Auditory Time-Frequency Masking: Psychoacoustical Data and Application to Audio Representations

Thibaud Necciari[1,2], Peter Balazs[1], Richard Kronland-Martinet[2], Sølvi Ystad[2], Bernhard Laback[1], Sophie Savel[2], and Sabine Meunier[2]

[1] Acoustics Research Institute, Austrian Academy of Sciences,
Wohllebengasse 12–14, A-1040 Vienna, Austria
http://www.kfs.oeaw.ac.at
[2] Laboratoire de Mécanique et d'Acoustique, CNRS-UPR 7051,
31 chemin Joseph Aiguier, 13402 Marseille cedex 20, France
http://www.lma.cnrs-mrs.fr

Abstract. In this paper, the results of psychoacoustical experiments on auditory time-frequency (TF) masking using stimuli (masker and target) with maximal concentration in the TF plane are presented. The target was shifted either along the time axis, the frequency axis, or both relative to the masker. The results show that a simple superposition of spectral and temporal masking functions does not provide an accurate representation of the measured TF masking function. This confirms the inaccuracy of simple models of TF masking currently implemented in some perceptual audio codecs. In the context of audio signal processing, the present results constitute a crucial basis for the prediction of auditory masking in the TF representations of sounds. An algorithm that removes the inaudible components in the wavelet transform of a sound while causing no audible difference to the original sound after re-synthesis is proposed. Preliminary results are promising, although further development is required.

Keywords: auditory masking, time-frequency representation, Gabor, wavelets

1 Introduction

The main goal of this study was to collect data on time-frequency (TF) auditory masking for stimuli with maximal concentration in the TF plane. The results of such measurements served as a basis to examine the accuracy of simple TF masking models currently implemented in some perceptual audio codecs like MP3 and develop a perceptually relevant *and* perfectly invertible audio signal representation. The latter aspect was achieved by improving an existing algorithm [3] designed to remove the inaudible components in a TF transform while causing no audible difference to the original sound after re-synthesis.

1.1 Representations of Sound Signals in Audio Signal Processing

In the field of audio signal processing, many applications involving sound analysis-synthesis (*e.g.*, virtual reality, sound design, sonification, perceptual audio coding) require specific tools enabling the analysis, processing, and re-synthesis of non-stationary signals. Most of them are based on linear TF representations such as the Gabor and wavelet transforms. These transforms allow decomposing any natural sound into a set of elementary functions or "atoms" that are well localized in the TF plane (for a review on TF analysis see, *e.g.*, [4, 10, 35]). For the cited applications, obtaining a perceptually relevant (*i.e.*, providing a good match between signal representation and human auditory perception) and perfectly invertible signal representation would be of great interest. To that end, the long-term goal of the present study is to propose a signal representation being as close as possible to "*what we see is what we hear*". Because the achievement of such a sparse representation would facilitate the extraction and reconstruction of perceptually relevant sound features, it would also be of great interest for music information retrieval applications.

To date, two approaches exist to obtain a perceptually motivated *time versus frequency* representation of an audio signal. The first approach includes models of auditory processing providing an "internal" representation of sound signals like in [17, 26, 27]. While this approach is useful to improve our understanding of auditory processing, it does not enable reconstruction of the input signal. Thus, it is not useful as a TF analysis-synthesis tool. The second approach includes TF transforms whose parameters are tuned to mimic the spectro-temporal resolution of the auditory system [1, 16, 24]. While this approach is useful for audio signal analysis, the cited algorithms feature some limitations. In particular, [1,16] can only approximate the auditory resolution because the temporal and spectral resolutions cannot be set independently. The use of a bilinear transform in [24] overcomes this limitation but, on the other hand, this method does not allow reconstruction of the input signal. More recently, Balazs *et al.* [3] proposed a new approach to obtain a perceptually relevant and perfectly invertible signal representation. They introduced the concept of the "irrelevance filter", which consists in removing the inaudible atoms in a perfectly invertible Gabor transform while causing no audible difference to the original sound after re-synthesis. To identify the *irrelevant* atoms, a simple model of auditory spectral masking was used. A perceptual test performed with 36 normal-hearing listeners in [3] revealed that, on average, 36% of the atoms could be removed without causing any audible difference to the original sound after re-synthesis. The work described in the present paper can be considered as an extension of the irrelevance filter. Mostly, we attempt to improve it in two aspects: (i) overcome the fixed resolution in the Gabor transform by using a wavelet transform and (ii) replace the simple spectral masking model by using psychoacoustical data on auditory TF masking for stimuli with maximal concentration in the TF plane.

1.2 State-of-the-Art on Auditory Masking

Auditory masking occurs when the detection of a sound (referred to as the "target") is degraded by the presence of another sound (the "masker"). This effect is quantified by measuring the degree to which the detection threshold of the target increases in the presence of the masker.[3] In the literature, masking has been extensively investigated with simultaneous and non-simultaneous presentation of masker and target (for a review see, *e.g.*, [20]).

In simultaneous masking, the masker is present throughout the presentation time of the target (*i.e.*, the temporal shift between masker and target, ΔT, equals zero) and the frequency shift (ΔF) between masker and target is varied, resulting in the *spectral masking* function. To vary the ΔF parameter, either the target frequency (F_T) is fixed and the masker frequency (F_M) is varied, or vice versa. When F_T is fixed the masking function measures the response of a single auditory filter (*i.e.*, the filter centered on F_T).[4] As a result, such functions (called "psychoacoustical tuning curves" or "filter functions" depending on whether the masker or the target is fixed in level) are commonly used as estimates of auditory frequency selectivity [20, Chap. 3]. When F_M is fixed it is common to plot the target level at threshold (L_T) or amount of masking (see Footnote 3) as a function of F_T for a fixed-level masker, which is called a "masking pattern". Because a masking pattern measures the responses of different auditory filters (*i.e.*, those centered on the individual F_Ts), it can be interpreted as an indicator of the spectral spread of masking produced by the masker. The physiological mechanisms of simultaneous masking can be of two origins [5]: excitation and suppression. Excitation refers to the spread of excitation produced by the masker to the place responding to the target on the basilar membrane (BM). In other terms, the spread of excitation masks the BM response to the target. Suppression refers to the suppression or "inhibition" of the BM response to the target by the masker, even if the masker does not produce excitation at the place responding to the target (for an illustration of the excitatory and suppressive masking phenomena see [5, Fig. 1]). Excitation and suppression are not mutually exclusive. Their relative contributions depend on the frequency and level relationships between masker and target.

In non-simultaneous masking, ΔF most often equals zero and ΔT is varied, resulting in the *temporal masking* function. Backward masking (the target precedes the masker, $\Delta T < 0$) is weaker than forward masking (the masker

[3] The detection threshold of the target measured in presence of the masker represents the masked threshold, whereas the detection threshold of the target measured in quiet represents the absolute threshold. The difference between the masked threshold and the absolute threshold (in dB) represents the "amount of masking".

[4] The spectral resolution in the auditory system can be approximated by a bank of bandpass filters with a constant relative bandwidth. Each of these filters, named *auditory filters*, is characterized by its equivalent rectangular bandwidth (ERB) in Hz. This concept led to the definition of the so-called "ERB scale" that allows to plot signals or psychoacoustical data on a frequency scale related to human auditory perception [20, Chap. 3].

precedes the target, $\Delta T > 0$). The amounts of backward and forward masking depend on masker duration [6, 7, 36]. Although the mechanisms underlying backward masking remain unclear, the most accepted explanation is that it is of peripheral origin. Backwards masking would be caused by the temporal overlap of the BM responses to masker and target at the outputs of the auditory filters [6]. The amount of overlap depends on the "ringing" time of the BM, *i.e.*, the length of the impulse response of the BM, which itself depends on signal frequency.[5] Note, however, that backward masking studies often reported large inter-listener differences (*e.g.*, [7]) and that trained listeners often show little or no backward masking. Thus, backward masking may also reflect some confusion effects between masker and target [20]. Forward masking can be attributed to three mechanisms. The first is the temporal overlap of the BM responses to masker and target at the outputs of the auditory filters as a consequence of the filters' ringing. This phenomenon is more likely to be involved with small values of ΔT. The second is short-term adaptation or "fatigue": the exponential decay of masker-induced excitation over time in the cochlea and in the auditory nerve can reduce the response to a target presented shortly after the extinction of the masker [6,7,31,36]. The third is temporal integration or "persistence" of masker excitation: the neural representation of the masker is smoothed over time by an integration process so that the representation of the masker overlaps with the representation of the target at some stage in the auditory system [27, 28]. To date, however, the distinction between short-term adaptation and temporal integration as the most probable explanation to forward masking is still a matter of debate [25].

Because of the specific demands in the simultaneous and non-simultaneous masking experiments reported in the literature, the experimental stimuli were almost always broad either in the temporal domain (*e.g.*, long-lasting sinusoids), the frequency domain (*e.g.*, clicks), or both.

A few studies investigated how masking spreads in the TF domain (*i.e.*, by measuring masking patterns for various ΔTs) [7, 18, 21, 31]. Those studies involved relatively long (duration $\geqslant$ 100 ms) sinusoidal maskers, that is, maskers with good concentration in frequency but not in time. Overall, little is known about the spread of TF masking for a masker with good concentration both in time *and* frequency.

1.3 Models of Auditory Masking

The results of the spectral and temporal masking experiments reported in the literature were used to develop models of either spectral, temporal, or TF masking. Masking models are useful to both the fields of psychoacoustics and sound signal processing. In psychoacoustics they allow to improve our understanding of auditory processing. In signal processing they allow to exploit auditory masking in some applications, for instance in perceptual audio coding. To reduce the

[5] An estimation of the ringing time at a given frequency can be obtained by considering the inverse of the ERB (in Hz) of the auditory filter centered on that frequency.

digital size of audio files, audio codecs decompose sounds into TF segments and use masking models to reduce the bit rates in those segments (for a review on audio coding techniques see [32]).

Masking models can be classified into two groups: excitation pattern-based models and auditory processing-based models. Excitation pattern-based models transform the short-term spectrum of the input signal into an excitation pattern reflecting the spread of excitation induced by the signal on the BM. This approach is based on the power-spectrum model of masking in which the auditory periphery is conceived as a bank of bandpass filters (see Footnote 4). Masking is then determined by the target-to-masker ratio at the output of each filter. This group of models mostly includes spectral masking models (*e.g.*, [9, 13]) and is the technique most frequently employed in audio codecs [32].

In contrast, auditory processing-based models attempt to simulate the effective signal processing in the auditory system. Such models consist of a series of processing stages and a decision stage on which the prediction of masking is based [17, 27]. The model described in [27] is able to predict temporal and TF masking data. The model described in [17] is able to predict temporal and spectral masking data but has not been tested on TF conditions. Because auditory models usually have a high computational complexity and are not invertible, they are rarely used in audio processing algorithms.

To obtain a perceptually relevant audio signal representation based on a perfectly invertible transform, we propose to exploit masking in TF representations of sounds, that is, predict the audibility of each TF atom in the signal decompositions. To do so, a model of TF masking is required. There exist some models of TF masking that are currently implemented in audio coding algorithms [11, 12, 14, 34]. In the cited studies, the predictions of TF masking are based on a simple superposition of spectral and temporal masking functions (typically, only forward masking is considered). The decay of forward masking is modeled with a linear function of $\log(\Delta T)$ [11, 12, 34], or with an exponential function of the form $e^{-(\Delta T/\tau)}$ where τ is a time constant depending both on frequency (ΔF) and level [14]. Given the highly nonlinear behavior of cochlear mechanics (*e.g.*, [29]), such a simple combination of spectral and temporal masking functions is unlikely to correctly predict TF masking. Accordingly, the results presented in Sec. 2.3 reveal that such approaches are not adequate for predicting the audibility of TF atoms. To do so, it seems more appropriate to use masking functions that are based on the spread of TF masking produced by a maximally-compact masker. Because previous psychoacoustical studies mostly focused either on temporal or on spectral masking and used stimuli with temporally and/or spectrally broad supports, the spread of TF masking for a signal that is maximally compact in the TF plane cannot easily be derived from available data and therefore has to be measured.

1.4 Outline of the Present Study

The present paper consists of two main parts. In the first part, the results of psychoacoustical experiments on masking using maximally-compact stimuli are

presented. To best fulfill the requirement of maximum compactness in the TF plane, we used Gaussian-shaped sinusoids (referred to as Gaussians) as masker and target stimuli. Three experiments were conducted. The spectral and temporal masking functions for Gaussian maskers were measured in Experiments 1 and 2, respectively. In Experiment 3, the TF spread of masking was measured. We then tested with which accuracy the results from Exp. 3 can be predicted based on the results from Exps. 1 and 2 (assuming a simple superposition of spectral and temporal masking effects). This allowed us to examine the accuracy of simple TF masking models currently implemented in some perceptual audio codecs.

In the second part, the "extended" irrelevance filter based on psychoacoustical data on TF masking is described. Then, preliminary results are presented and discussed.

2 Psychoacoustical Measurements of Masking Using Gaussian Stimuli

2.1 General Methods

Stimuli. Masker and target were Gaussian-shaped sinusoids (Gaussians) defined by [23, 30]

$$s(t) = \sqrt{\Gamma} \sin\left(2\pi f_0 t + \frac{\pi}{4}\right) e^{-\pi(\Gamma t)^2} \tag{1}$$

where f_0 is the carrier frequency, Γ defines the equivalent rectangular bandwidth (ERB), and Γ^{-1} defines the equivalent rectangular duration (ERD) of $s(t)$. In our experiment, Γ was set to 600 Hz, corresponding to $\Gamma^{-1} = 1.7$ ms. The f_0 value varied depending on ΔF. By introducing the $\pi/4$ phase shift, the energy of the signal is independent of f_0. Since a Gaussian window has infinite duration, the signals were windowed in the time domain using a Tukey window. The "effective duration" (defined as the 0-amplitude points duration) of the stimuli was 9.6 ms and the cutoff in the frequency domain was located at the 220-dB down points. The sound pressure level (SPL) of the Gaussian was specified by measuring the SPL of a long-lasting sinusoid having the same frequency (f_0) and maximum amplitude as the carrier tone of the Gaussian.

Procedure. Thresholds were estimated using a three-interval, three-alternative forced-choice procedure with a 3-down-1-up criterion that estimates the 79.4%-correct point on the psychometric function. Each trial consisted of three 200-ms observation intervals visually indicated on the response box, with a between-interval gap of 800 ms. The masker was presented in the three intervals and the target was presented with the masker in one of those intervals, chosen randomly. The listener indicated in which interval he/she heard the target by pressing one of three buttons on the response box. Immediate feedback on the correctness of

the response was visually provided to the listener. The target level varied adaptively by initial steps of 5 dB and 2 dB following the second reversal. Twelve reversals were obtained. The threshold estimate was the mean of the target levels at the last 10 reversals. A threshold estimate was discarded when the standard deviation of these 10 reversals exceeded 5 dB. Two threshold estimates were obtained for each condition. If the standard deviation of these two estimates exceeded 3 dB, up to four additional estimates were completed. The final threshold was the average across all estimates (maximum = 6).

Apparatus. A personal computer was used to control the experiments and generate the stimuli. Stimuli were output at a 48-kHz sampling rate and a 24-bit resolution using an external digital-to-analog converter (Tucker-Davis Technologies (TDT) System III), attenuated (TDT PA5) and passed to a headphone buffer (TDT HB7), and to the right ear-pad of a circumaural headphone (Sennheiser HD545).The headphones were calibrated so that levels were considered as SPL close to the eardrum. Listeners were tested individually in a double-walled, sound-attenuated booth.

Listeners. Six normal-hearing listeners participated in Exps. 1 and 2. Four of the listeners (L1–L4) participated in Exp. 3.

Experimental Conditions. Throughout the experiments, the carrier frequency of the masker was fixed to 4 kHz. Its sensation level (*i.e.*, the level above the absolute threshold of the masker for each listener, see Footnote 3) was fixed to 60 dB, which corresponded to SPLs of 81–84 dB across listeners.

Spectral Masking. Masker and target were presented simultaneously ($\Delta T = 0$). Masking patterns were measured for 11 values of ΔF, defined in the ERB scale: -4, -3, -2, -1, 0, +1, +2, +3, +4, +5, and +6 ERB units.[6] To prevent cochlear combination products from being detected and thus from producing irregularities in the masking patterns [20], a continuous background noise was added for all ΔFs > 0 [23]. Each session contained conditions measured either with or without background noise. The order of sessions (with noise; without) was counterbalanced over days. Within a session, ΔF was chosen randomly.

Temporal Masking. Masker and target had the same carrier frequency ($\Delta F = 0$). Because a pilot experiment indicated very little backward masking for such short maskers, we focused on forward masking. ΔT, defined as the time shift between masker onset and target onset, was 0, 5, 10, 20, or 30 ms. Within a session, ΔT was chosen randomly.

[6] The target frequencies corresponding to these ΔFs were 2521, 2833, 3181, 3568, 4000, 4480, 5015, 5611, 6274, 7012, and 7835 Hz, respectively.

Time-Frequency Masking. Both ΔT and ΔF were varied. Masked thresholds were measured for 30 out of 40 possible $\Delta T \times \Delta F$ combinations (*i.e.*, 5 ΔTs from Exp. 2×8 ΔFs from Exp. 1). Although the effect of cochlear combination products is usually ignored in forward masking studies, we used a background noise identical to that of Exp. 1 to mask potential cochlear combination products (when $\Delta F > 0$) because of the small ΔT values. The whole set of conditions was split into two groups: frequency separations measured with and without background noise. Then, experimental blocks were formed that contained the ΔT conditions for each ΔF. The order of blocks and groups was randomized across sessions. Within a session, the target frequency was fixed and ΔT was chosen randomly.

2.2 Results

This section presents the data with respect to their applications described in Secs. 2.3 and 3. A more thorough description and interpretation of the data can be found in [23].

Experiment 1: Spectral Masking. Figure 1 presents the individual and mean amounts of masking (in dB) as a function of ΔF (in ERB units). First, in some listeners a dip (L1, L3 and L4) or a plateau (L5) was observed instead of a peak at $\Delta F = 0$.[7] It has to be considered that this represents a special condition, where masker and target were exactly the same stimuli presented at the same time. Thus, the listeners could only use as a cue the intensity increase in the interval containing the target. In other words, the listeners performed an intensity discrimination task in this condition [8, 22, 23].

Second, for all listeners and $|\Delta F| \geqslant 2$ ERB units, the amount of masking decreased as $|\Delta F|$ increased. The decrease was more abrupt for $F_T < F_M$ than for $F_T > F_M$. Regression lines computed for each side of the masking patterns and listener (straight lines in Fig. 1) indeed show that, on average, the slopes for $F_T < F_M$ (mean slope = +60 dB/octave) are 1.6 times those for $F_T > F_M$ (mean slope = -39 dB/octave). This steeper masking decay for $F_T < F_M$ is consistent with that reported in classical spectral masking studies (see, *e.g.*, [20, 22] for a review).

Experiment 2: Temporal Masking. Figure 2 presents the individual and mean amounts of masking as a function of ΔT on a logarithmic scale. On average, masking decreased from 50 dB for $\Delta T = 0$ to about 6 dB for $\Delta T = 30$ ms. The data for $\Delta T > 0$ are well fitted with straight lines, a result consistent with almost all previous forward masking studies using various types of maskers (*e.g.*, [6, 7, 36]). A straightforward description of these data is provided by

[7] In simultaneous masking, the greatest amount of masking, also referred to as the "maximum masking frequency", is classically located at $F_T = F_M$, which results in a peak in the masking pattern at $\Delta F = 0$ [22].

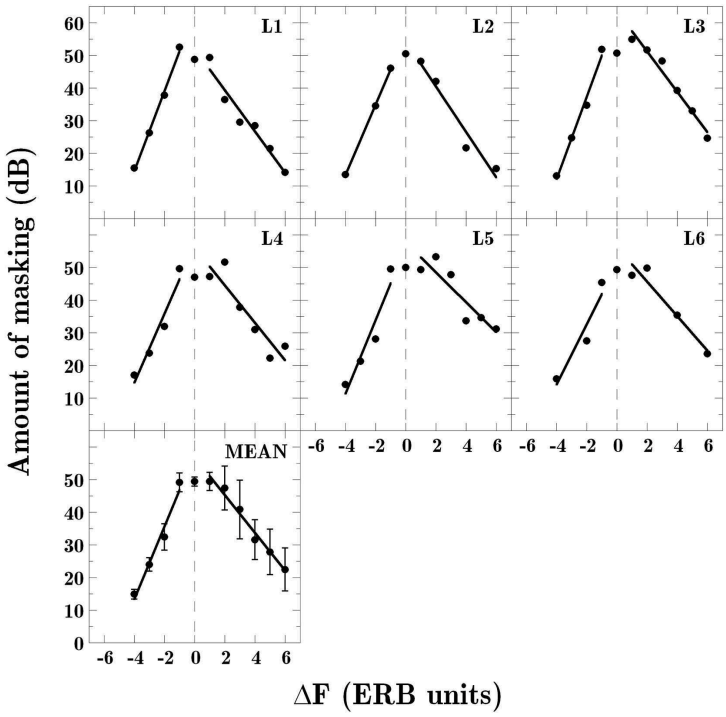


Fig. 1: Results of Experiment 1: amount of masking (in dB) as a function of ΔF (in ERB units). Data were fitted with linear regression lines on each side of the masking patterns (excluding the point at $\Delta F = 0$). The bottom panel shows the mean data with ± 1 standard deviation bars [23].

$$AM = \alpha \log(\Delta T) + \beta \tag{2}$$

where AM is the amount of masking, α is the slope of the forward masking decay, and β is the offset of the forward masking decay. Table 1 lists the values of α and β determined by applying a weighted-least-squares fit of (2) to the data for $\Delta T > 0$. To take the variability of each data point into account in the estimation of parameters α and β, the weight of each data point corresponded to the reciprocal of the variance of the measurement.

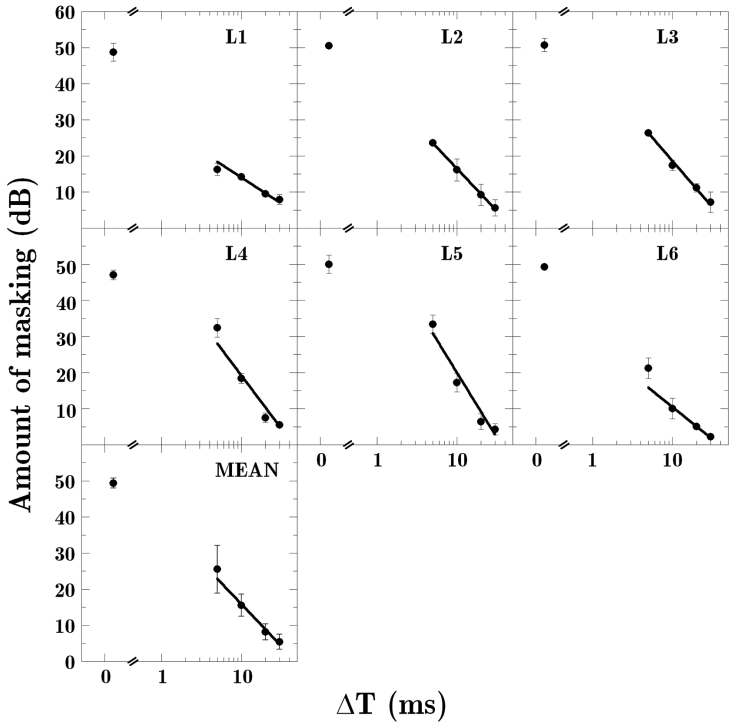


Fig. 2: Results of Experiment 2: amount of masking (in dB) as a function of ΔT (in ms) on a logarithmic scale with straight-line fits to the data for $\Delta T > 0$ according to (2). The fit parameters are listed in Tab. 1. Error bars in the individual panels indicate ± 1 standard deviation across measurements. The bottom panel shows the mean data with ± 1 standard deviation bars [23].

Experiment 3: Time-Frequency Masking. Figure 3 presents the results as simultaneous and forward masking patterns, that is, the amount of masking as a function of ΔF with ΔT as the parameter. For all ΔFs, the largest amount

Table 1: Values of parameters α (in dB/log(ΔT)) and β (in dB) determined by fitting (2) to the data for $\Delta T > 0$ in Fig. 2 using a weighted-least-squares criterion. Note that because the ΔT axis is logarithmically scaled in Fig. 2, the values of β correspond to the y-intercepts at $\Delta T = 1$ ms. The last column indicates r^2 values.

Listener	α	β	r^2
L1	−14.39	28.46	0.98
L2	−23.38	39.93	1.00
L3	−25.61	44.28	0.99
L4	−29.37	48.60	0.95
L5	−36.00	56.01	0.96
L6	−17.76	28.26	0.97
MEAN	−23.18	39.12	0.97

of masking was obtained in the simultaneous condition ($\Delta T = 0$). Masking dropped as ΔT increased to 5 ms. For ΔTs > 10 ms, masking was generally less than 10 dB for all ΔFs. To assess whether the patterns broadened or narrowed with increasing ΔT, we estimated the quality factors at the -3-dB bandwidth (Q_{3dB}) [18]. The mean values of Q_{3dB} are 12, 3, and 2 for $\Delta T = 0$, 5, and 10 ms, respectively, *i.e.*, the patterns flattened as ΔT increased. The mean masking patterns in Fig. 3 are asymmetric for all ΔTs. Finally, the dip/plateau observed in listeners L1, L3, and L4 at $\Delta F = 0$ for $\Delta T = 0$ (see also Fig. 1) disappeared when ΔT increased. For $\Delta T > 0$, listeners L1 and L3 exhibited a peak at $\Delta F = +1$ instead of 0. In other terms, these two listeners revealed a shift in the maximum masking frequency towards F_Ts > F_M.

Our pattern of results is consistent with the few preceding studies that measured TF masking patterns with long maskers [7, 18, 21, 31] in that (i) masking patterns flatten with increasing ΔT, (ii) the masking patterns' asymmetry remains for $\Delta T > 0$, and (iii) a shift in the maximum masking frequency towards F_Ts > F_M is observed in some listeners for $\Delta T > 0$. However, because TF masking is affected by nonlinear processes in the cochlea [19, 20, 23], the present data could not have been deduced from existing data for long maskers.

Our results are summarized in the three-dimension plot in Fig. 4. To provide a smooth and "complete" representation of TF masking (*i.e.*, one that reaches 0 dB of masking), the ΔT axis was sampled at 1 kHz and the data for ΔFs below -4 and above +6 ERB units were then extrapolated based on a two-dimensional cubic spline fit along the TF plane. Overall, the function shown in Fig. 4 represents the TF spread of masking produced by a Gaussian TF atom.

2.3 Accuracy of Simple Time-Frequency Masking Models Used in Perceptual Audio Codecs

To examine the accuracy of simple TF masking models currently used in some audio codecs, we tested two prediction schemes assuming a linear combination

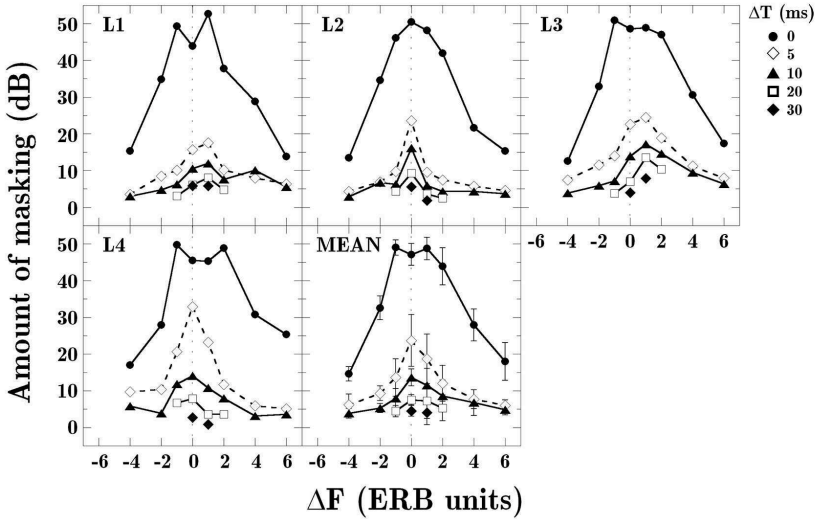


Fig. 3: Results of Experiment 3: amount of masking (in dB) as a function of ΔF (in ERB units) obtained for five ΔTs [23].

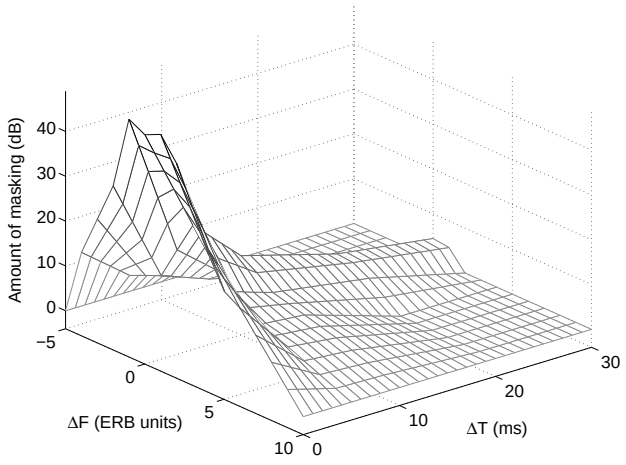


Fig. 4: Mean TF masking data extrapolated and plotted in the TF plane [23]

of spectral and temporal masking. Specifically, we tested with which accuracy the results of Exp. 3 can be predicted based on the results of Exps. 1 and 2. The general idea of the prediction is that the spread of TF masking caused by a masker can be described by the spectral masking pattern combined with the decay of forward masking from each point of the masking pattern. In the following, let $AM(\Delta T, \Delta F)$ denote the amount of masking produced by the masker on a target separated from the masker by ΔT and ΔF in the TF plane ($\Delta T > 0, \Delta F \neq 0$).

Simple Superposition of Spectral and Temporal Masking Functions. We first considered a simple superposition of the spectral and temporal masking functions to predict TF masking. A similar approach is used in [11,12,34]. This prediction scheme, referred to as "Prediction A", is given by

$$AM(\Delta T, \Delta F) = AM(0, \Delta F) - \big(AM(0,0) - AM(\Delta T, 0)\big) \tag{3}$$

where $AM(0, \Delta F)$ represents the "initial" spread of masking produced by the masker at the target frequency (read from Fig. 1) from which is subtracted the temporal decay of forward masking over time ΔT (read from Fig. 2). The mean masking patterns predicted with Prediction A for ΔT values of 5, 10, and 20 ms are depicted in Fig. 5 (crosses, solid lines). It is clear from the figure that Prediction A overestimates the amount of masking for small frequency separations ($|\Delta F| \leqslant 2$ ERB units) and underestimates masking for larger ΔFs. One obvious reason for the inefficiency of Prediction A is the fact that it does not take into account the ΔF dependency of the forward masking decay.

Superposition of Spectral Masking Function and Level-Dependent Temporal Masking Function. An approach which takes into account the ΔF dependency of forward masking is used in [14]. In this approach (referred to as "Prediction B"), each point of the spectral masking pattern with $F_M \neq F_T$ is considered as a hypothetical forward masker with $F_M = F_T$ but with a lower level. This consideration follows from the analogy reported between the ΔF- and level-dependency of forward masking [23]. Prediction B has the form

$$AM(\Delta T, \Delta F) = AM(0, \Delta F) - \alpha' \log(\Delta T) \tag{4}$$

with $\alpha' = AM(0, \Delta F)/\log(\Delta T_{0dB})$, ΔT_{0dB} being the ΔT-axis intercept, or 0-dB masking point, at which the forward masking functions converge. Given the parameters α and β determined by fitting (2) to the temporal masking data presented in Fig. 2, $\Delta T_{0dB} = 10^{-\beta/\alpha}$ (see Tab. 1). As for Prediction A, $AM(0, \Delta F)$ is determined from the spectral masking data. The mean masking patterns predicted with Prediction B for ΔT values of 5, 10, and 20 ms are depicted in Fig. 5 (filled diamonds, dashed lines). It can be seen that the shape of the masking patterns with Prediction B is more similar to the data than that with Prediction A. Nevertheless, it is clear that Prediction B overestimates the amount of masking in almost all conditions, the overestimation being particularly large for small

frequency separations ($|\Delta F| \leqslant 2$ ERB units). Overall, both prediction schemes failed in accurately predicting TF masking data for Gaussian stimuli. This confirms that TF masking is not predictable by assuming a simple combination of temporal and spectral masking functions.

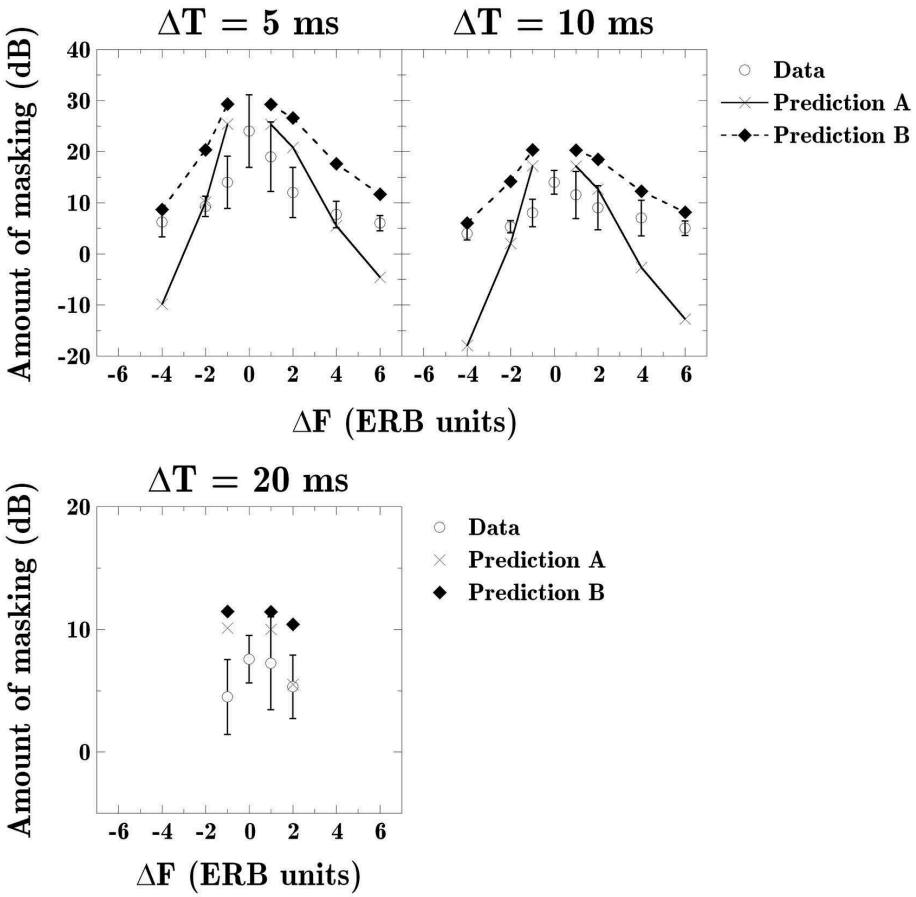


Fig. 5: Mean forward masking patterns for ΔT = 5, 10, and 20 ms predicted with Prediction A ($\times$, solid lines) and Prediction B ($\blacklozenge$, dashed lines). Because only one point was measured for $\Delta F < 0$ and ΔT = 20 ms, only symbols are used in the bottom left panel. Error bars show ± 1 standard deviation.

2.4 Interim Summary

To obtain a measure of the TF spread of masking produced by a signal with maximal concentration in the TF plane, three experiments were conducted that involved Gaussian-shaped sinusoids with fixed bandwidth (ERB = 600 Hz) and duration (ERD = 1.7 ms) both as masker and target. In all experiments, the masker had a carrier frequency of 4 kHz and a sensation level of 60 dB. The target was shifted relative to the masker either in frequency, in time, or both.

The results of the frequency- and time-shift conditions showed that the superposition of spectral and temporal masking effects, as currently implemented in some perceptual audio codecs, does not provide an accurate representation of the measured TF masking effects for Gaussian maskers. These results suggest that audio coding algorithms using such an approach provide rather erroneous predictions of TF masking.

The results of the TF conditions provide the TF spread of masking produced by a Gaussian TF atom. These new data constitute a crucial basis for the prediction of auditory masking in audio TF representations. This is addressed in the following section.

3 Exploiting Time-Frequency Masking in a Time-Frequency Transform: Improvement of the Irrelevance Filter Algorithm

The concept of the irrelevance filter was first introduced in [3]. It consists in removing the inaudible atoms in a Gabor transform while causing no audible difference to the original sound after re-synthesis. The algorithm first determines an estimation of the masked threshold based on a simple model of spectral masking. The masked threshold is then shifted in level by an amount determined experimentally, which results in the "irrelevance threshold". This shift gives a conservative way to deal with uncertainty effects resulting from removing TF atoms and with inaccuracies in the masking model. Next, all TF atoms falling below threshold are removed. Although a perceptual test performed in [3] with 36 normal-hearing listeners indicated that, on average, 36% of the atoms can be removed without causing any audible difference to the original sound after re-synthesis, the irrelevance filter algorithm can be improved. The main limitations of the algorithm are the fixed resolution in the Gabor transform and the use of a simple spectral masking model to predict masking in the TF domain.

In this section, a preliminary version of the extended irrelevance filter is presented. Because the algorithm presented below differs from the original algorithm in many aspects, including signal representation, masking model, and irrelevance threshold calculation, no direct comparison can be established between the two versions. Moreover, because the new algorithm is still being developed, it has not been formally evaluated yet (*e.g.*, by conducting perceptual listening tests with natural sounds). Thus, we merely evaluate the performance of the new algorithm based on preliminary results with deterministic signals and informal listening by the authors.

3.1 Choice of the Signal Representation: Wavelet Transform

To mimic the spectral resolution of the human auditory system, a signal representation allowing a variable frequency resolution is required. The continuous wavelet transform (CWT) fulfills this requirement, unlike the Gabor transform that allows only a fixed TF resolution (*e.g.*, [10]). Thus, the CWT was chosen as the TF analysis-synthesis scheme used in this paper. In the following we summarize some general theory on wavelets (for a more detailed description see, *e.g.*, [4, 35]) and describe the practical implementation of the CWT used in our algorithm.

The CWT results from the decomposition of a signal into a family of functions that are scaled versions of a prototype function ("mother wavelet") $g(t)$ according to

$$g_a(t) \;=\; \frac{1}{\sqrt{a}}\, g\left(\frac{t}{a}\right) \quad (a \in \mathbb{R}^{*+}) \tag{5}$$

where a is a *scale factor* allowing to compress ($a < 1$) or dilate ($a > 1$) the mother wavelet $g(t)$ ($a = 1$). This parameter defines the time and frequency resolution in the TF plane. The scale is linked to the frequency according to $\omega = \omega_0/a$ where ω_0 is the pulsation of the mother wavelet. Then, for any signal $x(t)$,

$$\begin{aligned} CWT_x(b,a) &= \langle\, g_{a,b}, x \,\rangle \\ &= \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\, \overline{g\left(\frac{t-b}{a}\right)}\, \mathrm{d}t \end{aligned} \tag{6}$$

provides a two-dimensional representation of $x(t)$ in the time-scale plane, $b \in \mathbb{R}$ being the time variable. Using Parseval's relation, (6) can also be written in the frequency domain

$$CWT_x(b,a) \;=\; \frac{\sqrt{a}}{2\pi} \int_{-\infty}^{+\infty} \hat{x}(\omega)\, \overline{\hat{g}(a\omega)}\, e^{jb\omega}\, \mathrm{d}\omega \tag{7}$$

where $\hat{x}(\omega)$ and $\hat{g}_a(\omega)$ denote the Fourier transforms of $x(t)$ and $g_a(t)$, respectively. The CWT is invertible if and only if $g(t)$ fulfills the admissibility condition

$$C_g \;=\; \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{|\hat{g}(\omega)|^2}{\omega}\, \mathrm{d}\omega \;<\; \infty \tag{8}$$

which implies that $g(t)$ be of finite energy. This is usually fulfilled in practice since $g(t)$ must be a function oscillating in time (hence the name "wavelet"). Additionally, $g(t)$ must be of zero mean. Finally, the reconstruction formula is

$$x(t) \;=\; \frac{1}{C_g} \iint\limits_{a\,>\,0,\,\mathbb{R}} CWT_x(b,a)\, g_{a,b}(t)\, \frac{\mathrm{d}a\mathrm{d}b}{a^2} \tag{9}$$

which reflects the "atomic" decomposition of $x(t)$ into wavelets.

The CWT has the properties of being linear and ensuring energy conservation. Another property is the "reproducing kernel", which states that

$$CWT(a',b') = \frac{1}{C_g} \iint_{\mathbb{R}} K_g(a',b',a,b) CWT(a,b) \frac{\mathrm{d}a\mathrm{d}b}{a^2} \tag{10}$$

where $K_g(a',b',a,b) = \langle\, g_{a,b}, g_{a',b'} \,\rangle$ is called the reproducing kernel. Equation (10) means that the reproducing kernel ensures a strong correlation between all components in the CWT. In other words, any component at location a', b' depends upon the remote component at location a, b through the reproducing kernel. This is reflected by the fact that the CWT is highly redundant.

The numerical implementation of the CWT requires the discretization of the time and scale variables b, a and the choice of the mother wavelet $g(t)$. In our implementation we chose the following discretization $(a_j, b_k) = (a_0^j, kb_0)$ where $a_0 = 2$ and $b_0 = 1/F_S$ (F_S being the sampling frequency) are two constants defining the size of the sampling grid. Furthermore, we opted for a sub-sampling of the scale in voices and octaves such that

$$a_0^j = 2^{\frac{m}{\mathcal{D}_v} + n} = a_{m,n}$$

where $m \in [0, \ldots, \mathcal{D}_v - 1]$, $n \in [0, \ldots, \mathcal{D}_o - 1]$, and $j \in [0, \ldots, \mathcal{D}_v \mathcal{D}_o - 1]$, $\mathcal{D}_v$ and $\mathcal{D}_o$ being the number of voices and octaves, respectively. This discretization yields scale factors $a_0^j \geqslant 1$ with increment steps of $2^{1/\mathcal{D}_v}$. Moreover, it provides two parameters, $\mathcal{D}_v$ and $\mathcal{D}_o$, for determining the total number of scales in the representation.

Regarding the choice of the mother wavelet, the accurate prediction of masking in the time-scale domain requires that the spectro-temporal characteristics of the wavelets match the spectro-temporal characteristics of the masker used in the psychoacoustical experiment. Thus, a straightforward solution for $g(t)$ was to use a Gaussian-shaped sinusoid similar to that defined in (1). Because the CWT was computed in the frequency domain according to (7), we defined the following function for the mother wavelet

$$\hat{g}(\omega) = \frac{1}{2i\sqrt{\Gamma}}\, e^{-\pi\left(\frac{\omega-\omega_0}{\Gamma}\right)^2} \tag{11}$$

where $\Gamma = \mu f_0 = \mu \frac{\omega_0}{2\pi}$, μ being the shape factor of the Gaussian window. To provide a Γ value of 600 Hz at $f_0 = 4$ kHz as used in the psychoacoustical experiment (see Sec. 2), μ was set to 0.15. Note that $\hat{g}(\omega)$ corresponds to the positive-frequency components of the Fourier transform of $s(t)$ in (1).

The frequency of the mother wavelet (f_0) was set to $3F_S/8$. Because we used only scale factors $a_0^j \geqslant 1$ (*i.e.*, we used only compressed versions of $\hat{g}(\omega)$), f_0 defines the highest center frequency in the signal representation. To cover the whole spectrum of audible frequencies (*i.e.*, 0.02–20 kHz) while maintaining a large overlap between wavelets, thus to avoid loosing details in the signals, we used 108 scales split into 9 octaves ($\mathcal{D}_o = 9$) and 12 voices per octave ($\mathcal{D}_v = 12$). At $F_S = 44.1$ kHz and $\mu = 0.15$, the highest-frequency analysis filter had a center frequency $f_0 = 16.5$ kHz and a bandwidth of 2.5 kHz. The lowest-frequency filter had a center frequency $f_0/a_0^{\mathcal{D}_v\mathcal{D}_o-1} = 33.8$ Hz and a bandwidth of 5 Hz.

3.2 Implementation of the Irrelevance Filter

The gathered TF masking data were used to predict masking in the time-scale domain. More specifically, the TF masking function in Fig. 4 was used as a *masking kernel* in the time-scale domain. Accordingly, this function had to be discretized in time and scales. Because we conserved all time samples of the signal, the ΔT axis was sampled at F_S. The ΔF axis (in ERB units) had to be matched to the scale axis (in voices and octaves). Considering that the ERB of an auditory filter corresponds to approximately one third of octave [20] and the present analysis counts 12 voices per octave, one ERB unit was associated with 4 voices. The TF masking kernel in Fig. 4 covers a range of 15 ERB units (ΔF = -5 to +10). Thus, the ΔF axis should be divided into 61 voices. This was achieved by interpolating the ΔF axis at a sampling rate of 4 voices per ERB unit based on a two-dimensional cubic spline fit along the TF plane.

In the following, we denote by $X(a,b)$, $a = \{a_j; j = 0, \ldots, \mathcal{D}_o\mathcal{D}_v - 1\}$ the discrete wavelet transform (DWT) of the input signal $x(k)$, k being the discrete time variable such that $t = kT_S$ (all signals were sampled at $F_S = 44.1$ kHz). The representation from which components have been removed is referred to as $\widetilde{X}(a,b)$. Accordingly, the output signal (reconstructed from $\widetilde{X}(a,b)$) is referred to as $\tilde{x}(k)$. $\mathcal{M}(a,b)$ refers to the discrete masking kernel in dB.

The structure of the irrelevance time-scale filter is shown in Fig. 6. The algorithm includes three main steps:

1. Scale the modulus of the DWT in dB SPL. The difficulty in the SPL normalization is that the actual playback level remains unknown during the entire signal processing. We considered that an amplitude variation of ± 1 bit in the signal is associated with a SPL of 0 dB, while a full-scale signal is associated with a SPL close to 92 dB [32].
2. Identify local maskers, *i.e.*, local maxima in the transform that fulfill

$$|X(a,b)| \geqslant Tq(a,\cdot) + 60 \quad \text{(in dB SPL)}$$

   where $Tq(a)$ is an analytic function approximating the absolute threshold of a normal-hearing listener in dB SPL. It is given by [33]

$$T_q(a) = 3.64\,(f_0/a)^{-0.8} - 6.5\,e^{-0.6\left(\frac{f_0}{a}-3.3\right)^2} + \frac{(f_0/a)^4}{1000}$$

   with f_0 in kHz. More precisely, Step 2 selects the components whose SPL exceeds the absolute threshold by 60 dB. This selection rule follows from the masker sensation level of 60 dB used in the experiment (see Sec. 2). Let Ω_M denote the set of maskers selected in Step 2.
3. Apply the masking kernel $\mathcal{M}(a,b)$ (in dB) to each masker in order of descending SPL and iteratively compute the output wavelet transform as

$$\widetilde{X}(a,b) = \begin{cases} X(a,b) & \text{if} \quad |X(a,b)| \geqslant Tq(a,\cdot) + \mathcal{M}(a,b) \quad \text{(dB SPL)} \\ 0 & \text{otherwise} \end{cases}$$

   until Ω_M is empty.

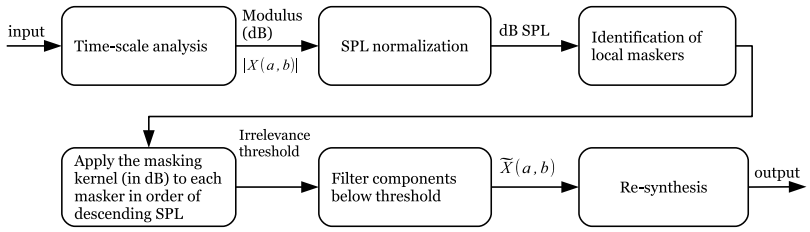


Fig. 6: Structure of the irrelevance time-scale filter

Note that a more straightforward approach could have consisted in applying $\mathcal{M}(a,b)$ to the whole time-scale domain, that is, without identifying local maskers in the transform (Step 2). However, we opted for a local application of $\mathcal{M}(a,b)$ because the amount of masking highly depends on level [20]. Thus, applying the TF masking kernel derived from data measured with an average masker SPL of 84 dB to components with SPLs below 84 dB is likely to result in an overestimation of masking. This would in turn result in the removal of *audible* components. To process all components in the transform, a level-dependent masking kernel is required.

3.3 Results

We present below the results obtained when the irrelevance filter was applied on deterministic and musical signals.[8] Two conditions measured in Exp. 3 were tested: one condition ("Condition 1": $\Delta F = +4$ ERB units, $\Delta T = 10$ ms, target SPL (L_T) = 50 dB) where the target is not masked and another condition ("Condition 2": $\Delta F =$ -2 ERB units, $\Delta T = 5$ ms, $L_T = 15$ dB) where the target is masked. A test signal $x(t)$ composed of two Gaussians (see (1)) with time and frequency shifts was synthesized as follows:

$$x(t) = \underbrace{\mathfrak{g}_M(t)}_{\text{Masker}} + \underbrace{\mathfrak{g}_T(t-\Delta T)}_{\text{Target}} \tag{12}$$

with $\mathfrak{g}_l(t) = A_l \sin\left(2\pi f_l t + \frac{\pi}{4}\right) e^{-\pi(\Gamma t)^2}$, $l = \{M, T\}$ where A_l allows to control the signal SPLs[9]. Let $x_1(t)$ and $x_2(t)$ denote the test signals for Conditions 1 and 2, respectively. Their parameters are listed in Tab. 2.

Consider first $x_1(t)$. Because the target is not masked, the representation of $\mathfrak{g}_T(t)$ should not be removed from $X_1(a,b)$. Figure 7 depicts the original (Fig. 7a) and modified (Fig. 7b) representations of $x_1(t)$ in dB SPL. As expected, it can

[8] The sound files corresponding to each of the results can be downloaded as wav files at: `http://www.lma.cnrs-mrs.fr/~kronland/cmmr2011`.

[9] The SPL of the test signal was controlled by setting the signal amplitudes $A_l = 10^{(L_l-92)/20}$ where L_l is the desired SPL and 92 dB corresponds to the amplitude of a full-scale signal.

Table 2: Parameters used for test signals $x_1(t)$ and $x_2(t)$ to simulate experimental Conditions 1 and 2.

	$x_1(t)$	$x_2(t)$
Γ	600	600
f_M (kHz)	4.0	4.0
L_M (dB SPL)	80	80
f_T (kHz)	6.3	3.2
L_T (dB SPL)	50	15
ΔT (ms)	10	5
Target status	*not masked*	*masked*

be seen that the target was not removed from $X_1(a,b)$. However, the representation of $\mathfrak{g}_M(t)$ was roughly altered by the filter. To evaluate the amount of components filtered out from $X_1(a,b)$, we computed the binary representations associated with $|X_1(a,b)|$ and $|\widetilde{X}_1(a,b)|$. These representations (not shown) comprise pixels '1' where $|X_1(a,b)|$ (respectively, $|\widetilde{X}_1(a,b)|$) > -10 dB SPL and pixels '0' elsewhere. Comparing the DWT of the input representation $X_1(a,b)$ and the output representation $\widetilde{X}_1(a,b)$ indicated that about 45% components were removed. It has to be considered, however, that $|\widetilde{X}_1(a,b)|$ in Fig. 7b is *not* the actual representation of $\tilde{x}_1(t)$. Because of the reproducing kernel (see (10)), reconstructing the signal from the modified representation restores some of the removed components. To illustrate this effect, the modulus of the DWT of $\tilde{x}_1(t)$, *i.e.*, the analysis of the reconstructed signal, is represented in Fig. 7c. It can be seen that the masker components removed by the filter were restored by the reproducing kernel. Informal listening revealed no perceptual difference between $\tilde{x}_1(t)$ and $x_1(t)$, and the reconstruction error $(\tilde{x}_1(t) - x_1(t))$ was $< 10^{-4}$.

Consider next $x_2(t)$. In this case the target is masked, and thus the representation of $\mathfrak{g}_T(t)$ should be removed from $X_2(a,b)$. This was the case, as depicted in Fig. 8b. Computations of the binary representations indicated that about 57% components were removed. As for $\tilde{x}_1(t)$, the reproducing kernel restored the masker components removed by the filter. Informal listening revealed no perceptual difference between $\tilde{x}_2(t)$ and $\mathfrak{g}_M(t)$.

Finally, we applied the irrelevance time-scale filter to a musical sound (referred to as $x_3(t)$), namely a clarinet sound representing the note A3. The results are depicted in Fig. 9. Computations of the binary representations indicated that about 50% components were removed. Although the reconstruction error was $< 10^{-4}$, in that case, informal listening revealed some perceptual differences between $\tilde{x}_3(t)$ and $x_3(t)$. More specifically, the filter altered the attack of the note, which became noisy.

3.4 Discussion

The preliminary results obtained with deterministic sounds indicated that the irrelevance time-scale filter removes information (as predicted from experimental

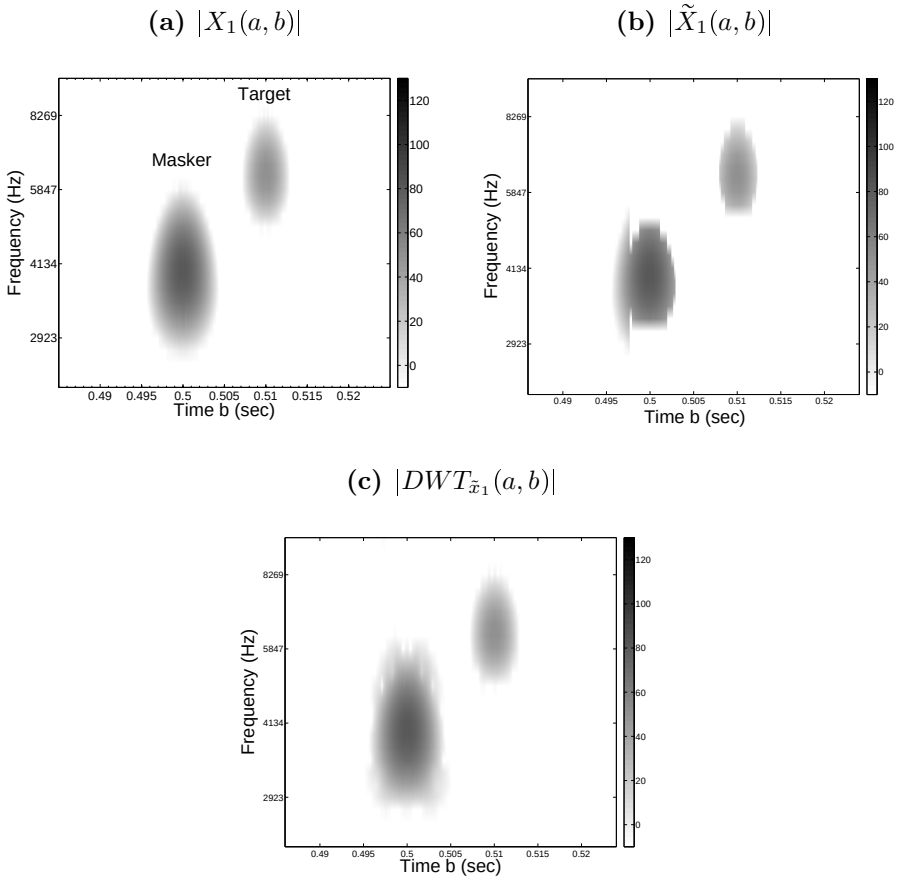
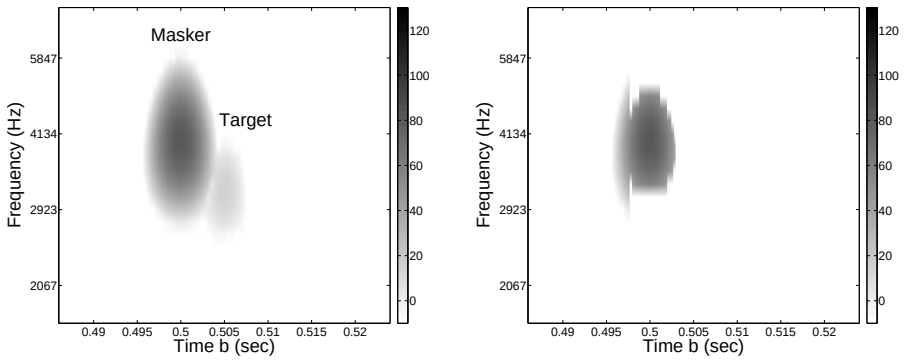


Fig. 7: Modulus of the DWT (in dB SPL) of test signal $x_1(t)$ (see Tab. 2) (a) at the input and (b) at the output of the irrelevance time-scale filter. (c) Modulus of the DWT (in dB SPL) of the reconstructed signal $\tilde{x}_1(t)$.

(a) $|X_2(a,b)|$ **(b)** $|\widetilde{X}_2(a,b)|$



(c) $|DWT_{\widetilde{x}_2}(a,b)|$

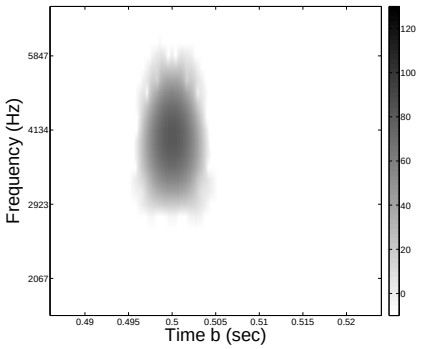


Fig. 8: As in Fig. 7 but for test signal $x_2(t)$.

(a) $|X_3(a,b)|$ **(b)** $|\widetilde{X}_3(a,b)|$

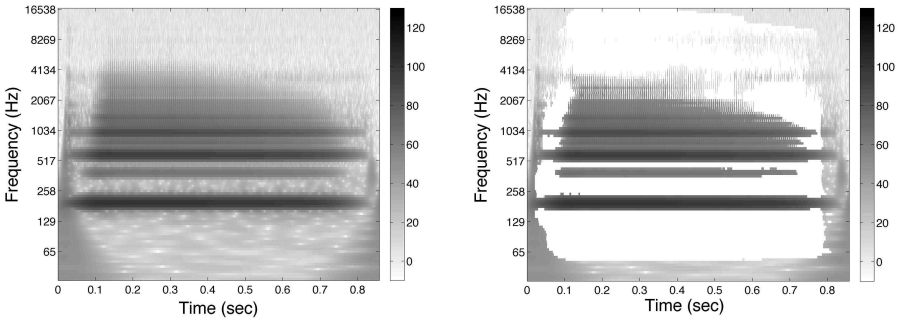


Fig. 9: As in Fig. 8 but for a musical sound ($x_3(t)$, clarinet note A3).

data) in the signal representation while causing little or no audible difference to the original sound after re-synthesis. The result obtained with a musical sound is more challenging: although the filter removed about 50% components, informal listening indicated that the output signal was not perceptually identical to the input signal. This suggests that the filter removed some *relevant* information. A possible explanation can be attributed to the methods employed itself. Indeed, the present algorithm removes components by setting their amplitudes to zero (Step 3). This operation affects the phase relationships between components, which can in turn result in audible effects in the reconstructed signal. Accordingly, the noisy attack of the re-synthesized clarinet sound is likely to result from phase effects. Moreover, it has to be considered that removing a component in a CWT is tricky. Because of the strong correlation between components (as a consequence of the reproducing kernel), removing a component in a CWT affects other components remote from that component and leads to a new representation which is not a CWT anymore (*i.e.*, it does not satisfy the reproducing kernel property). We were conscious of this problem when developing the algorithm but we found worth trying it. In [3], this problem was addressed using a conservative approach: to compensate for the inaccuracies of the masking model, the masking function was shifted in level by an amount determined experimentally. An alternative approach could consist in encoding the masked components on a smaller number of bits than the audible components, as currently done in perceptual audio codecs [32].

Furthermore, the determination of the irrelevance threshold is roughly dependent on the discretization of the CWT. The highly redundant sampling grid we opted for in the present study is likely to have caused an overestimation of masking. In future works, a more appropriate discretization should be chosen so as to represent a Gaussian with a shape factor $\mu = 0.15$ by a single atom. Such a discretization could be, for example, a dyadic grid [35]. Another possibility could be to exploit the recent theory on non-stationary Gabor frames [2, 15]. To overcome the limitation of fixed resolution in the Gabor transform, the non-stationary Gabor transform provides some freedom of evolution of resolution either in time or frequency and allows perfect reconstruction. This constitutes an interesting background for implementing our TF masking data and will be investigated in future works.

To avoid the constraint on the discretization, the discrete masking kernel could be replaced by an explicit function of TF masking allowing the prediction of masking for any TF coordinates. Furthermore, the masking kernel was designed based on TF masking data for a single Gaussian masker with fixed frequency and level. Because masking is highly dependent on frequency and level, additional data are required to develop a model able to accurately predict masking in real-world sounds. Studies are currently underway (*e.g.*, [19]) that investigate the additivity of masking arising from multiple Gaussian maskers shifted in time and frequency. It would be interesting to explore the extent to which these new data on the additivity of TF masking can be incorporated in the current algorithm. Combining data on the frequency- and level-dependency of spectral masking for

Gaussian atoms gathered in [23] and literature data on the level-dependency of temporal masking may allow designing a level-dependent TF masking kernel or function.

4 Summary and Conclusions

In this paper, the question of the development of a perfectly invertible audio signal representation being as close as possible to "what we see is what we hear" was addressed, with specific considerations to TF representations and auditory masking. The proposed approach consisted in predicting the audibility of each TF atom in the TF representations of sounds based on psychoacoustical data on TF masking. To achieve this approach, data on the spread of TF masking produced by a TF atom (*i.e.*, a signal with maximal concentration in the TF plane) were required. Because (i) a few psychoacoustical studies investigated TF masking and (ii) those studies used stimuli with temporally broad supports, their results could not be used to derive the spread of TF masking for one atom.

Therefore, three psychoacoustical experiments were conducted that involved Gaussian-shaped sinusoids with fixed bandwidth (ERB = 600 Hz) and duration (ERD = 1.7 ms) both as masker and target stimuli. The target was shifted either along the time axis, the frequency axis, or both relative to the masker. The same group or subgroup of listeners participated in all three experiments. The conclusions that can be drawn from our data are:

(i) The superposition of the temporal and spectral masking functions does not provide an accurate representation of the measured TF masking function for a Gaussian masker;
(ii) This suggests that audio coding algorithms using such an approach provide rather erroneous predictions of TF masking and that our data may allow to improve the estimation of TF masking in these systems;
(iii) These new data constitute a crucial basis for the prediction of auditory masking in the TF representations of sounds.

We proposed an algorithm (referred to as the "extended irrelevance filter") for removing the inaudible atoms in the wavelet transform of a sound while causing little or no audible difference to the original sound after re-synthesis. Preliminary results obtained with deterministic and musical signals are promising. Future works will include: development of a level-dependent model of TF masking, incorporation of the nonlinear additivity of masking, replacement of the CWT by the non-stationary Gabor transform, refinement of the methods to remove the inaudible components, and perceptual validation of the algorithm with calibrated natural sounds.

Acknowledgments. This work was supported by grants from Egide (PAI "Amadeus" WTZ 1/2006), the ANR (project "SenSons"), and the WWTF (project "MulAc" MA07025).

References

1. Agerkvist, F.T.: A time-frequency auditory model using wavelet packets. J. Audio Eng. Soc. 44(1/2), 37–50 (1996)
2. Balazs, P., Dörfler, M., Holighaus, N., Jaillet, F., Velasco, G.: Theory, implementation and applications of nonstationary Gabor frames. J. Comput. Appl. Math. (2011, in press)
3. Balazs, P., Laback, B., Eckel, G., Deutsch, W.A.: Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking. IEEE Trans. Audio Speech Lang. Process. 18(1), 34–49 (2010)
4. Daubechies, I.: Ten Lectures on Wavelets. CMB-NSF Lecture Notes nr. 61, SIAM, Philadelphia, first edn. (1992)
5. Delgutte, B.: Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers. J. Acoust. Soc. Am. 87(2), 791–809 (February 1990)
6. Duifhuis, H.: Consequences of peripheral frequency selectivity for nonsimultaneous masking. J. Acoust. Soc. Am. 54(6), 1471–1488 (1973)
7. Fastl, H.: Temporal masking effects: III. Pure tone masker. Acustica 43(5), 282–294 (1979)
8. Florentine, M.: Level discrimination of tones as a function of duration. J. Acoust. Soc. Am. 79(3), 792–798 (March 1986)
9. Glasberg, B.R., Moore, B.C.J.: Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds. J. Audio Eng. Soc. 53(10), 906–918 (October 2005)
10. Gröchening, K.: Foundations of time-frequency analysis. Birkhaüser, Boston, first edn. (2001)
11. Hamdi, K.N., Ali, M., Tewfik, A.H.: Low bit rate high quality audio coding with combined harmonic and wavelet representations. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP'96). vol. 2, pp. 1045–1048 (1996), Atlanta, GA, USA
12. He, X., Scordilis, M.S.: Psychoacoustic music analysis based on the discrete wavelet packet transform. Res. Let. Signal Process. 2008(4), 1–5 (January 2008)
13. van der Heijden, M., Kohlrausch, A.: Using an excitation-pattern model to predict auditory masking. Hear. Res. 80, 38–52 (1994)
14. Huang, Y.H., Chiueh, T.D.: A new audio coding scheme using a forward masking model and perceptually weighted vector quantization. IEEE Trans. Audio Speech Lang. Process. 10(5), 325–335 (July 2002)
15. Jaillet, F., Balazs, P., Dörfler, M.: Nonstationary Gabor frames. In: Proc. of the 8th international conference on Sampling Theory and Applications (SAMPTA'09) (May 2009), Marseille, France
16. Jeong, H., Ih, J.: Implementation of a new algorithm using the STFT with variable frequency resolution for the time-frequency auditory model. J. Audio Eng. Soc. 47(4), 240–251 (1999)
17. Jepsen, M., Ewert, S.D., Dau, T.: A computational model of human auditory signal processing and perception. J. Acoust. Soc. Am. 124(1), 422–438 (2008)
18. Kidd Jr., G., Feth, L.L.: Patterns of residual masking. Hear. Res. 5(1), 49–67 (1981)
19. Laback, B., Balazs, P., Necciari, T., Savel, S., Meunier, S., Ystad, S., Kronland-Martinet, R.: Additivity of nonsimultaneous masking for short Gaussian-shaped sinusoids. J. Acoust. Soc. Am 129(2), 888–897 (February 2011)
20. Moore, B.C.J.: An introduction to the psychology of hearing. Academic Press, London, fifth edn. (2003)

21. Moore, B.C.J., Alcántara, J.I., Glasberg, B.R.: Behavioural measurement of level-dependent shifts in the vibration pattern on the basilar membrane. Hear. Res. 163, 101–110 (2002)
22. Moore, B.C.J., Alcántara, J.I., Dau, T.: Masking patterns for sinusoidal and narrow-band noise maskers. J. Acoust. Soc. Am. 104(2), 1023–1038 (1998)
23. Necciari, T.: Auditory time-frequency masking: Psychoacoustical measures and application to the analysis-synthesis of sound signals. Ph.D. thesis, University of Provence Aix-Marseille I, France (October 2010)
24. O'Donovan, J.J., Dermot, J.F.: Perceptually motivated time-frequency analysis. J. Acoust. Soc. Am. 117(1), 250–262 (2005)
25. Oxenham, A.J.: Forward masking: Adaptation or integration? J. Acoust. Soc. Am. 109(2), 732–741 (February 2001)
26. Patterson, R.D., Allerhand, M.H., Giguère, C.: Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. J. Acoust. Soc. Am. 98, 1890–1894 (1995)
27. Plack, C.J., Oxenham, A.J., Drga, V.: Linear and nonlinear processes in temporal masking. Acta Acust. united Ac. 88(3), 348–358 (2002)
28. Plack, C.J., Oxenham, A.J.: Basilar-membrane nonlinearity and the growth of forward masking. J. Acoust. Soc. Am. 103(3), 1598–1608 (March 1998)
29. Robles, L., Ruggero, A.: Mechanics of the mammalian cochlea. Physiol. Rev. 81(3), 1305–1352 (July 2001)
30. van Schijndel, N.H., Houtgast, T., Festen, J.M.: Intensity discrimination of Gaussian-windowed tones: Indications for the shape of the auditory frequency-time window. J. Acoust. Soc. Am. 105(6), 3425–3435 (1999)
31. Soderquist, D.R., Carstens, A.A., Frank, G.J.H.: Backward, simultaneous, and forward masking as a function of signal delay and frequency. J. Aud. Res. 21, 227–245 (1981)
32. Spanias, P., Painter, T., Atti, V.: Audio Signal Processing and Coding. Wiley-Interscience, Hoboken, New Jersey, USA (2007)
33. Terhardt, E.: Calculating virtual pitch. Hear. Res. 1, 155–182 (1979)
34. Vafin, R., Andersen, S.V., Kleijn, W.B.: Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP'00). vol. 2, pp. 901–904 (2000), Istanbul, Turkey
35. Vetterli, M., Kovačević, J.: Wavelets and subband coding. Prentice Hall PTR, Englewood Cliffs, New Jersey (1995)
36. Zwicker, E.: Dependence of post-masking on masker duration and its relation to temporal effects in loudness. J. Acoust. Soc. Am. 75(1), 219–223 (1984)