective annoyance of aircraft noise. However, they noticed a poor performance of Aures' model when the aircraft noise is formed by tones evenly spaced across the frequency spectrum, suggesting further optimization of some parameters of the Aures tonality model.

In this paper, a new approach based on the original Aures (OA) method and called here the improved Aures (IA) method is proposed in order to estimate the tonal factor affecting the masking threshold model. The improvement is mostly related to the computation of the tonal weighting used in the method explained in Section 2.3, but the whole Aures' method is revisited. In order to validate the proposed IA method and compare it with the original Aures' and Johnston's methods, a first subjective test has been designed to obtain the perceived masking threshold of a multitone signal in the presence of a narrowband noise. The experimental results obtained in the test are then compared to the MTM computed with the three tonality metrics: original and improved Aures, and Johnston, and the proposed model is validated as the best fit. Afterwards, a second subjective test is carried out where an audio signal in the presence of broadband noise is perceptually equalized [7,8] using the improved Aures and Johnston models to compute the tonal factor. Again, the results validate the proposed model for its use with complex sounds as music and broadband noise.

The outline of the paper is as follows: Section 2 describes the improved Aures model used to estimate the tonal factor included in the tonality offset of the MTM, highlighting the main differences with respect to the original model [23]. Section 3 briefly describes the MTM in order to show how the tonality offset contributes to the masking process. Section 4 describes the first subjective test and discusses the results obtained, whereas Section 5 describes the second subjective test together with the results obtained and a brief discussion. Finally, Section 6 concludes the paper.

2. Improved Aures model to compute the tonal factor

The proposed tonality model is based on the method to obtain the tonal factor originally proposed by Aures [23,24], which we call here the original Aures (OA) model, and whose block diagram is shown in Fig. 1, based on the same block diagram that appears in [23,30], although we have modified some of their parameters, denoting this modified model as the improved Aures (IA) tonality model. The tonal factor obtained by means of Fig. 1 will be used in the masking threshold model explained in Section 3.

The input in Fig. 1 is the mth time frame $x_m(n)$ of M samples of the signal of interest, $x(n)$. Therefore, the output is the tonal factor of the mth frame, μ_m, which is computed through the two main branches shown in Fig. 1. The upper branch considers the frequency, bandwidth and SPL level of the tonal components of $x_m(n)$, while the lower branch takes into account its loudness separating the contributions of tone-like components from the noise-like ones.

First of all, the power spectrum of $x_m(n)$, denoted by $P_m(k)$, is estimated, with $k = 0, \ldots, \frac{N_{\text{FFT}}}{2}$, being N_{FFT} the Fast Fourier Transform (FFT) size. For this purpose, the Welch method [31] with a Hamming window and an overlap of 50% is applied, where the power spectrum average is computed using the three previous frames. We assume that the microphone has been previously calibrated, thus $P_m(k)$ represents the SPL distribution in dB of $x_m(n)$ in the frequency range of $[0 - f_s/2]$ Hz, where f_s is the sampling frequency. For the sake of clarity, the subscript m relative to the mth frame is omitted for the following blocks in Fig. 1.

2.1. Extraction of tonal and noise components

Once $P(k)$ enters the upper branch in Fig. 1, the first step is the extraction of their tonal and noise components, which is described as the *pitch extraction algorithm* in Terhardt *et al.* [32]. For this purpose, all the frequency bins of $P(k)$ are analyzed to determine whether they satisfy the following two conditions:

$$P(k-1) < P(k) \geqslant P(k+1), \tag{1a}$$

$$P(k) - P(k \pm \lambda) \geqslant T_{\text{H}}, \tag{1b}$$

where $P(k)$ is expressed in dB, T_{H} is a threshold stated to detect tonal components, and λ represents the neighboring components to be checked, excluding ± 1 since the first condition already considers them. Therefore, the kth component of $P(k)$ will be considered "tonal" if the following two conditions are fulfilled:

1. It must be the largest component considering its two nearest neighbors.
2. It must be, at least, T_{H} dB larger than its $\pm\lambda$-separated neighboring components, with $\lambda = 2, 3, \ldots$

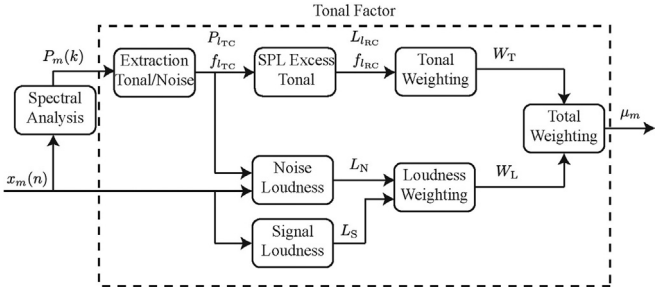


Fig. 1. Block diagram of Aures' method to estimate tonal factor μ_m.

Both λ and T_H values are chosen depending on the frequency resolution of $P(k)$. In [32] a frequency resolution of 12.5 Hz, a set of λ values given by $\lambda = \{2, 3\}$ and a threshold of $T_H = 7$ dB are used. This last value was empirically set from the analysis of different sounds [32].

Considering that input signals have been obtained with a sampling frequency of $f_s = 44100$ Hz, we have used an FFT of size $N_{FFT} = 4096$ to efficiently estimate $P(k)$. Therefore, the time frame $x_m(n)$ is 92.87 ms long and the frequency resolution is $f_s/N_{FFT} = 10.76$ Hz, quite similar to that used by Terhardt [32]. Regarding the neighborhood range in (1b), the maximum value of $\lambda = 3$ in [32] produced a minimum frequency bandwidth of $6 \cdot 12.5 = 75$ Hz for the tonal components. In order to keep as much as possible that frequency resolution, we have used a range of $\lambda = 2, 3, 4$, obtaining a minimum frequency bandwidth of $8 \cdot 10.76 = 86.08$ Hz for the tonal components. Similarly to the procedure followed in [32], the value of the threshold, T_H (1b), has been obtained through empirical tests resulting in a $T_H = 5.5$ dB.

Once the parameters of the "Extraction Tonal/Noise" block have been set, the sound pressure level $P_{l_{TC}}$ and the frequency $f_{l_{TC}}$ of the tonal components are obtained. The index l_{TC} indicates the tonal component ranging from $l_{TC} = 1, \ldots, M_{TC}$, where M_{TC} is the number of tonal components found.

2.2. Sound pressure level excess

This block tries to identify which of the tonal components $P_{l_{TC}}$ are aurally relevant. For this purpose, the SPL excess $L_{l_{TC}}$ is estimated for each tonal component as [32]

$$L_{l_{TC}} = P_{l_{TC}} - 10\log_{10}\left(\left[\sum_{\substack{\gamma=1 \\ \gamma \neq l_{TC}}}^{M_{TC}} 10^{\frac{L_{e\gamma}\left(f_{l_{TC}}\right)}{20}}\right]^2 + I_{N,l_{TC}} + 10^{\frac{L_{TH}\left(f_{l_{TC}}\right)}{10}}\right), \quad (2)$$

where $L_{e\gamma}\left(f_{l_{TC}}\right)$ is the excitation level over the l_{TC}-th tonal component due to the tonal component γ, $I_{N,l_{TC}}$ is the noise intensity in the critical band where the l_{TC}-th tonal component is located, and $L_{TH}\left(f_{l_{TC}}\right)$ is the level of the hearing threshold at the frequency $f_{l_{TC}}$.

The excitation level in (2) can be computed as [32]:

$$L_{e\gamma}\left(f_{l_{TC}}\right) = P_\gamma - \rho\left(f_\gamma, f_{l_{TC}}\right)\left(v_\gamma - v_{l_{TC}}\right), \quad (3)$$

where P_γ is the SPL of the tonal component γ expressed in dB, v_γ and $v_{l_{TC}}$ are the frequencies of the tonal components expressed in the Bark domain [33]:

$$v = 13\arctan\left(\frac{0.76f}{1000}\right) + 3.5\arctan\left(\frac{f}{7500}\right)^2, \quad (4)$$

where f is the frequency in Hz and v is the Bark index (or critical band index), being $v = 1, \cdots, N_c$ with $N_c = 24$ for $f_s = 44100$ Hz.

The parameter $\rho\left(f_\gamma, f_{l_{TC}}\right)$ in (3) specifies a masking pattern and is expressed as (in dB/Bark):

$$\rho\left(f_\gamma, f_{l_{TC}}\right) = \begin{cases} 27 & \text{if } f_{l_{TC}} \leqslant f_\gamma \\ -24 - \frac{230}{f_\gamma} + 0.2P_\gamma & \text{if } f_{l_{TC}} > f_\gamma. \end{cases} \quad (5)$$

The noise intensity $I_{N,l_{TC}}$ in (2) is obtained through the addition of all the components located within the critical band between $v_{l_{TC}} - 0.5$ and $v_{l_{TC}} + 0.5$, without including the tonal components l_{TC} and their neighbors (1b). Finally, the hearing threshold L_{TH} in (2) can be obtained as

$$L_{TH}\left(f_{l_{TC}}\right) = 3.64\left(\frac{f_{l_{TC}}}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f_{l_{TC}}}{1000} - 3.3\right)^2} + 10^{-3}\left(\frac{f_{l_{TC}}}{1000}\right)^4, \quad (6)$$

where $L_{TH}\left(f_{l_{TC}}\right)$ is expressed in dB.

Once each $L_{l_{TC}}$ (2) has been computed, only those tonal components, l_{TC}, such that $L_{l_{TC}} > 0$ will be considered aurally relevant and the rest of components will be discarded. In order to differentiate the relevant tonal components at the output of the "SPL Excess Tonal" block, we introduce a new subscript l_{RC} such that:

$$L_{l_{RC}} = L_{l_{TC}} \qquad \text{for } L_{l_{TC}} > 0, \quad (7)$$

where l_{RC} ranges from $l_{RC} = 1, \ldots, M_{RC}$, with M_{RC} as the number of relevant tonal components found.

2.3. Tonal weighting

The aurally relevant tonal components now enter the "Tonal Weighting" block of Fig. 1, where the tonal weighting W_T is estimated as:

$$W_T = \sqrt{\sum_{l_{RC}=1}^{M_{RC}} w_1^2\left(\Delta z_{l_{RC}}\right) w_2^2\left(f_{l_{RC}}\right) w_3^2\left(L_{l_{RC}}\right)}, \quad (8)$$

where the weighting functions w_n, with $n = 1, 2, 3$, are defined as

$$w_1\left(\Delta z_{l_{RC}}\right) = \frac{0.13}{\Delta z_{l_{RC}} + 0.13} \quad (9a)$$

$$w_2\left(f_{l_{RC}}\right) = \frac{1}{\sqrt{1 + 0.2\left(\frac{f_{l_{RC}}}{700} + \frac{700}{f_{l_{RC}}}\right)^2}} \quad (9b)$$

$$w_3\left(L_{l_{RC}}\right) = \left(1 - e^{\frac{-L_{l_{RC}}}{15}}\right), \quad (9c)$$

being $f_{l_{RC}}$ and $\Delta z_{l_{RC}}$ the frequency and the bandwidth of the relevant component l_{RC} in the Bark domain respectively. In the original Aures' method [23], the first weighting function was defined as $w_1\left(\Delta z_{l_{RC}}\right) = \left(\frac{0.13}{\Delta z_{l_{RC}} + 0.13}\right)^{\frac{1}{0.29}}$. This is the major modification that we have carried out onto Aures' method, and it is justified by the results of the subjective test that will be discussed in Section 4.

2.4. Loudness weighting

Once the tonal weighting W_T has been obtained in the upper branch of Fig. 1, in the following we describe the blocks of the lower branch that considers the effect of the loudness over the tonal factor and whose inputs are the signal frame $x_m(n)$ and the tonal components parameters $P_{l_{TC}}$ and $f_{l_{TC}}$. First of all, the method computes the "Noise Loudness" as the loudness level L_N of $x_m(n)$ once their l_{TC} tonal components have been eliminated, and the "Signal Loudness" as the loudness level L_S of the whole frame $x_m(n)$. Both loudness levels are estimated according to the Zwicker method [34]. The loudness weighting is estimated through the ratio of $\frac{L_N}{L_S}$ that corresponds to the loudness percentage of the noisy part of the $x_m(n)$. Therefore, we can state that the loudness weighting (W_L) represents the loudness percentage of the tonal part of $x_m(n)$:

$$W_L = 1 - \frac{L_N}{L_S}. \quad (10)$$

2.5. Total weighting

Finally, the tonal factor μ_m of the mth frame of $x(n)$ is obtained as the combination of the Tonal Weighting (8) and the Loudness Weighting (10) through

$$\mu_m = \min\left(C_{1k60} W_T^{0.29} W_L^{0.79}, 1\right), \tag{11}$$

where C_{1k60} is a calibration constant that sets $\mu_m = 1$ for a pure tone of 60 dB SPL located at 1 kHz. In [24], this constant is set as $C_{1k60} = 1.09$ under ideal conditions. The exponents of W_T and W_L are correction factors introduced in the model according to empirical experiments [23].

Summarizing, the parameter values that have been modified in our proposed improved Aures (IA) method with respect to those used by the original Aures (OA) method are:

1. The frequency resolution in the OA method for any spectral analysis performed in the model was 12.5 Hz whereas in the new IA method is 10.76 Hz due to the use of a sampling frequency of $f_s = 44100$ Hz and an FFT size of 4096.
2. To compute (1) in the "Extraction Tonal/Noise" block, the tonal threshold of the OA method was $T_H = 7$ dB and the minimum tonal bandwidth was 75 Hz, whereas in the new IA method the tonal threshold is $T_H = 5.5$ dB and the minimum tonal bandwidth is 86.1 Hz.
3. The weighting function w_1 in (9a) exhibited an exponent of 1/0.29 in the OA method, whereas in the new IA method its exponent is 1.

3. Masking threshold model

The new IA tonality model will be validated through subjective tests that assess the perceived masking threshold of complex sounds. Their corresponding objective masking thresholds will be computed using the MTM proposed in [7,8] and shown in Fig. 2. It can be seen that the tonal factor μ_m affects the masking threshold $T_m(\nu)$ by means of the tonality offset $O_m(\nu)$. Therefore, in this section we will briefly explain the MTM shown in Fig. 2, although further explanation can be found in [7,8]. It can be seen that the same input $x_m(n)$ to Fig. 1 is now considered the masker signal. The upper branch describes the procedure to compute the auditory masking, $S_m(\nu)$, whereas the lower branch describes the method to estimate the tonality offset, $O_m(\nu)$. The input to both branches is the same power spectrum, $P_m(k)$, detailed in Section 2 estimated by the "Spectral Analysis" block. The final output $T_m(\nu)$ is the masking threshold for every Bark band ν and frame m and is calculated as

$$T_m(\nu) = S_m(\nu) - O_m(\nu), \quad \nu = 1, \ldots, N_c. \tag{12}$$

3.1. Auditory masking

The upper branch in Fig. 2 labeled as *Auditory Masking* obtains the overall masking curve, $S_m(\nu)$, of the mth frame of the masker signal $x(n)$ [7]. This process is based on the perceptual division of the audible spectrum, from 20 Hz to 20 kHz, into different critical bands carried out by the human hearing system [33]. The critical bands are expressed using Bark scale. In Fig. 2, the block "Mapping to Bark" computes the energy per critical band defined as [5]:

$$E_m(\nu) = \sum_{k=\inf(\nu)}^{\sup(\nu)} P_m(k), \tag{13}$$

where $P_m(k)$ is in linear units and $\inf(\nu)$ and $\sup(\nu)$ correspond to the frequency bin of the lower and the upper boundary of the Bark band ν, respectively.

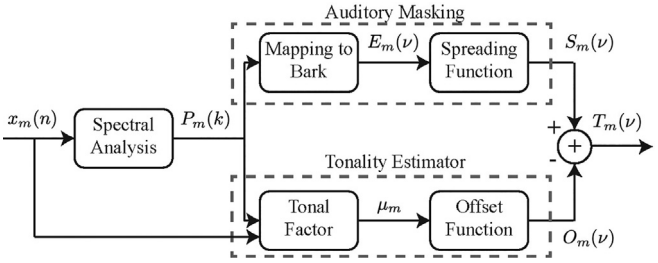


Fig. 2. Block diagram of the MTM to compute the frequency masking threshold $T_m(\nu)$.

The next block of the upper branch of Fig. 2 models the frequency masking effect of a single critical band over the rest of critical bands by means of a certain masking pattern, or *spreading function* [4,13]. The *spreading function* used here has been proposed in [4,5] and is expressed as:

$$B_\nu(\eta) = 15.81 + 7.5(\Delta_\nu(\eta) + 0.474) - 17.5 \times \sqrt{1 + (\Delta_\nu(\eta) + 0.474)^2}, \tag{14}$$

where ν is the maskee band, η is the masker band, $\Delta_\nu(\eta) = (\nu - \eta)$ and $B_\nu(\eta)$ is expressed in dB units. The value of $B_\nu(\eta)$ is used to compute the masking produced over the νth Bark band as

$$b_\nu(\eta) = 10^{\frac{B_\nu(\eta)}{10}} E_m(\eta), \quad \eta = 1, \ldots, N_c. \tag{15}$$

Once the masking pattern is applied to each critical band, their contribution is combined in order to obtain $S_m(\nu)$, based on the additivity of masking [2]. According to [15,35,36], the additivity of masking in the human hearing system is a complex operation that depends on several factors, such as the SPL value and frequency nature of the masker signal. We have investigated this issue in [8], where linear and non-linear combinations of different masking patterns were compared. The conclusions indicated that a linear summation of $b_\nu(\eta)$ (15) obtained similar masking values that those of non-linear combinations of alternative masking patterns. For this reason, we propose the following addition to obtain the overall masking curve (in dB units) as:

$$S_m(\nu) = 10\log_{10}\left(\sum_{\eta=1}^{N_c} b_\nu(\eta)\right), \quad \nu = 1, \ldots, N_c. \tag{16}$$

3.2. Tonality estimator

According to [2], the overall masking curve $S_m(\nu)$ suffers a decrease in energy that depends on the tonal characteristics of the masker signal, causing an asymmetry on the simultaneous masking. This effect is taking into account through the lower branch labeled as *Tonality Estimator* in Fig. 2. It is computed in two steps: The first step estimates the tonal factor μ_m, while the second step estimates the decrease in energy or tonality offset, $O_m(\nu)$.

A widely used tonal factor in the field of audio compression [37,38] was proposed by Johnston [5]. We use it here as an alternative to Aures' method and denote it as the Spectral Flatness (SF) method since it is based on the Spectral Flatness Measure (SFM) of a signal. The SF tonal factor is computed as

$$\mu_m = \min\left(\frac{\Upsilon_m}{\Upsilon_{1k60}}, 1\right), \tag{17}$$

where the SFM Υ_m is defined as the ratio between the geometric and arithmetic mean of $P_m(k)$ [5]:

$$\Upsilon_m = 10\log_{10}\left(\frac{\left[\prod_{k=1}^{N_{\mathrm{FFT}}} P_m(k)\right]^{\frac{1}{N_{\mathrm{FFT}}}}}{\frac{1}{N_{\mathrm{FFT}}}\sum_{k=1}^{N_{\mathrm{FFT}}} P_m(k)}\right), \tag{18}$$

and Υ_{1k60} is the SFM of a tone of 60 dB SPL located at 1 kHz. In [5], this constant is set as $\Upsilon_{1k60} = -60$ dB.

It can be seen that the SFM (18) ranges from a lowest level of $-\infty$ dB, which corresponds to a pure tonal signal, to a highest level of 0 dB corresponding to a white noise. Therefore, a masker signal would show a greater "whiteness" in its spectrum as its SFM level approaches 0 dB. This method has been slightly modified in [39] to provide a tonal factor per critical band. However, the method described by (17)–(18) has remained as the standard way to obtain the tonal factor of a masker sound [19,37,38,40].

As shown in Fig. 2, the estimated tonal factor μ_m is the input of the "Offset Function" block. According to [13], a pure tone signal located at the vth critical band suffers an offset of $(14.5 + v)$ dB over its masking curve $S_m(v)$, while the offset suffered by the overall masking curve of a white noise ranges between 3 and 6 dB, usually considered a constant value of 5.5 dB for the sake of simplicity [5,7]. This means that a pure tone provides a lower masking threshold $T_m(v)$ (12) compared to a white noise signal for the same energy $E_m(v)$ and Bark band. In the case of complex signals whose energy distribution along the Bark spectrum cannot be considered completely tone-like or noise-like, their offset (in dB units) can be estimated as [5]:

$$O_m(v) = \mu_m(14.5 + v) + (1 - \mu_m)5.5, \quad v = 1, \ldots, N_c. \tag{19}$$

4. Subjective experiments

A subjective test has been carried out in order to validate the performance of the different methods used to estimate the tonal factor in (19): IA, OA, and SF methods. This subjective test has been designed to evaluate the masking threshold of signals formed by a sum of pure tones in presence of a narrowband noise. For the purpose of validating the three methods of computing the tonal factor when dealing with complex sounds, the frequency masking threshold $T_m(v)$ has been computed for each method as the average value of (12) over all the time frames, and compared to the average masking threshold obtained in the subjective test. As a result, the averaged masking thresholds (MT) that will be referred to along this section are:

- $\widetilde{T}(v)$: MT obtained from the subjective test.
- $T^{\mathrm{SF}}(v)$: MT obtained by the SF method (17).
- $T^{\mathrm{OA}}(v)$: MT obtained by the OA method (11) (parameters as in [23]).
- $T^{\mathrm{IA}}(v)$: MT obtained by the IA method (11) (parameters as described in Section 2).

Fig. 2 shows that the computation of $S_m(v)$ does not depend on the tonal factor and, consequently, is a common procedure for all the MTs. Therefore, only the tonality offset, $O_m(v)$, computed according to the SF, OA or IA methods, will affect the masking threshold.

4.1. Generation of stimuli

The design of the stimuli for the subjective test involves two sets of sounds: multi-tone signals, which act as masker sounds, and narrowband noise signals acting as masked sounds. Each stimulus is formed by the combination of a multi-tone signal and a noise signal, as shown in Table 1. On the one hand, three multi-tone signals, each one composed by three different tones between 350 and 5800 Hz, have been used. On the other hand, nine narrowband noise signals have been generated with center frequency shown in Table 1 and a bandwidth of one critical band. All test signals are stationary, and thus, the average of the estimated masking threshold over time, $\widetilde{T}(v)$, can be considered unbiased for long enough frames. Both multi-tone and noise signals have been generated in Matlab with a sampling frequency of f_s = 44100 Hz and a duration of 5 s. During the test, they are presented in a loop (without any artifact between segments) in order to last as long as it takes until the user makes a decision. Table 1 shows that the chosen frequencies for both types of signals are the center frequencies of some specific critical bands. We have tried to use a wide part of the frequency spectrum, covering eight of the twenty-four critical bands, particularly those more sensitive according to the human auditory system.

There are only three different multi-tone signals, but they are used three times, one per narrowband noise centered at their tones' frequencies. Table 1 shows dashed lines to separate groups of stimuli with different multi-tone signals. The first multi-tone signal is formed by three tones lying in consecutive critical bands, the seventh, eighth and ninth, whereas the other two multi-tone signals are formed by three tones spread along the frequency spectrum. In this way, the design of the subjective test takes into account a diverse set of complex signals regarding whether their energy is more or less spread along the frequency spectrum. The 1000-Hz tone has been used in two multi-tone signals, once as the highest tone and once as the lowest, in order to analyze the masking produced on a single critical band by their adjacent (pre and post) critical bands.

Each of the tones of the multi-tone signals has been generated as a sine wave with unit amplitude and zero initial phase. The narrowband noise signals have been generated as the output of band-pass filters centered at the frequencies shown in the second column of Table 1, being their inputs white Gaussian noise. Their stop bands lie at the center frequencies of their adjacent critical bands, thus, each filter having a bandwidth of one critical band. Fig. 3 shows the power spectrum of the 8th stimulus of Table 1. The noise power has been chosen to match the peak power of the tone located at 2150 Hz, since the real noise power is decided by the listeners in the test. As it can be noticed, the noise power rapidly decays outside the 14th critical band. The frequency distribution of the rest of the stimuli is similar but using the corresponding critical bands shown in Table 1.

4.2. Apparatus and design

The perceptual test was carried out inside the listening room of the Audio Processing laboratory[1] of the Institute of Telecommunications and Multimedia Applications, whose reverberation time is 0.18 s. The reproduction system was formed by an M-Audio M-Track Quad soundcard and a pair of Sennheiser HD 600 headphones connected to a laptop.

The reproduction system was calibrated such that the headphones were placed on a Neumann KU100 dummy head and a tone of 1 kHz was reproduced through the headphones. Then, the reproduction system was adjusted to provide 60 dB SPL at the dummy head microphones, resulting in a digital amplitude of -14.7 dB with respect to the same tone ranging the full digital scale of $[-1, 1]$. Additionally, the multi-tone signals, shown in Table 1, have

[1] https://gtac.webs.upv.es

Table 1
List of stimuli generated for the subjective test.

Stimulus #	Center Frequency (Hz)		Critical Bands
	Multi-tonal signal	Noise signal	
1	700, 840, 1000	700	7, 8, 9
2	700, 840, 1000	840	
3	700, 840, 1000	1000	
4	350, 450, 5800	350	4, 5, 20
5	350, 450, 5800	450	
6	350, 450, 5800	5800	
7	1000, 2150, 3400	1000	9, 14, 17
8	1000, 2150, 3400	2150	
9	1000, 2150, 3400	3400	

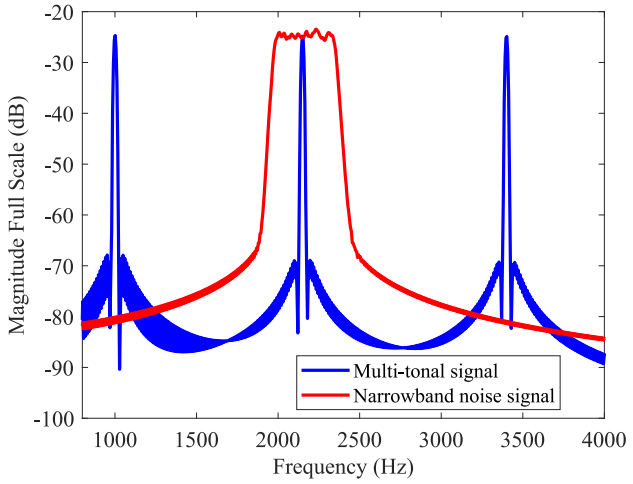


Fig. 3. Power spectrum of multi-tone signal (blue) and the narrowband noise signal (red) of the 8th stimulus.

been weighted to match their loudness with the loudness of the tone signal used for calibrating the system in order to make the perceptual test as comfortable as possible. Once the system was calibrated, Υ_{1k60} in (17) must be computed as the SFM (18) of a tone of 60 dB SPL located at 1 kHz, resulting in $\Upsilon_{1k60} = -54$ dB.

4.3. Participants

Nineteen people participated in the subjective evaluation. Before starting the subjective test, an audiometry was carried out to each participant using the same reproduction system available for the test. As the tests were carried out using Matlab, an ad hoc application was also programmed in the same software. The audibility curves of three subjects did not present normal hearing thresholds, so only the assessments obtained from 16 subjects were considered for the results. As a result, the jury panel was formed by seven male and nine female aged between 18 and 50 years, with the median value 27 years. Only three of them had previous knowledge on psychoacoustics.

4.4. Procedure

Once the stimulus is selected from Table 1, the perceptual test aims to obtain the masking threshold $T(\nu)$ of the corresponding multi-tone signal at the specific critical band where the narrowband noise is located. The presentation of the stimulus is such that the multi-tone signal is played with a constant SPL while the noise level is increased or decreased by the subject according to the method described in the following.

The interface of the ad hoc application designed in Matlab to carry out the perceptual tests is shown in Fig. 4. At its bottom left, the user can select the stimulus to be played, labeled as "Signal 1" in the figure. At its bottom center, the two buttons can increase ("+") or decrease ("-") the SPL of the corresponding narrowband noise. The application also provides two examples for training the subject, which can be selected through the buttons at the upper part of the interface. The examples contain two different multi-tone signals, each one presented alone ("clean") or mixed with an audible narrowband noise ("dirty"). None of the examples has been used or is similar to the true stimuli of the test. The subject employs the Matlab application during the entire duration of the test, but there is always an expert in the room in order to solve any question related to the application.

The flow diagram of the subjective test is shown in Fig. 5 and is described in the following:

1. The subject starts the test by selecting one of the nine stimuli from the list at the bottom left of Fig. 4. The stimuli correspond to those listed in Table 1 and are labeled "Signal 1" through "Signal 9".
2. The subject must then press "Play Signal" to play the selected stimulus. The multi-tone signal is presented with a constant SPL, while the noise is presented with an attenuation of 40 dB with respect to the SPL of a noise whose loudness would be equal to that of the multi-tone signal. In this way, we ensure that the noise is completely masked.
3. The participant then presses the plus button ("+") as many times as necessary to make the noise audible. At this point, the SPL of the noise increases in 5 dB steps each time the "+" button is pressed.
4. Once the participant is able to distinguish the noise, he or she presses the minus button ("-") as many times as necessary to make the noise inaudible again. The first time the "-" button is pressed, the noise level is reduced by 3 dB. The second and subsequent times the "-" button is pressed, the level values decrease in 2 dB steps. This reduction in step size as the number of selections increases is a common technique in subjective tests aimed at estimating a threshold [15,41].
5. The participant iteratively repeats steps 3 and 4 so that he or she is able to unmask (hear) the noise through step 3 and mask it through step 4. At this point, the increase and decrease of the noise level is always carried out by 2 dB steps. The loop ends when one of these two conditions is met:
   (a) The participant has pressed any "+" or "-" button 20 times.
   (b) The participant presses the sequence "- + -" or the alternative sequence "+ - +". These sequences indicate that the noise level has reached the masking threshold of the multi-tone signal within a margin of ± 2 dB.

   At this point, the number of clicks used to finish this step is saved for each participant and stimulus.
6. Once the evaluation of the first stimulus is finished, the noise signal is saved and the application allows the selection of a new stimulus by returning to step 1.
7. Once the participant has evaluated all the stimuli, he or she must click the "Finish Test" button located at the bottom right of Fig. 4.

Apart from the noise signal, the sequence of "+"/"-" clicks for each participant and stimulus has been saved and analyzed. This information has revealed that 12 out of 16 participants have finished two stimuli with 20 clicks, which represents 16% of the tests. Therefore, most of the masking thresholds were precisely identified within a margin of ± 2 dB.

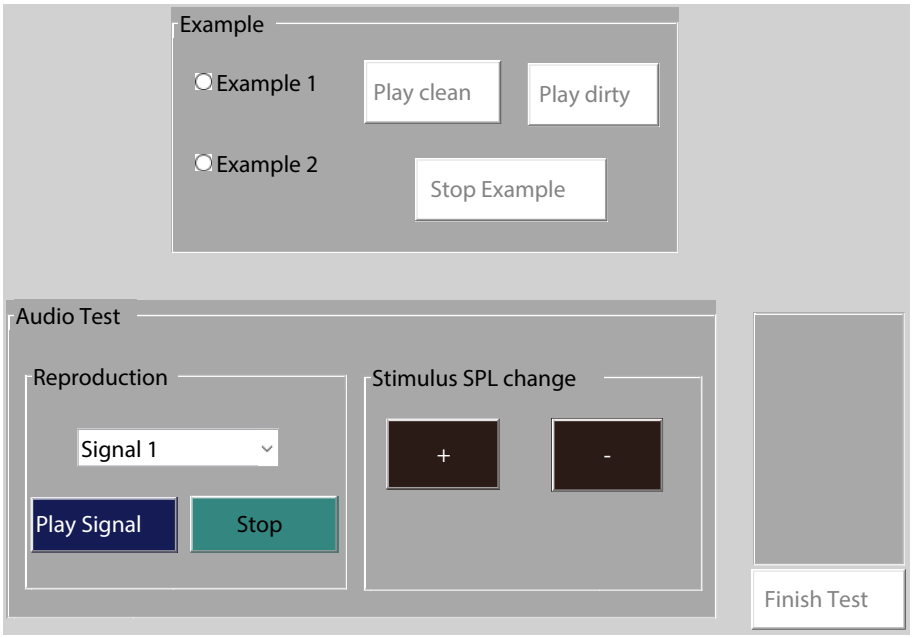


Fig. 4. Matlab interface specifically designed to run the subjective tests.

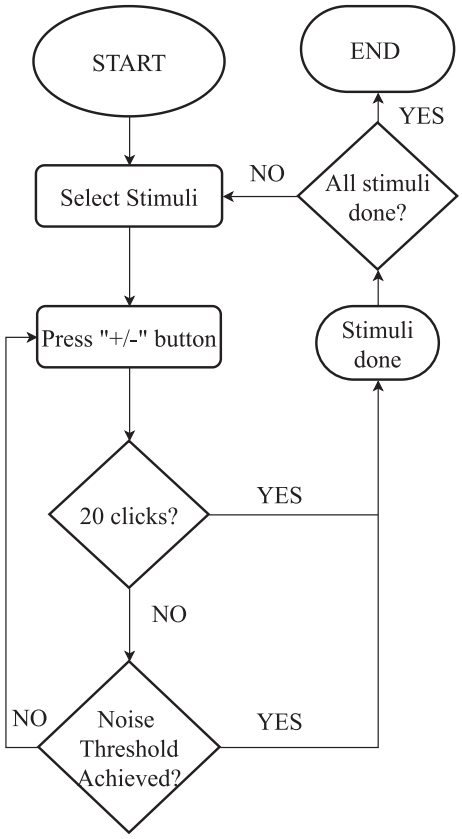


Fig. 5. Flow diagram of the subjective test.

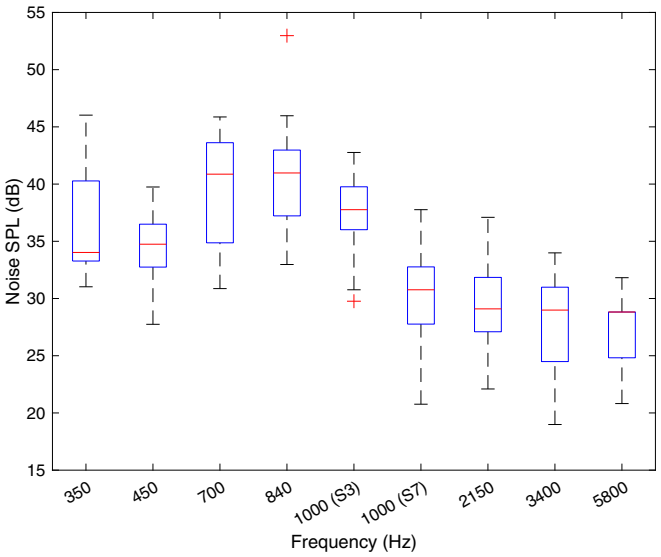


Fig. 6. Noise SPL obtained for each stimulus vs the center frequency of the corresponding noise critical band. Notice that stimuli "S3" and "S7" with the same noise critical band present different MT levels due to their different multi-tone signals.

4.5. Results and discussion

The SPL of the noise signal obtained in step 6 of the test for each stimulus and participant has been considered as the subjective masking threshold produced by that multi-tone signal over the critical band covered by the corresponding noise. Fig. 6 shows the box plot of the noise SPL obtained for each stimulus vs the center frequency of the corresponding noise critical band. The central red mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the two outliers are represented by red crosses. It can be noticed that the SPL obtained to mask the noise signal centered at 1000 Hz is 7 dB lower when the corresponding tone is accompanied by higher frequency tones (S7) than when is presented with tones of lower frequencies (S3).

Therefore, the mean SPL over the 16 participants will be considered as the subjective masking threshold (SMT) obtained from the

test for each stimulus. The mean SMT will be denoted by $\widetilde{T}^{(f_n)}$, where f_n indicates the center frequency of the corresponding narrowband noise in Table 1. The spectral distribution of each noise signal over the Bark scale has been computed as the averaged energy per Bark (13), and it is denoted by $\widetilde{T}^{(f_n)}(v)$. Its maximum value coincides with the SMT obtained in the test.

Each group of SMTs $\widetilde{T}^{(f_n)}(v)$ is compared to the MTs obtained by the different methods in Fig. 7. Fig. 7(a) shows the MT curves obtained by the SF, IA and OA methods for the multi-tone signal used in the first three stimuli of Table 1, whereas Fig. 7(b) and and Fig. 7(c) show the MT curves of the multi-tone signals used in the second and third set of stimuli. Additionally, each figure represents the spectral distribution of the three masking thresholds obtained from the perceptual test labeled as $\widetilde{T}^{(f_n)}(v)$. It can be noticed that T^{SF} provides a lower masking level than the MTs obtained in the perceptual test for the three multi-tone signals, especially in the range of middle and high frequency. In contrast, the OA and IA models based on Aures' method usually obtain masking levels above the subjective masking thresholds, although IA is always closer to the perceived levels.

Fig. 8 shows the SMT values obtained in the test with respect to the predicted values obtained by each method. Nine values per method are shown, corresponding to each of the stimuli shown in Table 1. SF method is represented by brown diamonds, IA method by blue circles and OA method by red squares. Additionally, the line $y = x$ is also plotted in order to compare the accuracy of the prediction for each method. The legend shows the mean error value for each method computed as:

$$\varepsilon^{\mathrm{method}} = \frac{1}{9}\sum_{n=1}^{9}\left(T^{\mathrm{method}}(f_n) - \widetilde{T}^{(f_n)}\right). \tag{20}$$

Results shown in Fig. 8 confirm that the MT values provided by the SF method are always smaller than the SMT (all of them fall in the region $x < y$), whereas the OA method presents the opposite behav-

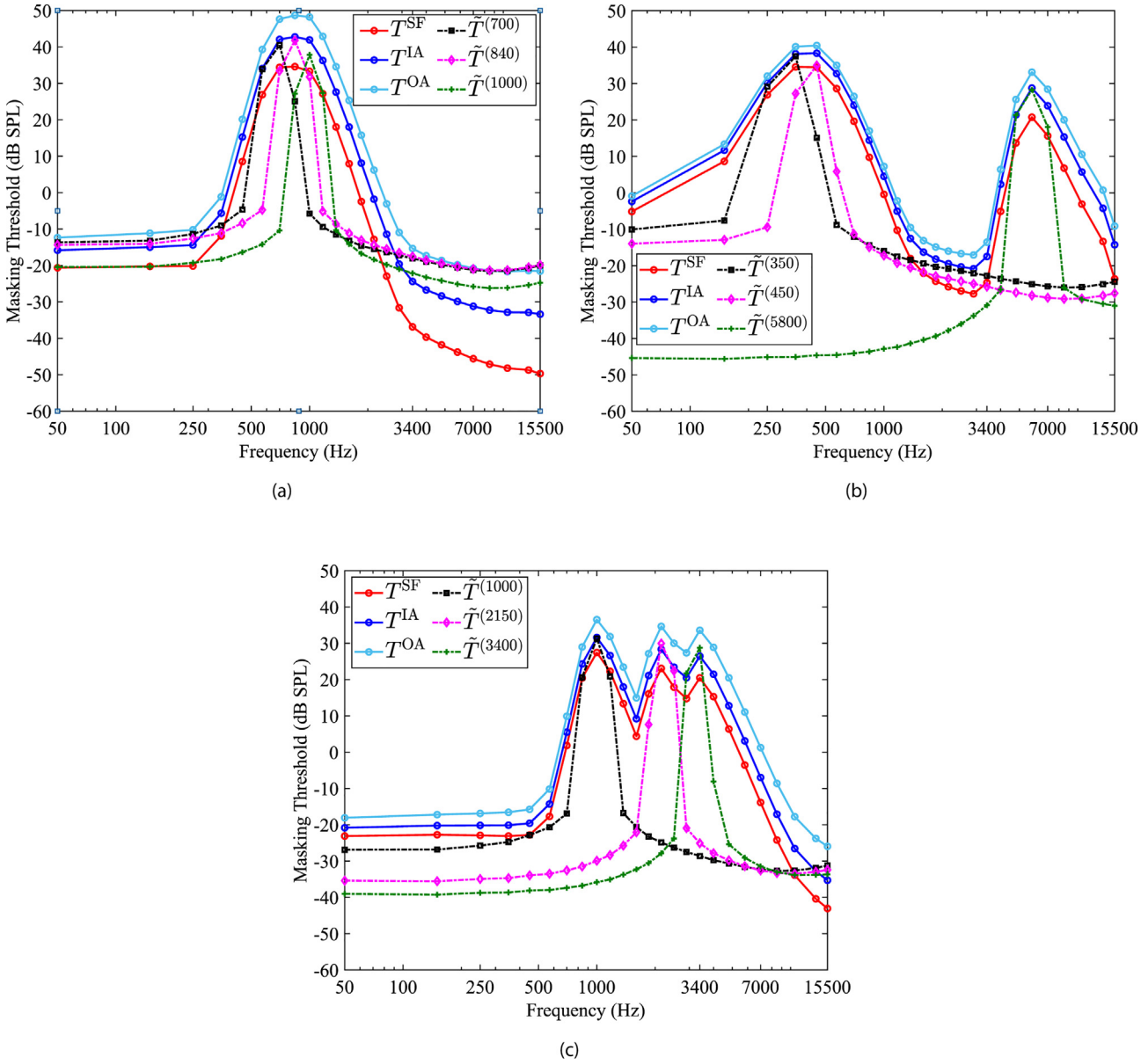


Fig. 7. Masking thresholds obtained by the SF (T^{SF}), the original (T^{OA}) and the improved (T^{IA}) Aures methods compared to the masking thresholds obtained in the subjective test for (a) the first multi-tone signal, (b) the second multi-tone signal and (c) the third multi-tone signal in Table 1.

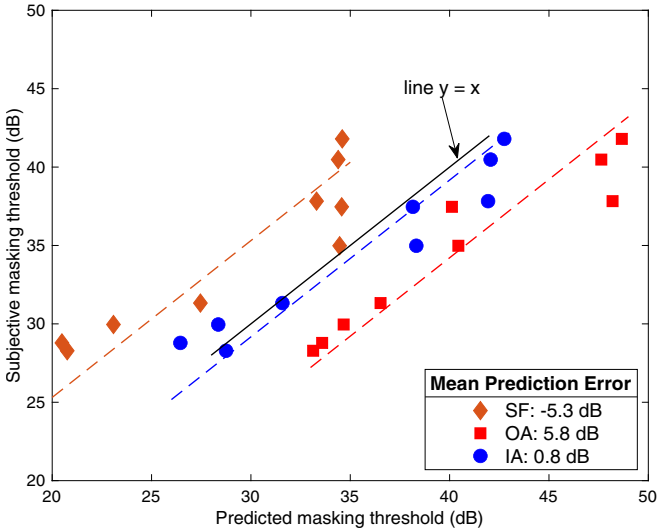


Fig. 8. SMT values obtained in the test with respect to the predicted values obtained by the SF (brown diamonds), OA (red squares) and IA (blue circles) methods. The black line represents $y = x$, whereas the other three lines represent $y = x - \varepsilon$, where ε is the mean prediction error shown in the legend.

ior and all their values are in the region $x > y$. Regarding the proposed IA method, their values are distributed within both sides of the $x = y$ curve, meaning that IA can predict more accurately the SMT of multi-tone signals than the SF and OA methods. Fig. 8 also shows three additional lines corresponding to $y = x - \varepsilon$, where the ε values (20) are shown in the legend. The brown line corresponds to the SF method, and the blue and red lines to the IA and OA methods respectively. It can be noticed that the methods that use Aures tonality (OA and IA) achieve a better distribution of the SMT if the error could be compensated. Regarding the higher MT values in particular, which corresponds to the most sensitive range of $[700, 1000]$ Hz in Fig. 6, the SF method presents a very poor prediction, providing an equal value of 34 dB for SMTs that range from 34 to 43 dB.

Since the overall masking curve $S_m(v)$ used in the MTM of Fig. 2 is the same for the three methods, we can state that the lower values in the SF masking threshold are produced by the overestimation of the tonal factor μ_m, which, in turn, increases the offset $O_m(v)$ (19) at those critical bands. On the other hand, the OA method underestimates μ_m, especially for middle frequencies identified as the red squares located at the highest values in the x-axis of Fig. 8. They present the highest error, visualized as the path length if ones moves from the red square towards the black line along the x-axis.

In contrast to the SF and OA methods, the IA method provides values slightly above of the $\widetilde{T}^{(f_n)}$ levels with a mean prediction error of $\varepsilon = 0.8$ dB. Its highest error (-4 dB) corresponds to the third stimulus, when the noise is masked by the 1000 Hz component. However, the SMT of the same frequency but for the seventh stimulus is predicted with an error of less than 1 dB.

Finally, the Pearson correlation coefficient R between the SMT and the predicted MT has been computed for each method, resulting in $R = \{92\%, 95.4\%, 96.8\%\}$ for the SF, OA and IA methods respectively. Their corresponding p-values and mean squared errors (MSE) are $\{4.3, 0.7, 0.2\} \cdot 10^{-4}$ and $\{2.55, 2.16, 1.33\}$ dB respectively, which confirm the statistical relevance of the three methods to predict the subjective masking threshold if the bias (ε) presented by the SF and OA methods could be compensated. However, we can conclude that only the proposed IA method, although being a slightly modified version of the original Aures method, provides accurate prediction of the masking threshold of multi-tone sounds in the presence of narrowband noise signals.

5. Perceptual equalization

A second subjective test has been carried out to validate the proposed tonality model when other complex signals, as music or speech, are involved. For this purpose, we have considered a scenario where an audio signal is played by a loudspeaker in presence of an undesired ambient noise. This scenario has been implemented in our laboratory as shown in Fig. 9, where the audio signal is emitted by the loudspeaker labeled "Spk 1", the ambient noise is emitted by the loudspeaker labeled "Ambient noise" and the microphone "Mic 1" captures the audio plus noise signal as it would be heard by a person seated there.

In this scenario, the perception of the audio signal may be severely impaired by the added ambient noise. Therefore, audio equalization is considered in order to boost the audio signal above the noise [42,43]. Since the equalizer levels depend on the perceived ambient noise, this kind of process is known as perceptual equalization [7,19,40]. We use here the perceptual equalizer that was implemented in our previous work using an acoustic node [8]. An acoustic node is a device formed by a microphone, a loudspeaker and a processing unit that can also communicate with other nodes within a network. For our experiment here, the set formed by the microphone "Mic 1" and the loudspeaker "Spk 1" is labeled "Node 1" in Fig. 9, meaning that they form an acoustic node able to process the signal captured by the microphone and generate the signal emitted by the loudspeaker.

The model of the discrete-time signal recorded by the microphone can be expressed as

$$x(n) = a(n) + z(n) = c(n) * s(n) + z(n), \tag{21}$$

where $c(n)$ is the electroacoustic path between loudspeaker "Spk 1" and microphone "Mic 1", which is modeled as a finite impulse response filter of L coefficients, $s(n)$ is the audio signal emitted by the loudspeaker, $(*)$ is the convolution operation, and $z(n)$ is the ambient noise at the microphone location. Assuming that $c(n)$ and $s(n)$ are known ($c(n)$ is usually estimated in a pre-set stage [44]), then the ambient noise $z(n)$ can be obtained as $z(n) = x(n) - c(n) * s(n)$.

5.1. Equalizer gains

Once the audio signal $a(n)$ and the ambient noise $z(n)$ are separated from the recorded signal $x(n)$ in (21), the processing unit of the acoustic node will compute the equalizer gains for every critical band according to the methodology described in Section III of [8], and considering the following two profiles of the perceptual equalizer.

5.1.1. Unmasked audio signal profile

The unmasked audio signal (UAS) profile aims to prevent the audio signal to be masked by the ambient noise [7], that is, its goal is to get the audio signal unmasked by the noise. Consequently, the equalizer gains are designed such that the energy of the audio signal is set above the masking threshold of the ambient noise. To this end, the equalizer gains are estimated as

$$g(v) = \max(T_z(v) - E_a(v), 0), \tag{22}$$

where $g(v)$ is the gain (in dB) of the critical band v such that $g(v) \geqslant 0$, $T_z(v)$ is the masking threshold (12) of the recorded ambient noise $z(n)$, and $E_a(v)$ (13) is the energy per critical band of the recorded audio signal $a(n)$, both expressed in dB.

5.1.2. Masked noise profile

The masked noise (MN) profile is intended to mask the ambient noise by increasing the energy of the audio signal [8]. To this end, the equalizer gains are estimated as

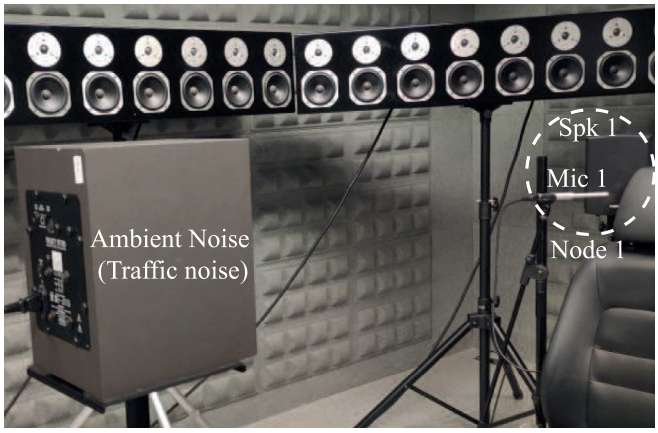


Fig. 9. Testbed for the validation of tonality models involving a perceptual equalizer of audio signals in the presence of ambient noise.

$$g(\nu) = \max(E_z(\nu) - T_a(\nu), 0), \tag{23}$$

where $E_z(\nu)$ is the energy per critical band of the recorded ambient noise $z(n)$ and $T_a(\nu)$ is the masking threshold of the recorded audio signal $a(n)$, both expressed in dB SPL. Analogous to the gains obtained in (22), also for this profile $g(\nu) \geqslant 0$.

5.1.3. Refinement of the equalizer gains

Additionally, to prevent saturation in the playing of the audio signal, $g(\nu)$ has been limited to 15 dB for all the critical bands. Furthermore, for each time-frame, the gains are set to 0 dB in those critical bands with very low energy, $E_a(\nu) \leqslant 0$ dB SPL.

As a last step in the equalizer design, a time averaging of the gain levels is carried out to provide a smoother transition between time frames:

$$g^{\text{av}}(\nu) = \xi g(\nu) + (1 - \xi) g^{\text{av}}(\nu), \tag{24}$$

where $g^{\text{av}}(\nu)$ is the averaged gain level (in dB) and ξ is a smoothing constant, which is set to a value of 0.2 when $g^{\text{av}}(\nu) > g(\nu)$ and a value of 0.7 otherwise. When the noise level increases, the equalizer adapts fast ($\xi = 0.7$) to mask the noise as soon as possible. In contrast, when the noise level decreases, the equalizer adapts slow ($\xi = 0.2$) to provide a smooth transition to the required gains, avoiding the annoying effect of continuous ups and downs in the audio level.

Summarizing, the equalizer gains are computed for each critical band according to the following two profiles: Unmasked Audio Signal (UAS) whose gains are computed by means of (22) and (24), and Masked Noise (MN), whose gains are computed by means of (23)–(24). Once the gains $g^{\text{av}}(\nu)$ have been computed, the audio signal $s(n)$ emitted by the loudspeaker is pre-equalized using the graphic equalizer proposed in [45,46] at each critical band ν.

5.2. Subjective test on perceptual equalization

The aim of the subjective test is to evaluate the performance of the SF and IA tonal factors in the masking threshold model when using more complex sounds than multi-tone sounds and narrow-band noise signals. For this purpose, we have carried out a perceptual pre-equalization of the audio signal emitted by loudspeaker "Spk 1" in Fig. 9 and we have recorded the audio signal contaminated by the noise at the position of "Mic 1". As seen before, the masking threshold is key in the estimation of the equalizer gains for both UAS and MN profiles. Their common goal is to find an equalization profile such that the noise signal is not perceived, or it is perceived as low as possible given the limitations in the gains' values.

5.2.1. Generation of the stimuli

The stimuli of the test are composed of five recorded signals:

- "MN-IA" and "MN-SF": Recorded signals when the MN strategy is used and the masking threshold of the audio signal $T_a(\nu)$ in (23) is computed using the IA (11) or the SF (17) methods.
- "UAS-IA" and "UAS-SF": Recorded signals when the UAS strategy is used and the masking threshold of the noise signal $T_z(\nu)$ in (22) is computed using the IA (11) or the SF (17) methods.
- "NONE": Recorded signal when no equalization is performed.

The audio signal is an excerpt of the song "Tell me something good" by Chaka Khan and the ambient noise is a traffic noise extracted from YouTube[2]. Both signals are sampled at $f_s = 44100$ Hz with a duration of 30 s.

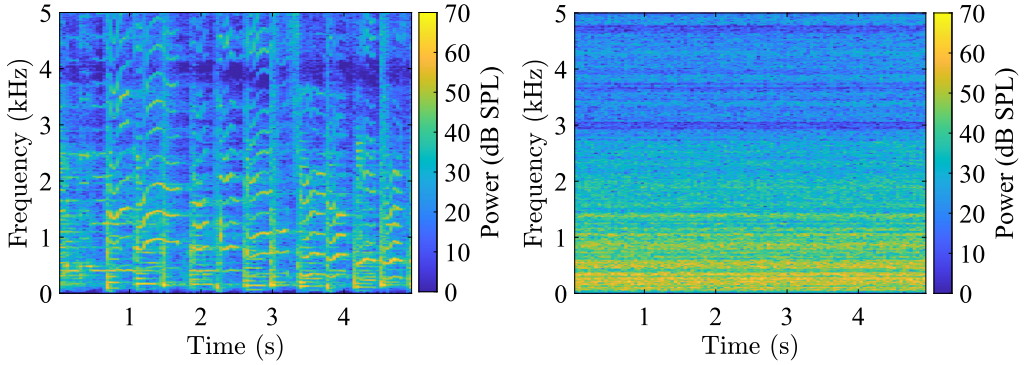
To produce the stimuli, the real-time acoustic responses of the system shown in Fig. 9 have been measured and the microphone of the real system has been calibrated. Then the five signals $x(n)$ (21) for the different equalization profiles have been simulated and converted to stereo signals. Although the original duration of the audio and noise was 30 s, the stimuli have been formed by selecting a segment of 5 s, in particular from the sixth to the eleventh second where the song presented remarkable level variations. The song and noise spectrograms of the segments are shown in Fig. 10, where only up to the frequency 5 kHz is shown since the traffic noise decreases significantly its power level above that frequency.

According to Fig. 10(b), the traffic noise presents an stationary behavior. In addition, it exhibits the highest power levels at frequencies below 1 kHz in contrast to the audio signal, and its power level decreases as the frequency increases, specially for frequencies above 3 kHz. Regarding the spectrogram of the audio signal shown in Fig. 10(a), it presents a non-stationary behavior and a harmonic structure that can be easily identified.

The gain levels computed for each profile over time are shown in Fig. 11 for the $3^{rd}, 10^{th}, 14^{th}$ and 19^{th} critical bands. The first two bands lie in the frequency range where the noise level is higher (centered at 250 Hz and 1 kHz respectively), whereas the 14^{th} band is centered at 2 kHz, where the music level is higher, and the 19^{th} critical band is centered at 4 kHz, where both music and noise levels are low. It can be seen from the four bands in Fig. 11 that "MN" profiles provide higher gains to the audio signals for both SF and IA methods than the "UAS" profiles. Moreover, the gains of the 3^{rd} critical band for the "MN-SF" profile are set to the maximum level of 15 dB all the time. On the other hand, the "UAS-SF" profile for high frequencies obtain gains close to 0 dB, meaning that the music is getting "unmasked" by the noise without any help. Comparing between tonality methods SF and IA within the same profile, Fig. 11 shows IA method introduces lower gains than the SF method for the "MN" profile, whereas the opposite behavior is observed for the "UAS" profile.

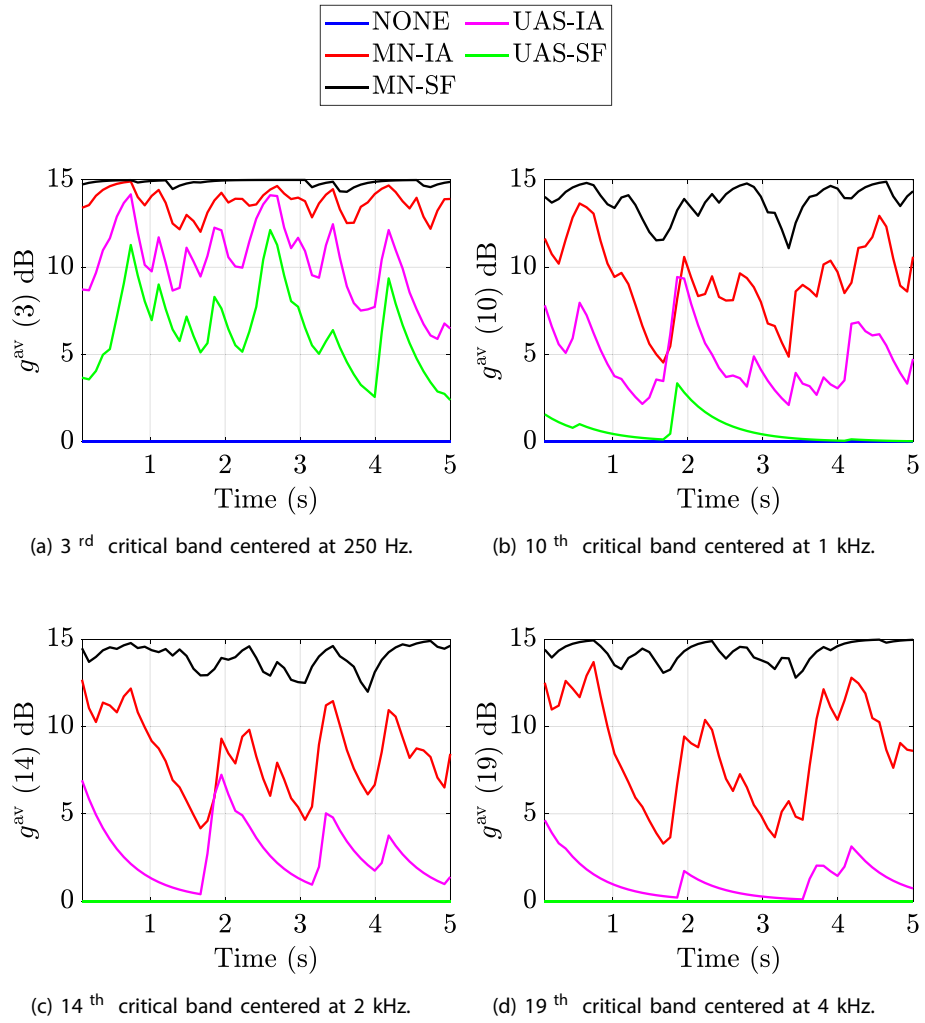
Fig. 12 shows the mean power spectrum of the music emitted by loudspeaker 1 for the different equalization methods, together with the original power spectrum labelled as "NONE". Although the masking signal is not stationary, see Fig. 10(a), its mean power spectrum presents the same behaviour as seen in Fig. 11: MN SF and IA profiles boot the energy of the music more than 10 dB with respect to their corresponding UAS profiles, especially for the frequency range above 1 kHz. Comparing SF and IA methods, the dif-

[2] https://www.youtube.com/watch?v=fh3EdeGNKus



(a) Excerpt of the song "Tell me something good". (b) Traffic noise signal.

Fig. 10. Spectrograms of the original signals as recorded at the microphone position.



(a) 3 rd critical band centered at 250 Hz. (b) 10 th critical band centered at 1 kHz.

(c) 14 th critical band centered at 2 kHz. (d) 19 th critical band centered at 4 kHz.

Fig. 11. Time variation of the gain values for each profile.

ference between the two curves obtained with SF-based equalizer gains (black and green lines) is greater than 10 dB except for the low frequency range. However, the IA-based spectra present a shift between methods in the range of $[2-8]$ dB. Therefore, we can conclude that IA method has a more stable behavior with respect to the type of selected profile.

5.2.2. Apparatus and design

The perceptual test was carried out in the same room as the tonality experiment of Section 4.2. The same setup has also been considered: an M-Track Quad sound card and a pair of Sennheiser HD 600 headphones connected to a laptop. The unprocessed audio signal has been weighted in loudness as for the tonality experi-

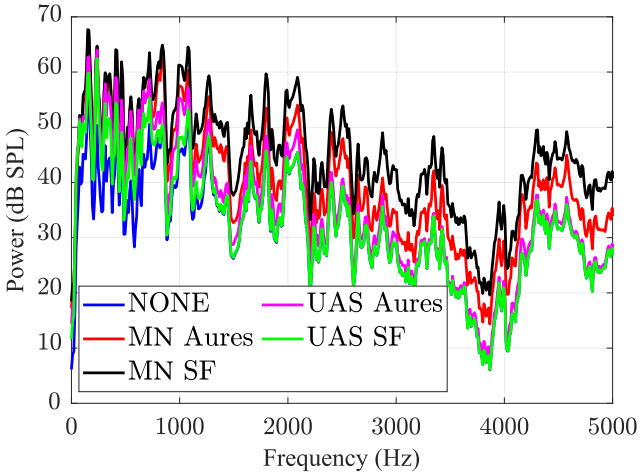


Fig. 12. Power spectrum of the audio signal for the different profiles used in the test.

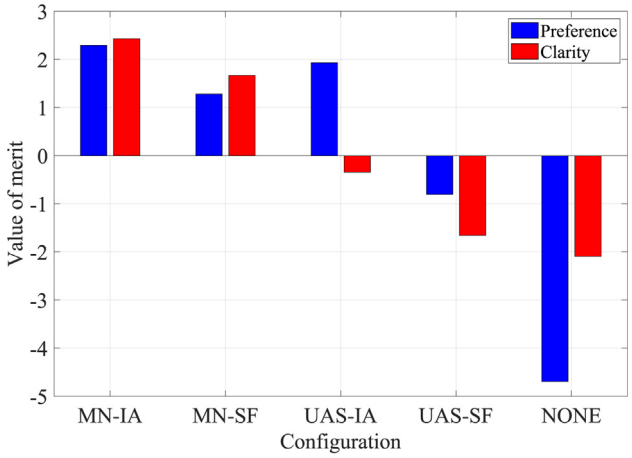


Fig. 13. Values of merit obtained by the subjective test on perceptual equalization.

ment (see Section 4.2), while the traffic noise has been weighted in order to provide an SNR of −3 dB regarding the audio signal when no equalization is performed.

5.2.3. Participants

The perceptual test was carried out by 13 participants, six males and seven females, aged between 20 and 40 years. All of them reported normal hearing and five of them were familiarized with the psychoacoustic area of research. All of the juries presented at least a 75% of repetitiveness and a 95% of consistency in their choices, thus, all the jury members were included in the test analysis.

5.2.4. Procedure

The subjective test was a paired comparison test [47] where two stimuli are presented at each step and their "clarity" and "preference" were evaluated by the jury. The subjects were asked to choose the "clearer" audio signal and they have to choose which is their preferred stimulus as well. For this purpose, before starting the test, each jury was invited to listen to several examples of an audio signal contaminated by noise, which were different from those used for the test.

The subjective test has been performed through an ad hoc designed application implemented in Matlab. The selected comparisons are shown in Table 2 marked with an "X", resulting in a total of twelve comparisons to be evaluated. As Table 2 shows, some combinations have been discarded in order to shorten the test and avoid the jury to get annoyed. Repeated comparisons as, for instance, "MN-IA" versus "MN-SF" and vice versa allow to evaluate the reliability of the jury.

5.2.5. Results and discussion

The values of merit for each combination of profile and method regarding the "clarity" (red bars) and "preference" (blue bars) are shown in Fig. 13. Notice that the values of merit for each characteristic must sum up to 1, and the greater the differences between the positive and negative values, the greater the agreement among the jury panel, i.e., most of the jury decided in the same way.

Regarding the "clarity" of the music signal in the presence of noise, the "MN" profiles are preferred with respect to the "UAS" profiles, which is an expected result taking into account that their gain levels are also higher as it is shown in Fig. 11. However, this logic does not apply to the comparison between the "MN-IA" and "MN-SF" methods, where IA is preferred even though a lower gain level is used throughout the entire duration of the music. Therefore, we can conclude that the "MN-IA" method better follows the effect of the tonality on the masking threshold of the audio signal, $T_a(\nu)$, than the "MN-SF" method. In this sense, the high gains provided by the "MN-SF" method may produce an annoying effect of increasing the music level without the required control, damaging in some way the perceived "clarity" of the audio signal.

Regarding the values of merit of the "preference" in Fig. 13, the IA method is preferred independently of the profile used to equalize the audio signal, although the "MN-SF" profile obtains a value of merit very similar to that of IA. Once again, a plausible explanation can be stated looking to the gain variations of the IA method shown in Fig. 11. It can be appreciated how the gain curves of the IA method have excursions of 6-7 dB between their minimum and maximum values, whereas the gain curves of the SF method presents excursions of 3-4 dB in the best case and flat curves for "MN" in low frequency bands and for "UAS" in high frequency bands. Therefore, the SF method is active only in the middle range of frequencies, whereas the IA method is able to control the equalizer gains along all the frequency range.

Therefore, from the results obtained in the subjective test, the IA method has proved to be more adequate than the SF method to compute the MT of the audio signal $T_a(\nu)$ in the "MN" strategy, and the MT of the noise $T_z(\nu)$ in the "UAS" strategy. The IA method has better respected the clarity of the music signal and has been selected as the preferred method independently of the strategy. Thus, we can conclude that the proposed improved Aures method to compute the tonal factor can model the tonality offset of complex sounds better than the two previous methods.

Table 2
Comparisons carried out in the perceptual test are marked (X).

MN-SF		X			X
MN-IA	X		X	X	
UAS-SF				X	X
UAS-IA	X	X	X		
NONE		X		X	
	MN-SF	MN-IA	UAS-SF	UAS-IA	NONE

6. Conclusions

An improved version of the original Aures (OA) model to estimate the tonal factor of the masking threshold has been introduced and compared to other models. The perception of the tonality has been extensively studied in simple sounds, but few studies have been carried out on complex sounds as multi-tone or music signals.

In this work, we have proposed the improved Aures (IA) method that includes minor changes of the OA from the mathematical point of view, but very relevant for the perception of the masking effect as the results of the subjective tests have shown.

As the tonality influences the masking threshold of a sound, we have evaluated in the first test the subjective masking of a multitone signal in the presence of a narrowband noise. This test has been designed to assess the accuracy of the proposed method (IA) in comparison to the OA and to the state-of-the-art method proposed by Johnston and denoted here by "spectral flatness" (SF) method. The results have shown that the proposed IA method presents a mean error of 0.8 dB compared to the masking threshold provided by the subjective test, while the mean error for the OA and SF methods were 5.8 and −5.3 dB, respectively. Thus, the best fit between the perceived and the estimated masking has been obtained by the new IA method, validating their accuracy for signals with multiple tonal components.

A second subjective test has been carried out to validate the IA method through the perceptual equalization of audio signals. This test has evaluated the "clarity" and "preference" of a music signal in the presence of a broadband noise when music has been pre-equalized according to the masking effect. In this second test only the new IA and the SF methods were compared, but two different strategies were used to compute the equalizer gains: one involving the masking threshold of the music and the other involving that of the noise. For both strategies, the equalizer based on the new IA method provided a "clearer" experience of the music and was preferred by the jury, validating the good accuracy of the proposed method to estimate the tonality offset of complex sounds.

CRediT authorship contribution statement

Juan Estreder: Software, Investigation, Validation, Data curation, Writing - original draft. **Gema Piñero:** Conceptualization, Formal analysis, Methodology, Supervision, Writing - review & editing. **Maria Diego:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Jussi Rämö:** Supervision. **Vesa Välimäki:** Conceptualization, Formal analysis, Supervision, Writing - review & editing.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially supported by the GVA Regional Government through PROMETEO/2019/109, the Spanish Government through BES-2016–077899 and RED2018-102668-T, and the European Union together with the Spanish Government through RTI2018-098085-BC41 (MCIU/AEI/FEDER). This research is also part of the activities of the "Nordic Sound and Music Computing Network" (NordicSMC), NordForsk project No. 86892. Juan Estreder's work has been carried out partly during his research stay at the Acoustics Lab of Aalto University.

References

[1] Fletcher H. Auditory patterns. Rev Mod Phys 1940;12(1):47. https://doi.org/10.1103/RevModPhys.12.47.
[2] Zwicker E, Fastl H. Psychoacoustics: Facts and models. second updated ed. Springer Science & Business Media, Heidelberg; 1990.
[3] Donley J, Ritz C, Kleijn WB. Multizone soundfield reproduction with privacy- and quality-based speech masking filters. IEEE/ACM Trans Audio Speech Lang Process 2018;26(6):1041–55. https://doi.org/10.1109/TASLP.2018.2798804.
[4] Schroeder MR, Atal BS, Hall J. Optimizing digital speech coders by exploiting masking properties of the human ear. J Acoust Soc Am 1979;66(6):1647–52. https://doi.org/10.1121/1.383662.
[5] Johnston JD. Transform coding of audio signals using perceptual noise criteria. IEEE J Sel Areas Commun 1988;6(2):314–23. https://doi.org/10.1109/49.608.
[6] Brandenburg K, Stoll G. ISO/MPEG-1 audio: A generic standard for coding of high-quality digital audio. J Audio Eng Soc 1994;42(10):780–92.
[7] Rämö J, Välimäki V, Tikander M. Perceptual headphone equalization for mitigation of ambient noise. In: Proc. 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada. p. 724–8. https://doi.org/10.1109/ICASSP.2013.6637743.
[8] Estreder J, Piñero G, Aguirre F, de Diego M, Gonzalez A. On perceptual audio equalization for multiple users in presence of ambient noise. In: 2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM), Sheffield, UK. p. 445–9. https://doi.org/10.1109/SAM.2018.8448591.
[9] Virag N. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans Speech Audio Process 1999;7(2):126–37. https://doi.org/10.1109/89.748118.
[10] You CH, Ma B. β-masking MMSE speech enhancement for speech recognition. In: Proc. 2017 IEEE 2nd Int. Conf. on Signal and Image Processing (ICSIP), Singapore. p. 341–5. https://doi.org/10.1109/SIPROCESS.2017.8124561.
[11] Gonzalez A, Ferrer M, De Diego M, Piñero G, Garcia-Bonito JJ. Sound quality of low-frequency and car engine noises after active noise control. J Sound Vib 2003;265(3):663–79. https://doi.org/10.1016/S0022-460X(02)01462-1.
[12] Doleschal F, Rottengruber H, Verhey JL. Influence parameters on the perceived magnitude of tonal content of electric vehicle interior sounds. Appl Acoust 2021;181:108155.
[13] Bosi M, Goldberg RE. Introduction to Digital Audio Coding and Standards. Norwell, US: Kluwer Academic Publishers; 2002.
[14] Green DM. Additivity of masking. J Acoust Soc Am 1967;41(6):1517–25. https://doi.org/10.1121/1.1910514.
[15] Lutfi RA. Additivity of simultaneous masking. J Acoust Soc Am 1983;73(1):262–7. https://doi.org/10.1121/1.388859.
[16] Hellman RP. Asymmetry of masking between noise and tone. Atten Percept Psychophys 1972;11(3):241–6. https://doi.org/10.3758/BF03206257.
[17] Gockel H, Moore BCJ, Patterson RD. Asymmetry of masking between complex tones and noise: The role of temporal structure and peripheral compression. J Acoust Soc Am 2002;111(6):2759–70. https://doi.org/10.1121/1.1480422.
[18] Brandenburg K, Faller C, Herre J, Johnston JD, Kleijn WB. Perceptual coding of high-quality digital audio. Proc IEEE Inst Electr Electron Eng 2013;101(9):1905–19. https://doi.org/10.1109/JPROC.2013.2263371.
[19] Christoph M. Noise dependent equalization control. In: Audio Engineering Society Conf.: 48th Int. Conf.: Automotive Audio, Audio Engineering Society. p. 77–86.
[20] Rämö J, Välimäki V, Alanko M, Tikander M. Perceptual frequency response simulator for music in noisy environments. In: Audio Engineering Society Conf.: 45th Int. Conf.: Applications of Time-Frequency Processing in Audio, Helsinki, Finland. p. 1–10.
[21] Taghipour A. Psychoacoustics of detection of tonality and asymmetry of masking: implementation of tonality estimation methods in a psychoacoustic model for perceptual audio coding. Friedrich-Alexander-Universität Erlangen-Nürnberg; 2016. Ph.D. thesis,.
[22] Hastings A, Davies P. An examination of Aures's model of tonality. Proceeding on Sound Quality Symposium, Dearborn, MI, USA, Vol. 2. p. 4–9.
[23] Aures W. Procedure for calculating the sensory pleasantness of various sounds. Acustica 1985;59(vol. 59):130–41.
[24] Hastings A, Lee KH, Davies P, Surprenant AM. Measurement of the attributes of complex tonal components commonly found in product sound. Noise Control Eng J 2003;51(4):195–209. https://doi.org/10.3397/1.2839715.
[25] More SR. Aircraft noise characteristics and metrics. Purdue University; 2010.
[26] Di G-Q, Chen X-W, Song K, Zhou B, Pei C-M. Improvement of Zwicker's psychoacoustic annoyance model aiming at tonal noises. Appl Acoust 2016;105:164–70.
[27] Merino-Martinez R, Pieren R, Schäffer B, Simons DG. Psychoacoustic model for predicting wind turbine noise annoyance. In: 24th International Congress on Acoustics. p. 1–8.
[28] Merino-Martinez R, Vieira A, Snellen M, Simons DG. Sound quality metrics applied to aircraft components under operational conditions using a microphone array. In: 25th AIAA/CEAS Aeroacoustics Conference. p. 1–16.
[29] Torija AJ, Roberts S, Woodward R, Flindell IH, McKenzie AR, Self RH. On the assessment of subjective response to tonal content of contemporary aircraft noise. Appl Acoust 2019;146:190–203. https://doi.org/10.1016/j.apacoust.2018.11.015.
[30] Shrestha M, Zhong Z, Sound quality user-defined cursor reading control-tonality metric, Master's thesis, Informatics and Mathematical Modelling,

Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Accessed: 2023–11-23 (2003). http://www2.compute.dtu.dk/pubdb/pubs/2385-full.html.
[31] Welch PD. The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 1967;15(2):70–3. https://doi.org/10.1109/TAU.1967.1161901.
[32] Terhardt E, Stoll G, Seewann M. Algorithm for extraction of pitch and pitch salience from complex tonal signals. J Acoust Soc Am 1982;71(3):679–88. https://doi.org/10.1121/1.387544.
[33] Zwicker E, Terhardt E. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. J Acoust Soc Am 1980;68(5):1523–5. https://doi.org/10.1121/1.385079.
[34] ISO Central Secretary, Acoustics – Method for calculating loudness level – Part 1: Zwicker method, Standard ISO 532-1:2017, International Organization for Standardization, Geneva, CH (June 2017). https://www.iso.org/standard/63077.html
[35] Moore BC. Additivity of simultaneous masking, revisited. J Acoust Soc Am 1985;78(2):488–94. https://doi.org/10.1121/1.392470.
[36] Humes LE, Jesteadt W. Models of the additivity of masking. J Acoust Soc Am 1989;85(3):1285–94. https://doi.org/10.1109/TASL.2009.2023164.
[37] Pan D. A tutorial on MPEG/audio compression. IEEE Multimedia 1995;2(2):60–74. https://doi.org/10.1109/93.388209.
[38] Painter T, Spanias A. Perceptual coding of digital audio. Proc IEEE Inst Electr Electron Eng 2000;88(4):451–515. https://doi.org/10.1109/5.842996.
[39] Taghipour A, Jaikumar MC, Edler B. A psychoacoustic model with partial spectral flatness measure for tonality estimation. In: Proc. of the 22nd European Signal Processing Conf. (EUSIPCO), Lisbon, Portugal. p. 646–50.
[40] Belyi V, Gan W-S. Integrated psychoacoustic active noise control and masking. Appl Acoust 2019;145:339–48.
[41] Nilsson M, Soli SD, Sullivan JA. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J Acoust Soc Am 1994;95(2):1085–99. https://doi.org/10.1121/1.408469.
[42] Oliver RJ, Jot JM. Efficient multi-band digital audio graphic equalizer with accurate frequency response control. In: Audio Engineering Society Convention 139. New York, NY, USA: Audio Engineering Society; 2015.
[43] Westerlund N, Dahl M, Claesson I. Speech enhancement using an adaptive gain equalizer with frequency dependent parameter settings. IEEE 60th Vehicular Technology Conference, 2004. VTC2004-Fall. 2004, Vol. 5. Los Angeles, CA, USA: IEEE; 2004. p. 3718–22. https://doi.org/10.1109/VETECF.2004.1404759.
[44] Farina A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. Audio Engineering Society Convention, 108, 2000. p. 1–15.
[45] Abel JS, Berners DP. Filter Design Using Second-Order Peaking and Shelving Sections. In: 30th Annual International Computer Music Conference Proceedings (ICMC), Miami, FL, USA. p. 1–4.
[46] Valimaki V, Liski J. Accurate cascade graphic equalizer. IEEE Signal Processing Letters 2017;24(2):176–80. https://doi.org/10.1109/LSP.2016.2645280.
[47] David HA. The Method of Paired Comparisons. 2nd Edition. London: Griffin/Oxford University Press; 1988.