Introducing Time-Frequency Sparsity by Removing Perceptually Irrelevant Components Using a Simple Model of Simultaneous Masking

Peter Balazs, *Member, IEEE,* Bernhard Laback, Gerhard Eckel, and Werner A. Deutsch *Member, IEEE,*

***Abstract*—We present an algorithm for removing time-frequency components, found by a standard Gabor transform, of a "real-world" sound while causing no audible difference to the original sound after resynthesis. Thus this representation is made sparser. The selection of removable components is based on a simple model of simultaneous masking in the auditory system. Important goals were the applicability to any real-world music and speech sound, integrating mutual masking effects between time-frequency components, coping with the time-frequency spread of such an operation, and computational efficiency. The proposed algorithm first determines an estimation of the masked threshold within an analysis window. The masked threshold function is then shifted in level by an amount determined experimentally, and all components falling below this function (the irrelevance threshold) are removed. This shift gives a conservative way to deal with uncertainty effects resulting from removing time-frequency components and with inaccuracies in the masking model. The removal of components is described as an adaptive Gabor multiplier. Thirty-six normal hearing subjects participated in an experiment to determine the maximum shift value for which they could not discriminate the irrelevance filtered signal from the original signal. On average across the test stimuli, 36 percent of the time-frequency components fell below the irrelevance threshold.**

***Index Terms*—simultaneous masking; irrelevance filter; spectral masking; sparse representation; Gabor filter; Gabor transform; time-variant filter; efficient algorithm; masking model;**
EDICS : AUD-AUDI Auditory Modeling and Hearing Aids, AUD-ACOD Broadband and Perceptual Coding; AUD-ANSY Audio Analysis and Synthesis

I. Introduction:

It is known in psychoacoustics that not all time-frequency components of a "real-world" acoustic signal can be perceived by the human auditory system. More precisely, it turns out that some time-frequency components mask other components, which are close in the time-frequency domain. Deleting these masked and thus perceptually irrelevant components makes the signal representation more sparse and the resynthesized signal would be expected to sound equivalently to the original signal.

A well-known technique to reduce the digital size of an audio file, the *MP3* audio codec [1], is based on a model of human auditory perception. This and similar perceptual audio codecs like AAC (see [2] for a review), allocate low bit rates to frequency channels which are subject to masking effects and thus have little or no perceptual relevance. This technique is very efficient in reducing the capacities required for transmitting and storing audio files.

The goal of the algorithm presented here, referred to as the *"irrelevance filter"*, is not to reduce the digital size of a sound. Rather, its goal is to remove those time-frequency components in a standard Gabor transform, whose removal causes no audible difference to the original signal after resynthesis. Note the difference to perceptual audio codecs; they use a low bit depth and thus introduce quantization noise in frequency bands were the signal falls below the masked threshold. In contrast, in the proposed model we want to either keep a component or remove it if irrelevant. Thus, we attempt to introduce "silence" in bands were the signal falls below the irrelevance threshold. In other words, the algorithm searches for a time-frequency representation, which is sparser but perceptually equivalent to the original representation after resynthesis.

The algorithm should work for most 'every-day' sounds, i.e. real-world music and speech signals, and no calibration should be necessary.

The proposed algorithm uses a simple model of simultaneous masking (also referred to as spectral masking) which is based on data from the psychoacoustic literature (see section III.A). The properties of simultaneous masking for simple stimuli (such as sinusoids or bandpass noises) have been studied extensively (see reviews by [3] and [4]). A basic model for the simultaneous masking effect, referred to as the *excitation pattern* model of masking [5], [6], [7], is that the auditory system can detect a target presented simultaneously with a masker only if the excitation pattern of target plus masker significantly differs from that of the masker alone. If the two excitation patterns do not differ in a way detectable by the auditory system the target cannot be perceived, it is masked [7]. This basic model allows for the prediction of the masked threshold of a target signal in the presence of a masker signal [7], with certain constraints upon the stimuli. The masked threshold is defined as the minimum level of the target at

Manuscript received June 26, 2008; revised November 14, 2008. This work was partly supported by the WWTF project MULAC ('Frame Multipliers: Theory and Application in Acoustics; MA07-025).

P. Balazs, B. Laback, and W. Deutsch are with the Acoustics Research Institute of the Austrian Academy of Sciences, Wohllebengasse 12-14, A-1040 Wien, Austria. (e-mail: {Peter.Balazs, Bernhard.Laback, Werner.Deutsch}@oeaw.ac.at)

G. Eckel is with the Institute of Electronic Music and Acoustics of the University of Music and Dramatic Arts, Infeldgasse 10/3, A-8010 Graz, Austria. (e-mail: eckel@iem.at)

which it is audible in the presence of the masker. In sections II.D and II.E we provide an overview of different properties of auditory masking and of different modeling approaches that have been shown to be successful in predicting simultaneous masking effects.

The concept of the excitation pattern has also been used in perceptual audio codecs to predict masking effects caused by individual spectral components of music or speech sounds (e.g. [8], [9], [2], [10], [11]). The aim is to calculate the masked threshold in each frequency channel of the analysis-resynthesis system to obtain a measure for the maximum tolerable level of the quantization noise in the respective channel. The level of the quantization noise is controlled by the allocated bit depth. In order to determine if the quantization noise in a given channel is audible or not, the quantization noise in that channel is considered as target and the total input sound is considered as masker. According to the excitation pattern model of masking, the target (i.e. the quantization noise) is considered to be audible as long as adding it to the masker (i.e. the total input sound) results in a significant change in the corresponding excitation pattern. This process is repeated for each channel to obtain an estimate of the bit depth required in each channel. In this way, reducing the bit depth for frequency channels that are perceptually less relevant due to masking effects allows to reduce dramatically the digital size required for encoding without quality loss compared to encoding at a fixed bit depth ([2]).

Given this knowledge from the literature, an apparently straight-forward solution to implement the irrelevance filter algorithm would be to first identify the components which are masked by other components, using a masking model like the excitation pattern model, and then to re-synthesize only those components that are not subject to masking. However, it became clear after initial considerations and heuristic pretests that using this approach is often not successful, since the resynthesized signal could often be discriminated from the original signal. Obviously, an important requirement of the irrelevance filter was not satisfied, i.e., the auditory representation of the filtered signal differs from that of the original signal. One problem leading to this is related to a general property of time-frequency representations. Removing a component will cause changes at time-frequency locations remote from that component. This can in turn lead to changes in the mutual masking effects between the components and thus in the auditory percept compared to the original sound.

Another problem arises from the application of the concept of the excitation pattern model of masking to identify and subsequently remove masked components from real-world signals. For such signals most often more than one components are subject to masking at a time, thus should be removed. This in turn leads to a violation of the assumption of the excitation pattern model that *all* components except for the target component are considered as maskers. The probable result is a lowered total masking effect within the resynthesized signal compared to the original signal for which the masking model has been applied and thus an audible difference between the two. In other words, removing more than one component at a time can result in unpredictable masking effects.

In summary, the irrelevance filter approach has properties and requirements that differ from those of established models that are used to predict the simultaneous masking effect of one signal on another signal. Let us stress the difference between an irrelevance and a masking approach again. A masking model gives an indication if adding a second signal (target) to a given signal (masker) can be perceived or not. In comparison an irrelevance model gives an estimation which of the components of the signal can be removed. In this paragraph let us use the word 'component' in the most general way as a, possibly complex, part of an additive synthesis model. While it is possible to use a masking approach for a two-component signal to determine if one of the two components is irrelevant or not, for a multi-component signal this would require the comparison of all possible combinations of two sets of components, as there is no clear distinction between target and masker. Such an iterative approach would be very time-consuming even for a small number of components. If no a-priori signal model (with only a few components) can be assumed, but instead a signal-independent representation like a Gabor or wavelet representation has to be chosen, this leads to a lot of components and a very infeasible scheme.

In order to deal with the specific problems associated with removing components from a signal, the following strategy was pursued. First, the masked threshold function was calculated, representing the basic simultaneous masking effect as described above. Then, the masked threshold function was shifted in level by a certain amount corresponding to the results of a perceptual experiment and all components falling below the shifted function (the irrelevance threshold) are removed. [1] At the level shift determined the subjects could not discriminate the irrelevance filtered signal from the original signal. Using this approach allows to cope with the uncontrolled effects of the above-mentioned properties associated with the removal of spectral components. Furthermore, it allows to cope with inaccuracies of the masking model itself. The masking model chosen for the current algorithm is simple considering the nonlinearities and complex interactions involved in auditory masking for real-world sounds. These include nonlinear additivity of masking [12], [13], across-frequency integration in signal detection [14], [15], and suppression [16], [17]. In this way a conservative criterion is provided for deciding which components can be removed without any audible difference to the original sound.

The primary task of such an irrelevance filter algorithm is to remove components of a Gabor transform which cannot be perceived by the human auditory system due to mutual masking effects. Since the masking effect depends on the signal itself, which is variable across time, it can be modeled as an adaptive non-linear filtering process. The process can be split into two steps, first the adaptive calculation of the operator G and second its application on the signal.

$$1.)\ x \mapsto G(x) \quad 2.)\ x \mapsto G(x) \cdot x \ .$$

[1]Note that both the masked threshold function value at a given Gabor coefficient and the level of the time-frequency component, with which it is compared to, represent levels that are integrated over a spectro-temporal region.

The first step is non-linear, whereas the second step is linear. The following section describes the time-variant filtering process known as a Gabor multiplier [18] as well as Gabor filter [19]. Then, the psychoacoustical background, the existing masking models, and the implementation of the algorithm are described. The last section describes the perceptual experiment to control the free parameter, the level shift of the irrelevance threshold, using the criterion of the discriminability between the original and the filtered signal.

The irrelevance filter presented here has first been developed and tested in [20] and implemented in ST^X [21], a signal processing software system designed at the Acoustics Research Institute of the Austrian Academy of Sciences. Practical experience indicates that the irrelevance filter is a very effective algorithm for real-world music and speech sounds. The resulting signal representation is more sparse and easier to interpret. Main applications are the facilitation of the synthesis of sounds and the ease of the interpretation of time-frequency properties of signals used in perception-related tasks.

II. Background: Notation and Preliminaries

A. Frames

To introduce the issue of perfect resynthesis a short summary of frame theory [22], [23] is given. Frame theory has been recognized as being important for signal processing in recent years [24], [25].

Let $\mathcal{H}$ be a Hilbert space $\mathcal{H}$ with inner product $\langle .,. \rangle$. The sequence (ψ_k) of elements in $\mathcal{H}$ for $k = 1, 2, \ldots$ is called a *frame*, if constants $A, B > 0$ exist, such that

$$A \cdot \|f\|^2 \leq \sum_k |\langle f, \psi_k \rangle|^2 \leq B \cdot \|f\|^2 \ \forall \ f \in \mathcal{H}. \qquad (1)$$

The constants A and B are called *lower frame bound* and *upper frame bound*, respectively. If these bounds can be chosen such that $A = B$, then the frame is called *tight*. If $A \approx B$, the frame is called *snug*. The operator $S : \mathcal{H} \to \mathcal{H}$, defined by

$$S(f) = \sum_k \langle f, \psi_k \rangle \cdot \psi_k \ \forall f \in \mathcal{H}$$

is called the *frame operator*. For every frame, the frame operator is self-adjoint, positive and invertible.

Frames are useful for signal processing applications, because they allow perfect reconstruction. With $\left(\tilde{\psi}_k\right) := \left(S^{-1}\psi_k\right)$, the so-called *canonical dual frame* every signal f can be expanded to

$$f = \sum_{k \in K} \left\langle f, \tilde{\psi}_k \right\rangle \psi_k \text{ and } f = \sum_{k \in K} \langle f, \psi_k \rangle \tilde{\psi}_k.$$

For a tight frame this reconstruction has a very simple form, $f = \frac{1}{A} \sum_{k \in K} \langle f, \psi_k \rangle \psi_k$. For snug frames this simple reconstruction given above gives a very good approximation of f.

This means that for a system, e.g. a filterbank, that constitutes a frame, we can find the perfect synthesis system by calculating the inverse frame operator and applying it on the original system. There are some algorithms for calculating this inversion in an efficient way [22], [26]. For tight and snug frames no complex calculation is needed.

B. Time-Frequency Analysis

The *Fourier Transformation*, denoted by $f \mapsto \hat{f}$, is a well known mathematical tool to analyze the frequency content of a signal f. Thanks to the very efficient *fast Fourier transformation (FFT)* [27], many discrete applications and developments have been made feasible. Listening to a sound, a voice or music, a listener does not hear spectral components and their amplitudes only, but also their dynamic evolution. A well known algorithm for a time-frequency representation is the *Short Time Fourier Transformation, STFT* [28]. One way to look at the STFT is to multiply the signal $f(t)$ with a window function $g(t - \tau)$ to obtain a version of the signal that is concentrated at the time τ (if the window is chosen accordingly). Then the Fourier transform is applied to the result:

$$STFT_g(f)(\tau, \omega) = \int_{-\infty}^{\infty} f(t) \overline{g(t-\tau)} e^{2\pi i \omega t} dt.$$

This can also be seen as a projection of the signal $f(x)$ on the time-frequency shifted Gabor atom $M_\omega T_\tau g(t)$, where T denotes the *translation* operator (i.e. $(T_\tau f)(t) = f(t-\tau)$) and M the *modulation* operator (i.e. $(M_\omega f)(t) = e^{2\pi i \omega t} f(t)$):

$$STFT_g(f)(\tau, \omega) = \langle f, M_\omega T_\tau g(t) \rangle .$$

If the STFT is not considered for continuous variables ω and τ, but in a sampled version, it is called a *Gabor transform*. A *Gabor system* with time shift parameter H, also called *hop size*, and frequency shift parameter ω_0 is given by:

$$\mathcal{G}(g, H, \omega_0) = \{M_{\omega_0 \cdot l} T_{H \cdot k} g : k, l \in \mathbb{Z}\} =$$
$$= \{e^{2\pi i \omega_0 l x} g(x - k \cdot H) : k, l \in \mathbb{Z}\}.$$

The Gabor transform is the projection on the Gabor system. The equivalence between Gabor analysis and corresponding filter-banks is a well-known fact [25].

From time-frequency representations like wavelet analysis [29] and the Wigner Ville representation [30] the STFT respectively the Gabor transform has been chosen as it is a linear transformation and fast algorithms directly connected to the frequency domain are available.

1) Synthesis: In order to perform modifications of a signal, a analysis system as well as a synthesis system is needed. The continuous STFT has an inverse, but it cannot be handled efficiently as it involves weak integrals. A series expansion, like in the case of Fourier series, would be much more desirable, both as a model and in an algorithmic sense. So a Gabor transform is used. If the *Gabor system* $\mathcal{G}(g, H, \omega_0)$ forms a frame, then every function is represented as infinite but discrete linear combination:

$$f(t) = \sum_{k,l \in \mathbb{Z}} STFT_g(f)(k \cdot H, l \cdot \omega_0) \cdot \left(e^{2\pi i l \omega_0 t} \tilde{g}(t - k \cdot H)\right)$$

for the canonical dual window $\tilde{g}$. Contrary to the Fourier transform the Gabor transform with a good time-frequency concentration cannot fulfill a orthonormal basis property, which leads to a conceptional difference between these two concepts. In particular synthesis coefficients are not unique anymore, and

perfect reconstruction can not always be achieved by using the same elements for analysis and synthesis. The Gabor frame theory provides a method to calculate a perfect reconstruction window and there are several algorithms available to do that in an efficient way [31], [32].

2) Gabor Filters: Time-invariant filtering has been used for many years [33]. This technique transforms the signal into the frequency domain and multiplies the spectrum with a fixed function. A generalization of this procedure is *time-variant filtering*, which has attracted more and more attention in the past years. Gabor multipliers, called *Gabor filters* in terms of signal-processing [19] or *time-frequency masks* in computational auditory scene analysis [34], are particular cases of time-variant filters. The signal to be processed is transformed into the time-frequency domain, the resulting coefficients are multiplied by a function on the same domain and the result of this multiplication is resynthesized. More formally, we can define [35], [18]:

Definition 1: Let $\mathcal{G}(g, H, \omega_0)$ be a Gabor frame. For a bounded sequence $(m_{(k,l)})$ of complex numbers, called the *symbol*, the *Gabor multiplier* or *Gabor filter* $\mathbf{G}$ is defined as the operator

$$\mathbf{G}f(t) = \sum_{k,l\in\mathbb{Z}} m_{(k,l)} STFT_g(f)(k\cdot H, l\cdot\omega_0)\cdot\left(e^{2\pi i l\omega_0 t}\tilde{g}(t-k\cdot H)\right).$$

Other possibilities to define and implement time variant filters, among them the Zadeh and Weyl Filter [19], have been considered. Finally the advantages of a Gabor filter are the following [19] :

(i) Easy implementation; the Gabor coefficients are multiplied and then resynthezised.
(ii) Computational efficiency compared to full STFT filters.
(iii) Easy interpretation in the time-frequency plane; the complex values at the time-frequency sampling points are simply multiplied.
(iv) Small time-frequency spread; only small time-frequency shifts are introduced with accordingly chosen windows (i.e. it is an underspread operator). This property can be improved, if the redundancy (see Section II-B4) is increased.

3) Discrete, Finite-Dimensional Data: Regarding the current application finite dimensional, discrete signals of length n are considered only, such as vectors denoted by $x = (x_0, x_1, \ldots, x_{n-1}) \in \mathbb{C}^n$. These vectors are regarded as periodic functions on $\mathbb{Z}$ (with period n), so $x_{i+k\cdot n} = x_i$ for all $i, k \in \mathbb{Z}$. In this case, the modulation and time shift operators are discretized, i.e., $T_l x = (x_{n-l}, x_{n-l+1}, \ldots, x_0, x_1, \ldots, x_{n-l-1})$ and $M_k x = \left(x_0\cdot W_n^0, x_1\cdot W_n^{1\cdot k}, \ldots, x_{n-1}\cdot W_n^{(n-1)k}\right)$ with $W_n = e^{\frac{2\pi i}{n}}$. As all vectors are periodic, the translation is a cyclic operator. In applications, as the one presented here, such vectors are normally samples of a continuous function. The indices correspond to the number of samples, so in this setting no units have to be used.

The convolution of two vectors in $\mathbb{C}^n$ is defined by

$$(x * y)_k = \sum_{i=0}^{n-1} x_i \cdot y_{k-i}.$$

As we regard all vectors as periodic, this is the cyclic convolution. The discrete Fourier transform convolution of two vectors corresponds to the element-wise multiplication of their discrete Fourier transform $\widehat{x * y} = \hat{x}\cdot\hat{y}$, where we denote the *Discrete Fourier Transformation (DFT)* again by $x \mapsto \hat{x}$.

We will consider the Gabor system $\mathcal{G}(g,a,b) = \left\{M_{bn}T_{ak}g : k = 0,\ldots,\tilde{a}; n = 0,\ldots,\tilde{b}\right\}$ for the window g. The parameters, a and b, are restricted to be factors of n such that the numbers $\tilde{a} = \frac{n}{a}$ and $\tilde{b} = \frac{n}{b}$ are integers. That is equivalent to sampling with period T and setting $\omega_0 = \frac{b}{nT}$ and $H = aT$. Note that $\tilde{b}$ denotes the number of frequency bins, written as $\tilde{b} = N_{FFT}$. In the discrete, finite-dimensional case, the Gabor frame operator has a special structure, the matrix S is zero except in every $\tilde{b}$-th side-diagonals and these side-diagonals are periodic with period a. This property can be directly seen by using the *Walnut representation* [36] of the Gabor frame matrix $S = (S_{p,q})_{n,n}$:

Theorem 2:

$$S_{p,q} = \begin{cases} \tilde{b}\sum_{k=0}^{\tilde{a}-1} \overline{g}_{p-ak}\cdot g_{q-ak} & \text{for } p-q \equiv 0 \bmod \tilde{b} \\ 0 & \text{otherwise} \end{cases}.$$

In this setting the Gabor transform at the sample l and the frequency bin k can be written as:

$$Gab(x)_{l,k} = STFT_g(x)_{la,kb} = \sum_{m=1}^{n} x_m g_{m-la} e^{\frac{-2\pi i k m}{N_{FFT}}}.$$

In this paper we also want to stress the formal distinction between Gabor transform and the full STFT even for the discrete case $\mathbb{C}^n$. The STFT can be seen as a limit case of the Gabor transform for $a = b = 1$ (samples). [2] The discrete STFT is always invertible with any window which is not perpendicular to the original one, in particular for standard windows (with positive values) this is always possible. For a Gabor transform a frame condition has to be checked to guarantee perfect reconstruction. Furthermore not every window allow reconstruction.

For the full STFT, although no weak integrals have to be considered as in the continuous case, it is still numerically infeasible as it involves n^2 data points. Using the Gabor transform the data points are reduced to $\frac{n^2}{ab}$. In case of small a and b still 'too many' data points have to be dealt with, in the sense described in the next point:

4) Redundancy: In practice of analysis-synthesis systems, reducing the amount of computation is essential, which means reducing the *redundancy* of the representation. The redundancy for a discrete Gabor transform is given by $red = \frac{n}{a\cdot b} = \frac{N_{FFT}}{a}$. compared to the limit case of the discrete STFT where $red = \frac{n}{1\cdot 1} = n$. This motivates why this value is called redundancy, as a signal of length n is represented by the STFT with n^2 data points.

Only if the Gabor system forms a frame the frame expansion can be applied to obtain perfect reconstruction. Gabor [37] proposed that in the case of Gaussian windows the redundancy

[2] This distinction is not always made. Sometimes the Gabor transform is also called a STFT, even if $a \neq 1$ and $b \neq 1$.

could be reduced to $red = 1$. It can be shown that the Gaussian windows constitute a *frame* for $red > 1$ [28]. The question if certain windowing functions form frames for certain redundancies has already been answered for many systems. It is clear that there is a kind of *"Nyquist criterion"* for Gabor frames, as it has been shown that no window function can be a frame for $red < 1$ [28].

Another approach to the concept of redundancy could be taken to include a perceptual viewpoint. If one is interested in perceptual feature extraction, any part of an audio signal that cannot be heard can be considered as redundant. If the signal is reduced to the perceptually relevant parts only, the representation can be made more sparse, compared to the perfect reconstruction scheme.

C. Physiological Background of Masking and Irrelevance

An comprehensive review of the physiology of the auditory system and psychoacoustics can be found in [38] respectively [3]. Sound waves arriving at the ear spread through the ear canal, pass the middle ear, and finally reach the *cochlea*. The mechanical vibrations are transformed into electrical action potentials to be transmitted to the brain via the auditory nerve. The transformation is performed by the hair cells located on the basilar membrane (BM), whose vibration pattern resembles a traveling wave, moving from the oval window to the apex of the cochlea. Maxima of vibration occur near the oval window for high-frequency tones and near the apex for low-frequency tones. This correspondence of frequency to place on the basilar membrane is called *tonotopy*. Physiological measurements of basilar membrane motion and auditory nerve activity have shown that the frequency of a pure tone is encoded temporally as well as tonotopically. Therefore, different spectral components of a broad-band signal end up in different neural channels of the auditory nerve.

The frequency selectivity of the auditory system, i.e. the ability to separate closely spaced frequency components, is limited. This can be understood by considering the activation pattern of the basilar membrane caused by a sinusoid. This so-called excitation pattern has the maximum at a specific place along the BM and decays towards both sides (see Figure 3). This implies that the sinusoid activates also neighboring areas at the BM. The slopes of the excitation pattern depend nonlinearly both on the absolute frequency and the amplitude of the signal. With increasing frequency region the spacing between the characteristic places in mm corresponding to a given frequency distance in Hz decreases [39]. The slopes of the excitation pattern in BM deflection per tonotopic distance (both in mm), however, are constant as a function of signal frequency.

Psychoacoustic frequency scales have been derived experimentally, reflecting the nonlinear mapping function of the signal frequency. The so-called Bark scale function[3] can be expressed analytically [41] as

$$b(f_{\mathrm{kHz}}) := 13\ \tan^{-1}(0.76 \cdot f_{\mathrm{kHz}}) + 3.5\ \tan^{-1}(\frac{{f_{\mathrm{kHz}}}^2}{7.5}), \quad (2)$$

[3]Another well-described scale is the so-called Equivalent Rectangular Bandwidth, ERB scale (e.g. [5]). The two scales are conceptually similar. see e.g. [40]

where f_{kHz} is the frequency in kHz. In a first approximation, the Bark scale resembles the tonotopy (refer to Figure 1).

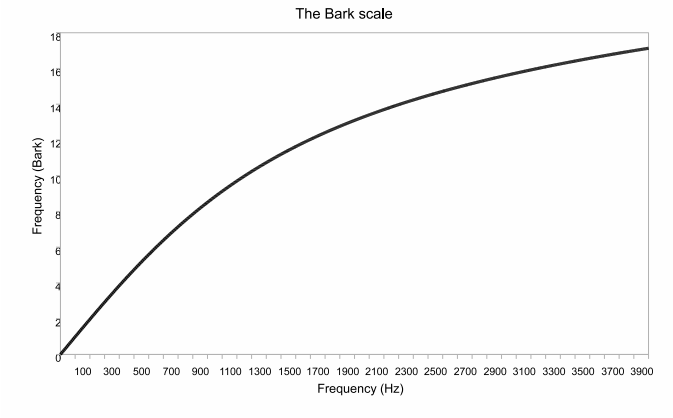

Fig. 1. The Bark scale: plot corresponding to Equation 2

Applying the Bark scale, the slopes of the excitation pattern of a sinusoid are constant as a function of the absolute frequency. This scale is based on the concept of *critical bands* in audition, which is related to the auditory filters. A basic definition of the critical band states that spectral components interact fundamentally differently within a critical band than across critical bands.

D. Masking

Masking refers to the process by means of which the threshold of audibility for one sound (the target) is raised by the presence of another sound (the masker). Masking can render the masked sound inaudible. Masking occurs in two main signal configurations; simultaneous occurrence of target and masker is referred to as *simultaneous, frequency* or *spectral masking* [3]; non-simultaneous occurrence of target and masker is referred to as *temporal masking*, e.g. [42].

Real-world sounds are broadband and therefore involve mutual masking effects between the individual narrow-band components into which the signal can be decomposed. This raises the question how the masking effects of more than one simultaneous masker on a target add up. To a first approximation, the masked thresholds elicited by two individual maskers have to be added linearly in the power domain to derive the combined masked threshold [12]. For two equally effective maskers this means that the masked threshold in the presence of both maskers is 3 dB higher than that for one masker alone. This rule may apply if side effects are ruled out, such as the detection of cochlear combination products, the detection of the target at a tonotopic place aside from the target frequency (so-called off-frequency listening: [43]), or listening for the signal in minima of the temporal envelope of the masker [12], [13], [44]. In many configurations, however, the additivity of masking can be larger than according to the linear addition rule; in [44] it has been shown that for spectrally non-overlapping maskers nonlinear additivity is the rule. Furthermore, little is known about the additivity of masking for more than two maskers [45].

Another effect complicating the prediction of masking effects for real-world sounds is that the auditory system integrates signal information across frequencies to detect a signal. As an example, for two simultaneously presented sinusoids equally contributing to detection, the masked threshold per sinusoid is about 2.5 dB lower than the masked thresholds for each sinusoid alone [14]. This implies that two (or more) spectral components of a broad-band signal may be audible even if each of them separately is below the masked threshold. In addition, the maximum bandwidth up to which spectral integration is efficient depends on the signal duration [14].

Furthermore, mutual suppression effects between individual spectral components of a sound may reduce the effective masking effect evoked by those components [46].

E. Masking Models

In psychoacoustics, two types of models have been developed that attempt to predict simultaneous masking. The first type are excitation pattern or loudness-based models. These models, in their initial formulation, transform a spectrally defined (thus stationary) signal into an excitation pattern [5]. This approach goes back to the power-spectrum model of masking [47], in which the auditory periphery is conceived as containing a bank of bandpass filters. Masking is then determined by the target-to-masker ratio at the output of the filters. The target is masked if this ratio does not exceed a certain value. Based on this basic approach, in [6] a method has been proposed to predict masking for arbitrary stationary maskers. In [7] this method was used to predict different psychoacoustic measures of simultaneous masking. The only moderate success of this model was attributed mainly to the fact that it does not represent the nonlinear behavior of auditory processing. Variants of this excitation-pattern type of models, intended to predict loudness perception, have been proposed (e.g. [48], [49]). They allow to predict the audibility of a sound in the presence of another sound by the assumption that audibility occurs at a fixed partial loudness. Still another variant of this model type was designed to predict audibility discrimination thresholds for spectral envelope distortions in vowel-like sounds ([50]).

The second type of models ([51], [52], [53], [54], [55]) attempts to simulate the effective signal processing in the auditory system. These models are intended to predict any more peripherally-located auditory effects. The main focus of the family of models presented in ([51], [52], [53], [54]) is, however, on the modeling of masking effects. The last version ([54]) consists of outer- and middle-ear transformations, nonlinear cochlear processing, hair-cell transduction, a squaring expansion, an adaptation stage, a low-pass modulation filter, a bandpass modulation filterbank, a constant-variance internal noise, and an optimal detector stage. The optimal detector stage represents a decision process, where a stored temporal representation of the signal to be detected is compared with the actual activity pattern. Note that the main difference to the excitation pattern-based models, besides the apparent diversity of the processing stages included, is the implementation of the decision process. The optimal detector allows the model to use a priori knowledge about the target in the detection stage, which appears to reflect the underlying process in a real observer. The model is able to accurately predict the result of a large variety of simultaneous and non-simultaneous masking experiments.

In summary, there exist several models that can predict the auditory masking effect. But as mentioned in the introduction a the specific goal of the irrelevance algorithm requires a modified approach, which is described in the next section.

III. The Irrelevance Filter Algorithm

The aim of the proposed algorithm, whose outline is shown in Figure 2, is to remove any components of a music or speech signal which do not contribute to the perception of the sound after resynthesis, i.e. which are *perceptually irrelevant*. This implies that the masking effect of each component on every other component has to be considered. In the introduction section we presented arguments why a model which just removes those components that are subject to masking will not lead to satisfactory results, even if a sophisticated masking model, like the ones presented above, is used. Thus, the current problem required the definition of a new threshold function, fulfilling the following conservative criterion: those components whose amplitudes do not exceed the threshold function can be removed while resulting in no perceptual difference to the original sound. Note that the conventional masked threshold function determines which components are masked whereas the new threshold function determines which components can be removed while causing no audible effects.

This new threshold function, referred to as *irrelevance threshold*, contains a level offset parameter whose optimal setting has been determined in an perceptual experiment with 36 subjects. The level offset allows to cope with uncontrolled effects associated with removing components from a sound. In addition, it allows to manage inaccuracies in the masking model applied, which is simple and unlikely to predict accurately the complex and nonlinear effects involved in masking, particularly in multi-component stimuli. The realization of the concept of the irrelevance threshold is described below.

A. The Spreading Function

The term spreading function is used here to functionally describe the spread of excitation induced by a sinusoid on the BM in the Bark scale [56]. An approximation of the spreading function is a triangle-like function (in the Bark scale with logarithmic amplitudes), see Figure 3. It was used in [20] to formulate a simple model of simultaneous masking and approximated by the function

$$B(\omega) = 13.94 + 1.5 \cdot (\omega + 0.03) - 25.5 \cdot \sqrt{0.3 + (\omega + 0.03)^2} \quad (3)$$

as a combination of two other models found in [56] and [57]. More general, the shape of the spreading function can be modeled by three parameters: the lower frequency slope l and the upper slope u (giving the absolute slope of the left respectively right part of the function in $dB/Bark$), as well as a non-negative parameter e that allows to control the

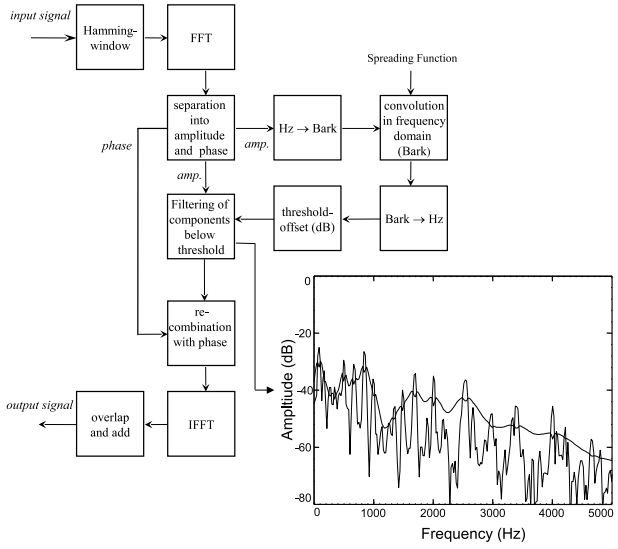

Fig. 2. The main stages of the irrelevance algorithm. The plot inserted into the graph shows a signal spectrum and the calculated irrelevance threshold. Time-Frequency components whose integrated level within a Gabor-bin falls below the irrelevance threshold are filtered out.

smoothness of the function at point zero. This parameter is introduced, as the model should predict the smooth excitation pattern of the BM for a single sinusoid. The function

$$F(x) = \frac{l-u}{2} \cdot x - \frac{l+u}{2} \cdot \sqrt{e+x^2} \tag{4}$$

is used as shape function. As

$$F'(x) = \frac{l-u}{2} - \frac{(l+u)\cdot x}{2\cdot\sqrt{e+x^2}}, \tag{5}$$

we get $\lim_{x\to\infty} F'(x) = -u$ and $\lim_{x\to-\infty} F'(x) = l$ as expected. In [20] the maximum of this function, (x_{max}, y_{max}) is found approximately, but can be calculated analytically as

$$x_{max} = \sqrt{\frac{e}{lu}} \cdot \frac{l-u}{2}, \quad y_{max} = -\sqrt{e\cdot u\cdot l}.$$

A new function $B(x)$, the spreading function, is built, such that the maximum is shifted to the point $(0,0)$.

$$B(\omega) = F\left(\omega + x_{max}\right) - y_{max}. \tag{6}$$

Setting $l = 27$, $u = 24$, $e = 0.3$ in Equation 6 (according to [56] respectively [57]) leads to Equation 3. These parameters have been used in the perceptual experiments described in Section III-D. Note that we included no level-dependency of the spreading function since we wanted to avoid the calibration of the input signal. In the implementation in ST^X[21] varying these parameters allows an heuristic estimation of their influence on the algorithm.

Furthermore, a new function $B_0(\omega)$, called the *spreading threshold kernel*, is introduced. For a given $\epsilon > 0$ we set

$$B_0(\omega) = \begin{cases} -\infty & \omega \in (0-\epsilon/2, 0+\epsilon/2) \\ B(\omega) & \text{otherwise} \end{cases}. \tag{7}$$

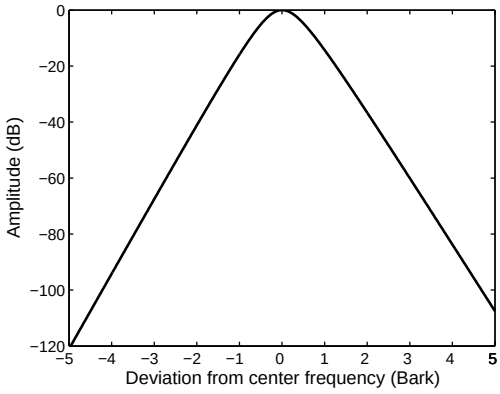

Fig. 3. The spreading function $B(\omega)$ in the Bark domain. Here the parameters ($l = 27$, $u = -24$, $e = 0.3$) are used.

B. The Irrelevance Function

This spreading function is used for the calculation of a threshold function, the irrelevance function, $\mathcal{I}_{l,k}$, in the following way.

1) Calculate the square of the absolute values of a Gabor transform of the signal x, $|Gab(x)_{l,k}|^2$.
2) Transform the columns, power spectra of the signal at regular temporal intervals, into the Bark scale.
3) Convolve with the involuted [4] spreading threshold kernel $B_0(\omega)$ in the Bark domain. Transform back into the Hz scale. Denote this, see also Appendix A, by
$$\left(|Gab(x)_{l,.}|^2 \overset{(b)}{*} B_0(-.)\right)_k.$$
4) Weight the result by the relative bandwidth $\frac{100}{CB(\xi)}$ at the corresponding frequency bin, where $CB(\omega)$ is the critical bandwidth, see Eq. 9.
5) Shift the result (in dB) by an level offset parameter o to get the irrelevance function $\mathcal{I}$.
$$\mathcal{I}_{l,k} = \frac{\frac{100}{CB\left(\frac{k\cdot b}{N_{FFT}}\right)} \cdot \left(|Gab(x)_{l,.}|^2 \overset{(b)}{*} B_0(-.)\right)_k}{10^{o/10}}$$
6) Use $\mathcal{I}$ as a threshold function to get the symbol for a Gabor filter.
$$m_{l,k} = \begin{cases} 1 & \text{if } Gab(x)_{l,k} \geq \mathcal{I}_{l,k} \\ 0 & \text{otherwise} \end{cases}.$$
Apply the Gabor filter on the signal.

More details on each step are provided in the following.

1) Gabor Transform: The current algorithm uses a *Hamming* window of length N_{win}, i.e.

$$g_k = \begin{cases} 0.54 - 0.46\cos\left(2\pi\frac{k}{N_{win}-1}\right) & 0 \leq k \leq N_{win}-1 \\ 0 & otherwise \end{cases} \tag{8}$$

A Gabor analysis is performed with the time shift $a = \frac{N_{win}}{8}$ and $N_{FFT} \geq N_{win}$ frequency bins. This corresponds to the *"painless non-orthogonal expansion"* [28] and ensures that the Gabor transform can be computed efficiently. For the

[4] *Involution* means mirroring around 0, i.e. flipping: $\tilde{f}(x) := f(-x)$.

continuous case it can be shown that the Hamming window always forms a tight Gabor frame for any hop size $H = \frac{N_{win}}{2^\eta}$ for any $\eta = 1, 2, \ldots$. This is not true for the sampled version of this window as in Eq. 8 , because the typically chosen boundary conditions, i.e. that the window has the same value at 0 and $N_{win} - 1$, does not support this in the finite dimensional case. But this half-point symmetrical window is typically used in applications. Nevertheless it would allow a more intuitive interpretation of the time-frequency coefficients, if the same window is used for analysis and synthesis. Also, to avoid the calculation of a dual window, we show that the discrete Gabor system forms a snug frame in the finite, discrete case. With a redundancy of 8 using Theorem 2 it can be shown that the Gabor matrix S is a diagonal matrix with a (nearly) constant diagonal (for the parameters used in the experiment, see Section III-D1 up to a relative error of $\epsilon = 4.0090 \cdot 10^{-5}$). This is an acceptable value, so the perfect reconstruction can be obtained up to a very small error if the same Hamming window is also applied as synthesis window, with an appropriate scaling ($1/810.7694$ for the parameters used in Section III-D1). So this window forms a snug frame.

2) Hz-to-Bark Transform: The transformation of the signal spectrum from the linear into the Bark frequency domain, $Hz \rightarrow Bark$, is performed according to a point-wise relation. For that purpose a fixed grid in the Bark scale is defined, with N_{Bark} bins. For every FFT-bin the nearest Bark bin corresponding to its Hz value is chosen and set to this value. Components not corresponding to FFT-bins are set to zero. This means that the number and values of the non-zero bins in the Bark scale correspond to the number and values of the associated FFT-bins. The number of bins in the Bark domain is chosen high enough, such that the resolution is always better than on the Hz scale, i.e. this transformation is one-to-one.

This point-wise relation can be seen as corresponding to a sinusoidal synthesis model and was chosen to be comparable to [57], which motivated the choice of the values for u and l. This choice for the transformation can also be found, for example, in the explanation of masking effects in [4].

3) Spreading by Convolution: The convolution can be implemented very efficiently using the FFT in an overlap-add (OLA) approach [33], applying zero-padding to avoid the aliasing effect due to circular convolution. A convolution model of auditory masking assumes linear additivity of masking in the power domain, i.e. linear summation of energy. Although power-law additivity may be more appropriate in certain signal and masker configurations [13], [12], [44], linear additivity gives a conservative estimate of the masking effect. Using convolution implies that all components, even those which may fall below the absolute or masked threshold function, are taken into account. This step is based on the assumption that even sub-threshold components may contribute to the combined masking effect. As seen in Equation 7 the spreading threshold kernel $B_0(\omega)$ is zero (on the linear scale) or negative infinity (on the logarithmic scale) within the interval $(0 - \epsilon/2, 0 + \epsilon/2)$. This reflects the assumption that a given frequency component cannot influence the irrelevance threshold at the same frequency position. In the algorithm $\epsilon > 0$ is set to the resolution of the discrete frequency analysis (FFT).

4) Weighting by Relative Critical Bandwidth: It can be shown, see Appendix A, that the spreading function calculated by convolution and using appropriate weightings is equivalent to the excitation pattern calculated according to the method described in [6].

The excitation pattern of a signal with constant amplitude and constant spectral density (e.g. a harmonic complex with equal amplitudes) grows with increasing frequency [5] due to the broadening of the auditory filter. This effect is modeled by the assumption of the shift-invariance of the spreading threshold kernel $B_0(\omega)$ in the Bark domain, as this implies a broadening of this kernel in the Hz domain. Furthermore, there is an effect caused by the change in the spectral density as a function of frequency. By the one-to-one approach for the Hz to Bark transformation, the spectral density is increasing with frequency in the Bark domain. As the resulting function is used as a threshold function the rising tilt has to be avoided, because it would result in overly masking of higher frequency components. Therefore, the threshold function is weighted by the relative critical bandwidth $100/CB$. This corresponds to using the same weighting function only in the backward direction, see Appendix A. The resulting function is referred to as *masked threshold function*. The formula for the critical bandwidth [41] is given by:

$$CB(\omega) = 25 + 75 \cdot \left(1 + 1.4 \cdot 10^{-6} \cdot \omega^2\right)^{0.69}. \qquad (9)$$

The masked threshold function is transformed to the Hz scale, again on a one-to-one basis.

5) Shifting by Offset: Finally, the masked threshold function is shifted in dB level according to an offset parameter o. The appropriate choice of the offset parameter ensures that any uncontrolled effects of signal processing and properties of masking not accounted for by the described masking model are coped with. The determination of the offset value is based on a conservative criterion derived from the perceptual tests described below, including a variety of real-world test stimuli. Finally, the shifted threshold function is called the *irrelevance threshold*.

6) Gabor Filter by Thresholding: The simultaneous masking algorithm is implemented as an adaptive filter. The irrelevance threshold function is calculated for each consecutive spectrum of a running signal. Only the components exceeding the threshold are included in the re-synthesis stage. This step is equivalent to multiplying each time-frequency point by 0 or 1. Fig. 4 shows the perceptually relevant TF components.

This procedure is an example of a Gabor filter with a symbol consisting of zeros and ones. First the *irrelevance threshold* is determined based on the signal, which is clearly an adaptive and therefore non-linear process. The filtering stage itself, a time-variant filter, is a linear process again. Introducing this model, the underspread property (see Section II-B2) is important, since the induced time-frequency shift should be as 'local' as possible. The approximation process, in which only single time-frequency points are removed from the signal, was performed as accurately as possible. The goal was to obtain an operator with good time-frequency localization, i.e. an underspread operator [58]. To achieve that goal and following

Gabor theory, a high redundancy has been chosen, $red = 8$. At high redundancy, short on/off cycles of single components that are close to the irrelevance threshold are smoothed out, which is desirable from a psychoacoustical point of view as sharp on/off edges cause audible artifacts.

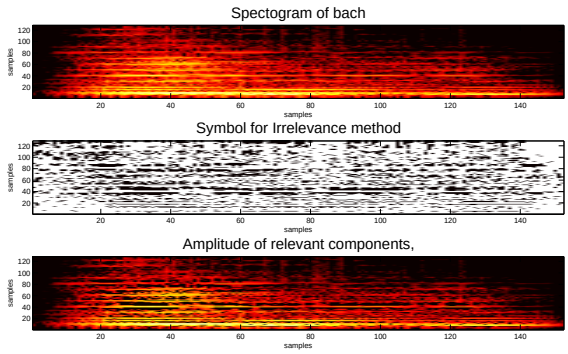

Fig. 4. TOP: The spectrogram of test signal 'bach' (classical music by J. S. Bach), high amplitude is displayed brightly, low darkly; MIDDLE: The symbol of the Gabor filter for the irrelevance filter, white = 1, black = 0. BOTTOM: The result of the point-wise multiplication of these two sets of coefficients, representing the amplitude of relevant components.

C. Numerical Complexity

Using a simple linear model, a convolution and a simple thresholding approach leads to a fast algorithm: For a signal of length n with hop size a we get $\frac{n}{a}$ spectra in the Gabor transform. For each of these spectra we have the following calculations:

1) Gabor transform: $\mathcal{O}(N_{FFT} \cdot log(N_{FFT}))$.
2) Hz-to-Bark transform: $\mathcal{O}(N_{Bark})$.
3) Spreading by convolution: $\mathcal{O}(N_{Bark} \cdot log(N_{Bark}))$.
4) Weighting by Relative Critical Bandwidth: $\mathcal{O}(N_{Bark})$
5) Bark-to-Hz transform: $\mathcal{O}(N_{Bark})$.
6) Shifting by Offset: $\mathcal{O}(N_{FFT})$.
7) Thresholding: $\mathcal{O}(N_{FFT})$.
8) inverse Gabor transform: $\mathcal{O}(N_{FFT} \cdot log(N_{FFT}))$.

As $N_{Bark} \geq N_{FFT}$ for the whole signal we have an estimation of the number of operations by

$$\mathcal{O}(\frac{n}{a} \cdot N_{Bark} \cdot log(N_{Bark})).$$

D. Experimental Evaluation of the Proposed Algorithm

The algorithm was evaluated in a listening experiment [20]. The aim was to find the value of the free parameter o, determining the level offset of the threshold function, for which normal hearing listeners cannot detect any difference between the processed signal and the original for a broad range of signals. The higher the level of the threshold function, the more spectral components fall below the threshold and are filtered out. For o = 0, the threshold function is not shifted. Positive and negative values of o correspond to upward and downward shifts in level, respectively. A downward shift of the threshold function allows to account for potential overestimation of masking effects, either due to inaccuracies in the masking model, e.g. due to spectral integration in signal detection, or due to unpredictable effects associated with the removal of components.

1) Method: Thirty-six normal hearing subjects completed the experiment. The majority of them were students of the University of Vienna. The test stimuli were derived from 25 music recordings, covering a wide variety of musical styles and musical instruments, and one speech recording obtained from a female news speaker. From each of the 26 sounds, segments with three different durations (300, 600, and 1200 ms) were extracted. Linear ramps with a duration of 90 ms were applied. The segment borders were determined pseudo-randomly within a preselected range. Segments for which the random process led to truncation of musical phrases were discarded. [5] The sound level of the stimuli was set to yield a comfortable loudness. The stimuli were stored on computer hard disc and output via a DAC converter (Siemens, ADC 16/12-15), an amplifier (Kenwood KA-7100), and a circum-aural headphone (AKG K 240 DF). Only the right channel of the recordings was presented to the subjects. The subjects were seated in a double-walled sound booth (IAC 1202A). The sampling frequency of 16 kHz and a digital word length of 16 bit was used.

Based on the results of pilot tests, four processing conditions were selected for the main tests (Table I): Condition 1 represented the original (unprocessed) signal. Conditions 2, 3, and 4 corresponded to the values of the parameter o (the level offset) -6.59, -4.59 and -2.59 dB, respectively. Condition 1 corresponded to $o = -\infty$.

Condition	Offset o
1	$-\infty$ dB
2	-6.59 dB
3	-4.59 dB
4	-2.59 dB

TABLE I
THE FOUR CONDITIONS TESTED IN THE PERCEPTUAL EXPERIMENT. THE dB VALUES SPECIFY THE LEVEL OFFSET PARAMETER o.

The offset parameter values were chosen to encompass the transition from chance rating to significant discrimination. All other parameters of the algorithm were held constant:

sampling rate	16 kHz
window length N_{win}	256 samples
FFT length N_{FFT}	256 samples
hop size a	32 samples
lower slope of the spreading function l	27 dB / Bark
upper slope of the spreading function u	24 dB / Bark
damping factor e	0.3
length of Bark scale N_{Bark}	512 samples

Each of the 26 sounds was presented once at all combinations of three durations and four processing conditions, resulting in a total number of 312 test stimuli.

A double-paired comparison task was used to obtain percent scores on the discriminability between original and processed

[5] This was intended to introduce a kind of controlled randomness into the selection process.

stimuli. This task represents a four-interval, two-alternative forced-choice procedure. One pair contained two identical stimuli (the original signal) and the other pair contained the processed signal and the original. The temporal position of the processed condition within the four possible signal intervals was randomized. The subjects had to indicate which pair contained different stimuli. The subjects were allowed to repeat the stimulus five times at maximum before giving a response. The stimulus intervals were indicated visually on a computer screen. Visual feedback on the correctness of the response was provided after each trial by indicating if the correct pair was chosen. The inter-stimulus interval of each pair was 0.5 s and between the two pairs 1 s. The order of stimulus conditions was randomized and the same order was used for all subjects. The stimuli were split into two blocks, each lasting about 40 minutes. Before the start of data collection, a practice period of maximally 25 items was completed.

2) Results: The mean percent correct scores for the four processing conditions at each of the three durations are shown in Fig. 5. The error bars show 95% confidence intervals around the mean scores across the 36 subjects. A two-way repeated-measures analysis of variance (RM ANOVA) (factors: processing condition, duration) was performed. The percent correct scores were transformed using the rationalized arcsine transform proposed in [59] to not violate the homogeneity of variance assumption required for an ANOVA. The RM ANOVA showed that the main effects were significant (processing condition: $p < 0.0001$; duration: $p = 0.002$) as well as their interaction ($p = 0.036$). Tukey HSD post-hoc tests revealed differences to be significant between all combinations of processing conditions except between conditions 1 and 2. The main effect of the factor duration (and its interaction with the factor processing condition) was found to be caused by significant differences between durations 0.3 vs. 0.6 s for processing condition 3 and by significant differences between durations 0.3 vs. 1.2 s and 0.6 vs. 1.2 s for processing condition 4. Thus, there were significant improvements with increasing duration, but only for conditions 3 and 4.

The obtained percent correct scores are binomially distributed. To obtain a statistical measure, if a particular score represents sensitivity of the listener to discriminate the original from the processed sound (H1) or falls into the range of chance performance (H0), a test based on the binomial distribution is required. For each subject and processing condition there were $N = 3 \cdot 26 = 78$ trials. Since $N > 60$, in which case the binomial distribution can be approximated by the normal distribution, the *u-test* can be used to determine the probability that a given score is obtained by chance. In the specific case, we calculated the minimum score which has to be obtained to exceed chance performance at the given N of 78. Scores exceeding 64.1 and 60 percent correct indicate discrimination performance above chance at alpha levels of 0.01 and 0.05, respectively. Table II depicts the percentage of subjects for which this was fulfilled for each of the four processing conditions. In case of condition 1, no subject exceeded the range of chance performance and for condition 2 one subject just reached the 5 percent significance level. However, obtaining one significant result in 36 test repetitions with an alpha level of 5 percent is likely to occur by chance. In case of condition 3, 36.1 percent of the subjects reached the 5 percent and 19.4 percent reached the 1 percent significance level. In case of condition 4, 75 and 61.1 percent of the subjects obtained significant scores at the two significance levels, respectively.

To obtain a statistical measure of sensitivity for the sample of subjects as a whole, the mean percent correct score across all subjects was analyzed for each test condition. For $N = 36 \cdot 3 \cdot 26 = 2808$ trials, the *u-test* reveals that scores exceeding 52.5 and 51.9 percent correct indicate discrimination performance above chance level (alpha levels of 0.01 and 0.05, respectively). As can be seen in Table II the performance exceeds these critical values for condition 3 and 4, but not for conditions 1 and 2. The aim of the experiment was to find the highest value of o for which the listeners could not discriminate the processed from the original signal. Hence, condition 2 is considered as the irrelevance threshold. Please note that the term irrelevance threshold refers to a signal processing function and not to a psychophysical threshold.

For condition 2, 35.8 percent of the Gabor coefficients, on average across all stimuli and windows, fell below the irrelevance threshold and hence were set to zero. The standard deviation of the percentages across the stimuli was 8.5 percent. No absolute hearing threshold criterion was applied for the calculation of these percentages. This means that it is possible that a portion of the discarded coefficients fell below the absolute threshold of hearing. Only a small percentage of the signal energy has been removed (depending on the signal between 0.2 and 1.2 %, statistically 0.47 ± 0.26 %). Please note that not only the components with the lowest amplitudes were removed.

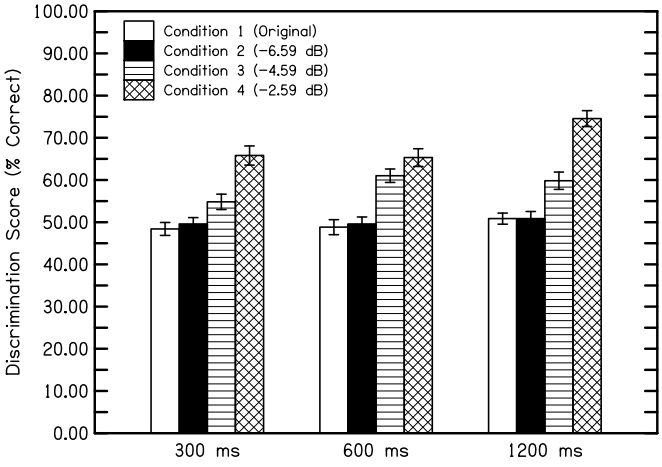

Fig. 5. Percent correct discrimination scores obtained from the perceptual experiment for the four signal processing conditions and three signal durations. The error bars show 95% confidence intervals around the mean values across the 36 listeners.

	Processing Condition			
	1	2	3	4
Mean % Correct Score	49.4	50.5	58.6	68.6
Percentage of subjects exceeding chance performance at $p < 0.05$	0.0	2.8	36.1	75
Percentage of subjects exceeding chance performance at $p < 0.01$	0.0	0.0	19.4	61.1

TABLE II
PERFORMANCE MEASURES OBTAINED FROM THE PERCEPTUAL EXPERIMENT. THE MEAN PERCENT CORRECT DISCRIMINATION SCORES ARE AVERAGED OVER THE THREE SIGNAL DURATIONS.

E. Masking Pattern Simulation

In this section we use a simple simulation approach to examine the appropriateness of the masking model to mimic the basic simultaneous masking effect as revealed by the simultaneous masking pattern. The simultaneous masking pattern provides a measure of the spread of masking caused by a narrow-band masker and is obtained by measuring masked thresholds of narrow-band targets placed at different frequencies around the masker. Both sinusoids and narrow-band-noises have been used as maskers and targets in the psychoacoustic literature and it has been shown that the shape of the masking pattern depends on the specific stimulus combination (e.g., [60]). For our simulation approach, both the masker and the targets were sinusoids. We did not test noise bands since the model is not designed to take into account effects of temporal fluctuations. 300-ms masker and target stimuli were fed into the algorithm and the resynthesized signal was inspected with a discrete Fourier transform. The masker frequency was 1000 Hz and 10 target frequencies surrounding the masker were chosen, ranging from 250 to 4000 Hz. In order to simulate the masked threshold at a given target frequency, the target level was systematically varied in steps of 2 dB in a level region around the expected masked threshold. For target levels above the irrelevance threshold, a given decrease in target level results in the same level decrease of the target in the resynthesized signal. As soon as the target level falls below the irrelevance threshold, however, the target is not resynthesized any more. Thus, by tracking the amplitude in the FFT bins surrounding the target frequency, we can determine the level of the target where it passes the irrelevance threshold, and this level represents the simulated masked threshold. Due to the properties of the analysis-resynthesis system, there is, however, a transition region where the target level is increasingly dampened until it disappears. Therefore, the masked threshold was defined as the input target level at which the resynthesized target level was at least 10 dB dampened relative to the input target level. This procedure was performed for each target frequency. Note that we did not simulate the condition where the target frequency is equivalent to the masker frequency. The reason is that there is no straight-forward way to simulate the masked threshold for this condition.

Figure 6 compares the results of the simulated masking pattern with psychoacoustically measured masking patterns reported in the literature, using the same stimulus type and masker frequency and a comparable masker level. Data from the study of [60] are shown for a masker level of 65 dB SPL and from the study of [61] for a masker level of 60 dB SPL. Note that [60] used a three-interval forced-choice task whereas [61] used a Békésy tracking procedure. The results of those two studies coincide well for the lower edge, including the peak of the pattern. However, the upper edge of the masking pattern from [61] is much flatter. The simulated masking pattern was shifted in level to coincide with the two psychoacoustically measured masking patterns at the lower edge, thus where the data from those two studies themselves coincide. There is good agreement of the simulated masking patterns with the data from [60]. At very low levels, however, the simulation does not show the flattening of the pattern that appears in the data from [60]. For example, at target frequencies of 500 and 1500 Hz the simulated masked thresholds fall below zero sensation level (not shown) whereas the data from [60] still yield 3 and 12 dB of masking, respectively. Nevertheless, the main part of the simulated masking pattern agrees well with the masking pattern measured psychoacoustically using a forced-choice task.

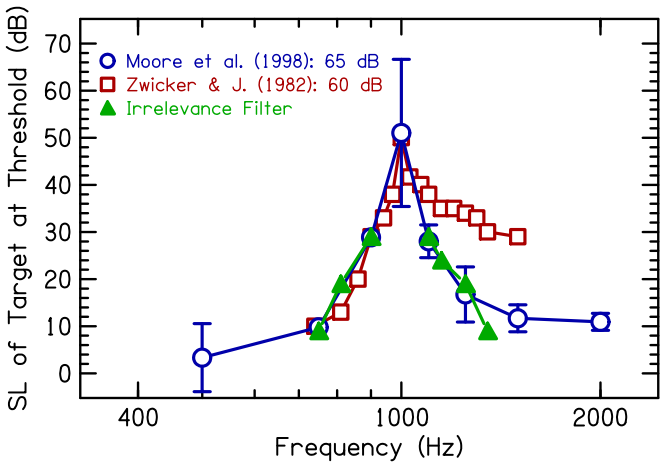

Fig. 6. Comparison between simulated masking pattern (filled triangles) and psychoacoustically measured masking patterns redrawn from [60] (circles) and [61] (squares). The sensation level (SL) of the target at threshold is shown as a function of frequency. The data from [60] are mean values of three listeners and include $\pm$ 1 standard deviation of the mean; the data from [61] are mean values of eight listeners.

IV. Discussion

The results of the perceptual experiment showed that the subjects could not discriminate the irrelevance filtered sound from the original sound for a value of the level offset parameter o of -6.59 dB. The transition from chance rating to significant discrimination performance falls between the processing conditions with offset values of -6.59 and -4.59 dB. These results suggest that the irrelevance threshold for real-world signals can be obtained by convolution of the Bark-transformed signal spectrum with a spreading function of simultaneous masking and a downshift by about 6.6 dB.

We observed improvements in discrimination scores for increasing signal durations (ranging from 0.3 to 1.2 s) for those processing conditions (3 and 4) that were discriminable from the unprocessed sounds. The results indicate that using signal durations larger than 0.3 s increases the probability to discriminate the processed from the unprocessed sound. This result is consistent with the idea that at longer signal durations the auditory system has the advantage of observing more time instances containing potential differences to be detected.

There is a general limitation of the approach to shift the masked threshold in level until no difference can be heard between the original and the processed signal. It does not allow to evaluate the contribution of different signal components to the perceptual degradation. For example, one frequency region could have contributed more to the perceptual degradation than other regions. This issue should be addressed in future advancements of the algorithm. One possibility would be to test the effect of the level shift of the masked threshold function systematically in different frequency regions. Another possibility would be to introduce an iterative approach, where a model of auditory processing is applied after signal re-synthesis and the auditory representation of that signal is compared to that of the original signal. This would allow to correct wrong decisions, i.e. the removal of either too many or of too few components. Such an approach would be, however, computationally expensive. In any case, the current algorithm should be considered as a starting point that hopefully motivates further advancements of the general approach.

We examined the appropriateness of the proposed irrelevance filter algorithm to mimic the basic simultaneous masking effect by simulating masking patterns and comparing them with data from two psychoacoustic studies from the literature ([60], [61]). But first, it is important to discuss the finding that the data from these two studies agree at the lower edge, including the peak of the pattern, but show a much flatter slope at the upper edge in [61]. This could be due to the fact that [61] used a Békésy tracking method whereas [60] used a three-interval forced-choice task. The latter task allows the subjects to "home in" on the optimal detection cue for each masker-signal combination and thus may lead to lower thresholds. As discussed in [60], the perception of combination products and beats likely influenced the thresholds at the upper edge of the pattern. Now turning back to the main comparison, the simulated masking pattern was shown to agree well with the pattern from [60] for the main part. Thus, the masking model used in our study gives a conservative estimate of simultaneous masking effects involved in sinusoidal stimuli.

Although there exist very sophisticated models that allow one to accurately predict simultaneous masking effects (e.g. [52], [54], [2]), no study is known to the authors that followed the approach of our study, i.e. to remove time-frequency components of a real-life signal while causing no audible difference to the original signal. As has been outlined in the introduction section, removing time-frequency components from a signal involves special effects that are related to the properties of time-frequency representations and to the properties of the masking model. These effects were handled by the introduction of a level shift of the masked threshold function. Because of the specifity of our approach, it is not possible to compare our results with results from published masking models or perceptual audio codecs.

The masking model applied in our study was deliberately a simple one, which does not consider complex and nonlinear effects such as suppression [16], [17], nonlinear additivity of masking for spectrally non-overlapping maskers [44], or the level-dependence of auditory filters [6]. Furthermore, the current masking model does not consider the dependence of simultaneous masking on the temporal characteristics of the stimuli. For example, the amount of masking differs between tonal and noise maskers, depending on the fluctuation rate of the masker [62], [4], and co-modulation of masker components across frequency bands is known to cause release from masking, an effect that has been termed co-modulation masking release [63]. Another effect not directly incorporated into the current masking model is the across-frequency integration in signal detection [14], [15]. Removing more than one spectral component could result in an audible change even if each of the components separately falls below the masked threshold. The result that the irrelevance threshold was found at a negative value of the level offset parameter might indicate that across-frequency integration effects were involved. However, it could also result from other inaccuracies of the masking model or from uncontrolled effects resulting from removing time-frequency components.

Some of the complex masking effects mentioned above might have been involved in the present study. Incorporating them into the masking model might result in a higher efficiency in terms of the number of removable components and will be considered for future refinement of the algorithm. In any case, the current approach for removing perceptually irrelevant components provided a safe criterion. By level-shifting the masked threshold function, the filter criterion, so that listeners just heard no difference to the original signal, some of those effects may have indirectly been taken into account.

When interpreting the results it has to be kept in mind that the stimuli used were dynamic in their spectral and temporal characteristics. This could have made it difficult to detect subtle sound differences at certain instances of time. In case of steady-state stimuli such as harmonic complexes, the irrelevance threshold might have been found at even lower values of the offset parameter. For such stimuli, it may be easier for listeners to focus their attention on specific spectral

regions. A related aspect is that the subjects had no possibility to improve their performance over time for the specific stimuli since each stimulus was tested only once. [6] If instead a small number of stimuli had been presented repeatedly, it might have been easier for the subjects to detect slight differences since they would have been familiarized with the cues to be detected. While these issues are interesting for future studies, it is important to note that the stimuli and procedures used in this study were selected to represent realistic listening situations. We think that due to the relatively broad range of sounds used in the experiments the results should be generalizable to other real-life music and speech sounds.

The high redundancy in the analysis-synthesis system results in smoothing and thus reduces the efficiency of the algorithm. Components whose levels vary around the irrelevance threshold from one analysis interval to the next are not completely removed. To obtain a true on/off switching, the redundancy would have to be very low (near 1), but as a consequence the loss of localization in the time-frequency plane would have to be accepted. Therefore, the smoothing in the resynthesis process appears inherent to time-variant systems.

The main application of the described irrelevance filter lies in removing perceptually irrelevant components from real-world sounds in order to obtain more sparse and simple frequency representations of perceptual relevance, to facilitate the sound synthesis and design. Furthermore, it may ease the interpretation of time-frequency properties of signal used in perception-related tasks.

V. Perspectives

Several parts of the current algorithm could be improved. For example, the high redundancy of the Gabor transform could be reduced by using the canonical dual Gabor window for resynthesis instead of relying on snugness. With the theory of Gabor multipliers, i.e. Gabor filters, the window and the parameters could be chosen such that the smoothing and underspread property of the filter are kept for lower redundancies. Furthermore sparsity is currently a prominent topic in signal processing, e.g. under the designation “compressed sensing” [64], [65], and we will look at a way to combine our approach with that one.

There is large room for refinements of the simultaneous masking model currently implemented in the algorithm. A more accurate model of peripheral auditory processing may increase the efficiency of the algorithm in terms of the amount of removable components. Possible improvements based on known and well-studied properties of auditory processing include the inclusion of outer/middle ear transfer function [6], the inclusion of the absolute hearing threshold, the level-dependence of auditory filters [6], the nonlinear additivity of masking depending on the spectral relations of the masker components [44], spectral integration effects in signal detection ([14], [15], [11]), or the dependency of the amount of masking on the degree of tonality of the masker as well as of the target [4], [62].

In the context of *Gabor filters*, ways to combine simultaneous masking and temporal masking will be explored to extend the current algorithm to a true *time-frequency masking* algorithm. In [66], a basic model for a simple time-frequency masking algorithm based on the algorithm presented here, has been proposed. Work on an extension of the algorithm, the evaluation of its applicability, as well as on basic psychoacoustic experiments on time-frequency masking using Gaussian-shaped tones is currently underway [45].

[6] disregarding the optional stimulus repetition, that provided no response feedback.

Acknowledgments

The authors would like to thank Matthew Goupell, Florent Jaillet, Monika Dörfler, Solvi Ystad and the anonymous reviewers for helpful comments, Toni Noll for a lot of help with implementations, Wolfgang Kreuzer for help with LaTeX and proofreading.

The first author would like to thank the hospitality of the LATP, CMI and the LMA, CNRS, both Marseille, France, where part of this work was prepared.

Appendix

A. Connection to the Excitation Pattern

The *excitation pattern model* [5], [6] uses the concept of the auditory filter bank to calculate a model for the BM activation, the *excitation pattern*. We will show that our spreading function is equivalent to the excitation pattern. We denote by $EP_{(f)}(\omega)$ the excitation pattern at frequency ω of a signal f with Fourier transform $\hat{f}$. Let $AF(\eta,\xi)$ be the auditory filter as a function of the frequency ξ, with center frequency η in the power spectrum with maximum 1. The excitation pattern can be regarded as the response of the auditory filter bank using power spectra, i.e

$$EP_{(f)}(\xi) = \left\langle \hat{f}^2(\cdot), AF(\xi,\cdot) \right\rangle = \int_{\mathbb{R}} AF(\xi,\nu)\hat{f}^2(\nu)d\nu. \quad (10)$$

For further motivation consider a signal, which consists of a single complex sinusoid at frequency ν_0 (i.e. $\hat{f}(\omega) = \delta_{\nu_0}(\omega)$). Then

$$EP_{(\check{\delta_{\nu_0}})}(\omega) = AF(\omega,\nu_0). \quad (11)$$

Compare this to the description in [5] p. 752.

In Equation 10 we can choose another frequency scale, in our case the Bark scale (Equation 2). We want to represent the formula in the Bark scale, using the notation $f^{(b)}(\omega) = f\left(b^{-1}(\omega)\right)$. Then let

$$\begin{aligned} EP^{(b)}_{(f)}(\omega) &= EP_{(f)}(b^{-1}(\omega)) = \\ &= \int_{\mathbb{R}} AF(b^{-1}(\omega),\nu)\hat{f}^2(\nu)d\nu. \end{aligned} \quad (12)$$

Substituting $\nu = b^{-1}(\zeta)$ and using integration by substitution we get

$$EP^{(b)}_{(f)}(\omega) = \int_{\mathbb{R}} AF(b^{-1}(\omega), b^{-1}(\zeta))\hat{f}^2(b^{-1}(\zeta))\frac{d\ b^{-1}(\zeta)}{d\zeta}d\zeta$$

where we denote the first derivative of a function $f(t)$ by $f'(t) = \frac{df}{dt}(t)$.

The derivative of the function b^{-1}, i.e. the transformation from the Bark to the Hz scale, can be approximated well by the critical bandwidth.[7] This is motivated in the following way: Let $CB(\omega)$ denote the *critical bandwidth* at (linear) frequency ω, defined as the distance of the two linear frequencies corresponding to $b(\omega) - \frac{1}{2}$ and $b(\omega) + \frac{1}{2}$:

$$CB(\omega) = b^{-1}\left(b(\omega) + \frac{1}{2}\right) - b^{-1}\left(b(\omega) - \frac{1}{2}\right).$$

Using the mean value theorem this is equivalent to

$$CB(\omega) = \left(b^{-1}\right)'(\zeta_0) \cdot \left(\frac{1}{2} + \frac{1}{2}\right)$$

for a $\zeta_0 \in \left(b(\omega) - \frac{1}{2}, b(\omega) + \frac{1}{2}\right)$. Applying Taylor's theorem to expand b^{-1} around $b(\omega)$ leads to the following approximation

$$CB(\omega) \cong \left(b^{-1}\right)'(b(\omega)).$$

Therefore

$$EP^{(b)}_{(f)}(\omega) = \int_{\mathbb{R}} AF(b^{-1}(\omega), b^{-1}(\zeta))\hat{f}^2(b^{-1}(\zeta)) \cdot CB\left(b^{-1}(\zeta)\right) d\zeta. \tag{13}$$

For $CB(b^{-1}(\zeta))$ a good approximation by an analytical formula is known, see Equation 9.

In the Bark scale all auditory filters can be approximated by triangular functions having equal slopes at different center frequencies. This means they are just shifted versions of a shape function, denoted by $\mathcal{AF}(\omega)$. Using the above mentioned translation operator T_τ this can be expressed as

$$AF(b^{-1}(\eta), b^{-1}(\zeta)) = T_\eta \mathcal{AF}(b^{-1}(\zeta)).$$

Therefore the original formula can be simplified to:

$$Eq.13 = \int_{\mathbb{R}} \left[T_\omega \mathcal{AF}\left(b^{-1}(\zeta)\right)\right] \hat{f}^2(b^{-1}(\zeta)) \cdot CB(b^{-1}(\zeta))\, d\zeta =$$
$$= \int_{\mathbb{R}} \left[\mathcal{AF}\left(b^{-1}(\zeta - \omega)\right)\right] \cdot \left(\hat{f}^2(b^{-1}(\zeta)) \cdot CB(b^{-1}(\zeta))\right) d\zeta.$$

Denote the involution by $\widetilde{\mathcal{AF}}(\omega) = \mathcal{AF}(-\omega)$ and the convolution in the Bark scale by

$$\left(f \overset{(b)}{*} g\right)(\omega) := \int_{-\infty}^{\infty} f\left(b^{-1}(\omega - \nu)\right) \cdot g\left(b^{-1}(\nu)\right) d\nu.$$

Then [8]

$$EP^{(b)}_{(f)}(\omega) = \left[\widetilde{\mathcal{AF}} \overset{(b)}{*} \left(\hat{f}^2 \cdot CB\right)\right](\omega).$$

For the transformation back to the Hz scale we set

$$EP_{(f)}(\xi) = \frac{1}{CB(\xi)} \left[\widetilde{\mathcal{AF}} \overset{(b)}{*} \left(\hat{f}^2 \cdot CB\right)\right](b(\xi)).$$

[7]For the very similar ERB scale [5] it can be easily shown that this relation is exactly true for the analytical approximation of the ERB scale and the ERB bandwidth [67].

[8]Notice that b^{-1} is symmetric.

The weighting by $\frac{1}{CB(\xi)}$ is chosen to keep Eq. 11 valid, because

$$\int_{\mathbb{R}} AF(b^{-1}(\omega), b^{-1}(\zeta))\delta_{b^{-1}(\zeta_0)}(b^{-1}(\zeta)) \cdot CB\left(b^{-1}(\zeta)\right) d\zeta =$$
$$= AF(b^{-1}(\omega), b^{-1}(\zeta_0)) \cdot CB\left(b^{-1}(\zeta_0)\right).$$

In conclusion, the calculation of the excitation pattern using the auditory filter model is equivalent to a convolution model in the Bark scale, using the critical bandwidth function as a weighting factor twice, the original CB in the forward and its inverse in the backward frequency scale transformation.

B. Download

The masking algorithm is implemented in ST^X [21], a signal processing software system designed at the Acoustics Research Institute of the Austrian Academy of Sciences. The software can be downloaded at `http://www.kfs.oeaw.ac.at` and a free trial license can be obtained (for 3 months) by e-mail as explained on this webpage.

REFERENCES

[1] K. Brandenburg, "MP3 and AAC explained," in *Audio Engineering Society (AES) 17th International Conference on High QualityAudio Coding, Florence, Italy*, September 1999.

[2] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[3] B. Moore, *An Introduction to the Psychology of Hearing*. Academic Press Limited, London, 1989.

[4] E. Zwicker and H. Fastl, *Psychoacoustics*. Springer-Verlag, Berlin, 1990.

[5] B. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitationpatterns," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, Sept. 1983.

[6] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, Aug. 1990.

[7] M. van der Heijden and A. Kohlrausch, "Using an excitation-pattern model to predict auditory masking," *Hear Res.*, no. 1, pp. 38–52, Oct 1994.

[8] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.*, vol. 66, no. 6, pp. 1647–1652, 1979.

[9] ISO/MPEG Comittee, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s - part 3: Audio," *ISO/IEC*, pp. 11 172–3, 1993.

[10] J. Johnston, "Estimation of perceptual entropy using noise masking criteria," *Proc. ICASSP-88*, pp. 2524–2527, 1988.

[11] S. den de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2005.

[12] B. Moore, "Additivity of simultaneous masking, revisited," *J. Acoust. Soc. Am.*, vol. 78, no. 2, pp. 488–494, Aug. 1985.

[13] R. Lutfi, "Additivity of simultaneous masking," *J. Acoust. Soc. Am.*, vol. 73, no. 1, pp. 262–267, Jan. 1983.

[14] W. A. C. van den Brink and T. Houtgast, "Spectro-temporal integration in signal detection," *J. Acoust. Soc. Am.*, vol. 88, pp. 1703–1711, 1990.

[15] A. Langhans and A. Kohlrausch, "Spectral integration of broadband signals in diotic and dichotic maskingexperiments," *J. Acoust. Soc. Am.*, vol. 91, pp. 317–326, 1992.

[16] B. Delgutte, "Physiological mechanisms of psychophysical masking: observations from auditory-nervefibers," *J. Acoust. Soc. Am.*, vol. 87, pp. 791–809, 1984.

[17] A. J. Oxenham and C. J. Plack, "Suppression and the upward spread of masking," *J. Acoust. Soc. Am.*, vol. 104, pp. 3500–3510, 1998.

[18] H. G. Feichtinger and K. Nowak, *A first survey of Gabor multipliers*. in H. G. Feichtinger, T. Strohmer (Eds.), Advances in Gabor Analysis, 2003, ch. 5, pp. 99–128.
[19] G. Matz and F. Hlawatsch, *Linear Time-Frequency Filters: On-line Algorithms and Applications*. Boca Raton (FL): CRC Press, 2002, ch. 6 in 'Application in Time-Frequency Signal Processing', pp. 205–271.
[20] G. Eckel, "Ein Modell der Mehrfachverdeckung für die Analyse musikalischerSchallsignale," Ph.D. dissertation, University of Vienna, 1989.
[21] A. Noll, J. White, P. Balazs, and W. A. Deutsch, *STX - Intelligent Sound Processing, Programmer's Reference*, Acoustics Research Institute, Austrian Academy of Science, 2001.
[22] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, vol. 72, pp. 341–366, 1952.
[23] O. Christensen, *An Introduction To Frames And Riesz Bases*. Birkhäuser, 2003.
[24] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, London, 1998.
[25] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Processing*, vol. 46, no. 12, pp. 3256–3268, 1998.
[26] K. Gröchenig, "Acceleration of the frame algorithm," *IEEE Trans. SSP*, vol. 41/12, pp. 3331–3340, 1993.
[27] J. S. Walker, *Fast Fourier Transforms*. CRC Press, 1991.
[28] K. Gröchenig, *Foundations of Time-Frequency Analysis*. Birkhäuser, Boston, 2001.
[29] I. Daubechies, *Ten Lectures On Wavelets*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM Philadelphia, 1992.
[30] W. Mecklenbräucker and F. Hlawatsch, *The Wigner Distribution - Theory and Applications in Signal Processin*. Elsevier, Amsterdam, 1997.
[31] T. Strohmer, *Numerical Algorithms for Discrete Gabor Expansions*. Birkhäuser, Boston, 1998, ch. 8, pp. 267–294.
[32] P. Balazs, H. G. Feichtinger, M. Hampejs, and G. Kracher, "Double preconditioning for Gabor frames," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, December 2006.
[33] A. V. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Oldenbourg, 1999.
[34] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.
[35] P. Balazs, "Basic definition and properties of Bessel multipliers," *Journal of Mathematical Analysis and Applications*, vol. 325, no. 1, pp. 571–585, January 2007.
[36] D. F. Walnut, "Continuity properties of the Gabor frame operator," *J. Math. Anal. Appl.*, vol. 165, no. 2, pp. 479–504, 1992.
[37] D. Gabor, "Theory of communications," *J. IEE*, vol. III, no. 93, pp. 429–457, 1946.
[38] J. O. Pickles, *An introduction to the physiology of Hearing*. Academic Press, London, 2003.
[39] D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *J. Acoust. Soc. Am.*, vol. 87, no. 6, pp. 2592–2605, June. 1990.
[40] J. O. Smith and J. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. Sig. Proc.*, vol. 7, no. 9, pp. 697 – 708, Nov 1999.
[41] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth asa function of frequency," *J. Acoust. Soc. Am.*, vol. 68, no. 5, pp. 1523–1525, Nov. 1980.
[42] H. Fastl, "Temporal masking effects: I. broad band noise masker," *Acustica*, vol. 35, pp. 287–302, 1976.
[43] R. D. Patterson and I. N. Smith, "Off-frequency listening and auditory-filter asymmetry," *J. Acoust. Soc. Am.*, vol. 67, pp. 229–245, 1980.
[44] L. Humes, L. Lee, and W. Jesteadt, "Two experiments on the spectral boundary conditions for nonlinear additivity of simultaneous masking," *J. Acoust. Soc. Am.*, vol. 92, no. 5, pp. 2598–2606, Nov. 1992.
[45] B. Laback, P. Balazs, G. Toupin, T. Necciari, S. Savel, S.Meunier, S. Ystad, and R. Kronland-Martinet, "Additivity of auditory masking using gaussian-shaped tones," in *Proceedings Acoustics'08, Paris, June 29 - July 4, 2008*, July 2008.
[46] L. Humes and W. Jesteadt, "Models of the additivity of masking," *J. Acoust. Soc. Am.*, vol. 85, no. 3, pp. 1288–1294, Mar. 1989.
[47] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–65, 1940.
[48] B. J. C. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, 1997.
[49] B. R. Glasberg and B. J. C. Moore, "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 906–918, 2005.
[50] P. Rao, R. van Dinther, R. Veldhuis, and A. Kohlrausch, "A measure for predicting audibility discrimination thresholds for spectral envelope distortions in vowel sounds," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2085–2097, 2001.
[51] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. i. model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.
[52] ——, "A quantitative model of the "effective" signal processing in the auditory system. simulations and measurements," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3623–3631, 1996.
[53] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2892–905, 1997.
[54] M. Jespen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception." *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 422–438, 2008.
[55] R. Patterson and M. Allerhand, "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.*, vol. 98, no. 4, pp. 1890–1894, 1995.
[56] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.*, vol. 66, no. 6, pp. 1647–1652, 1979.
[57] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch saliance from complex tonal signals," *J. Acoust. Soc. Am.*, vol. 71, no. 3, pp. 679–688, 1982.
[58] W. Kozek, *Adaption of Weyl-Heisenberg frames to underspread environments*. Birkhäuser, Boston, 1998, ch. 10, pp. 323–352.
[59] G. A. Studebaker, "A 'rationalized' arcsine transform," *J. Speech Hear. Res.*, vol. 28, pp. 455–462, 1985.
[60] B. J. C. Moore, J. I. Alcantara, and T. Dau, "Masking patterns for sinusoidal and narrow-band noise maskers," *J. Speech Hear. Res.*, vol. 104, no. 2, pp. 1023–1038, 1998.
[61] E. Zwicker and A. Jaroszewski, "Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels," *J. Acoust. Soc. Am.*, vol. 71, no. 6, pp. 1508–1512, 1982.
[62] M. van der Heijden and A. Kohlrausch, "The role of envelope fluctuations in spectral masking," *J. Acoust. Soc. Am.*, vol. 97, no. 3, pp. 1800–1007, Mar. 1995.
[63] J. W. Hall, M. Haggard, and M. A. Fernandes, "Detection in noise by spectro-temporal pattern analysis," *J. Acoust. Soc. Am.*, vol. 76, pp. 50–56, 1984.
[64] D. L. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence ofnoise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
[65] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 2210–2219, 2008.
[66] P. Balazs, "Regular and irregular Gabor multipliers with application to psychoacoustic masking," PhD thesis, University of Vienna, June 2005.
[67] W. M. Hartmann, *Signals, Sounds, and Sensation*. Springer, New York, 1998.