A Low-Complexity Spectro-Temporal Distortion Measure for Audio Processing Applications

Cees H. Taal, Richard C. Hendriks, and Richard Heusdens

***Abstract*—Perceptual models exploiting auditory masking are frequently used in audio and speech processing applications like coding and watermarking. In most cases, these models only take into account spectral masking in short-time frames. As a consequence, undesired audible artifacts in the temporal domain may be introduced (e.g., pre-echoes). In this article we present a new low-complexity spectro-temporal distortion measure. The model facilitates the computation of analytic expressions for masking thresholds, while advanced spectro-temporal models typically need computationally demanding adaptive procedures to find an estimate of these masking thresholds. We show that the proposed method gives similar masking predictions as an advanced spectro-temporal model with only a fraction of its computational power. The proposed method is also compared with a spectral-only model by means of a listening test. From this test it can be concluded that for non-stationary frames the spectral model underestimates the audibility of introduced errors and therefore overestimates the masking curve. As a consequence, the system of interest incorrectly assumes that errors are masked in a particular frame, which leads to audible artifacts. This is not the case with the proposed method which correctly detects the errors made in the temporal structure of the signal.**

***Index Terms*—Audio coding, auditory modeling, perceptual model.**

I. Introduction

It is well-known that the properties of the human auditory system play an important role in the development of various audio and speech processing algorithms. One such example is transparent audio coding where, by reducing the bit-rate, errors are introduced to a signal such that the distorted signal is perceptually indistinguishable from the original [1]. Here, a typical approach is to shape the quantization error in the frequency domain, on a frame-by-frame basis, according to the so-called masking threshold per auditory band. As long as the error signal is below this threshold, the original signal will act as a masker on the error signal. This phenomenon, called auditory masking, is also exploited in the field of watermarking [2], where some type of information is embedded (the watermark) by means of adding noise in such a way that it is masked by the clean signal.

In order to determine whether an introduced error is audible, the system under test typically uses a perceptual model. A well-known perceptual model is the ISO/IEC 11172-3 (MPEG-1, layer I) psychoacoustic model 1 [3]. This perceptual model is typically used in the field of audio coding [1], [4], but is also applied in the field of other audio and speech processing applications like speech enhancement [5] and watermarking [2]. Here, the masking threshold per frequency band is found by first separating the signal in tonal and noise maskers, after which for each of these spectral components a spreading function is defined [1]. Then, by power addition of these spreading functions, a masking threshold is obtained. This method is based on the assumption that the detectability of a specific frequency component is only determined by the auditory filter centered around that particular frequency. However, this assumption is not in line with various results in literature (e.g., [6]), where it is suggested that the detectability of a specific frequency component is also determined by off-frequency auditory filters.

Van de Par *et al.* introduced a perceptual distortion measure, which we will refer to as the Par-model, including spectral integration [7]. That is, the detectability of a specific frequency component is also determined by off-frequency auditory filters. This method showed better correspondence with data from psychoacoustic listening tests than the MPEG-1 model. Moreover, it does not need to separate the signal into tonal and noise maskers. It has been shown that the Par-model leads to better coding results compared to the MPEG-1 model for various fixed bit-rates in the field of sinusoidal coding [7]. In addition, the Par-model is defined as a mathematical norm, which allows for incorporating perceptual properties in least squares optimization algorithms. Examples are found in sinusoidal coding [8] and residual noise modeling [9]. Note that in the field of speech processing, mathematical tractable distortion measures are also used, like the log-spectral distance or distortion measures based on linear prediction (see, e.g., [10] and [11] for an overview). Although these measures include some perceptual properties they do not account for auditory masking effects.

Many perceptual models, like the Par-model and the MPEG-1 perceptual model, assume that the introduced error occurs simultaneously with the clean signal within one short-time frame (20–40 ms) and, therefore, do not take any temporal information into account. The consequence is that if an error is introduced before an onset of the clean signal in the same frame, these spectral models will consider the error to be masked, which is actually not the case. In fact, this will lead to so-called pre-echoes which are unwanted perceptual artifacts [1]. Although some

Manuscript received April 28, 2011; revised November 07, 2011; accepted January 05, 2012. Date of publication January 17, 2012; date of current version March 16, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick A. Naylor.

C. H. Taal is with the Sound and Image Processing Lab, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden (e-mail: chtaal@gmail.com).

R. C. Hendriks and R. Heusdens are with the Delft University of Technology, Signal and Information Processing Lab, 2628 CD Delft, The Netherlands.

Digital Object Identifier 10.1109/TASL.2012.2184753

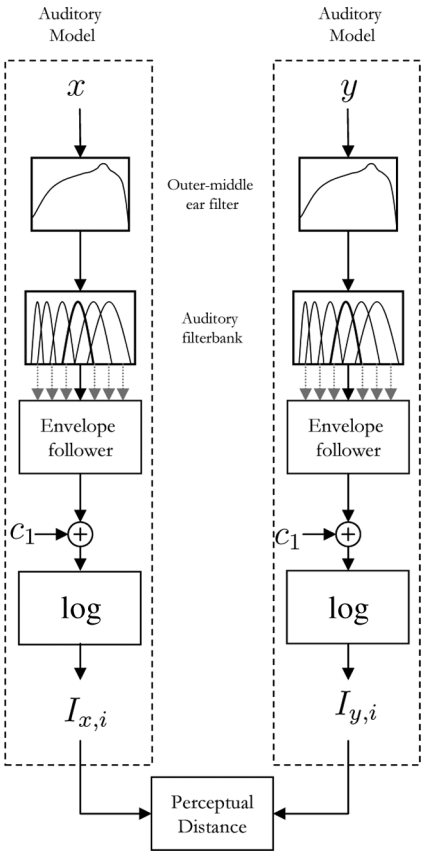


Fig. 1. Basic structure of the proposed model, which compares the internal representations I_x and I_y of the clean (x) and degraded (y) audio signal, respectively. First an outer middle ear filter is applied followed by an auditory filter bank. The haircell transduction staged is modeled by an envelope follower. Finally, a log-transform is applied to mimic the compressive properties of the outer haircells after which the internal representations are compared by means of applying a distance measure (see text below for more details).

backward masking may occur to mask the pre-echo, this is typically not sufficient since backward masking is only present a few milliseconds before the onset of the clean signal [12], [13]. A solution to prevent pre-echoes is called temporal noise shaping [14], which minimizes the squared error by means of frequency domain linear prediction. However, this method is not based on a perceptual model. Other solutions are window switching [1] and moving transient locations [15]. These methods are heuristic in nature and do also not take into account some type of perceptual model.

There are more advanced perceptual models available which do take into account time information. Examples can be found in the field of computational auditory modeling where neural firing patterns are obtained by modeling certain stages of the auditory periphery, e.g., [16], [17]. However, these approaches are not meant for optimization algorithms in (real-time) audio and speech processing applications and, as a consequence, may be computationally demanding. For example, in the advanced auditory model developed by Dau *et al.* [17], [18] (Dau-model) a masking threshold for a given error signal can only be found by using adaptive procedures [19], as is done in [18], and a closed-form analytic expression is not available. This means that when used in a coding environment, for each newly introduced quantization level the model must be applied several times in order to find an estimation of its masking threshold, which is computationally demanding. Another problem with these advanced models is that they are typically not defined for short-time frames, this in contrast to the Par-model and the MPEG-1 model. These properties make it difficult to use these advanced models in the applications we are interested in.

In this paper, a new distortion measure defined for short-time frames is presented based on a spectro-temporal auditory model. The measure is simplified under certain assumptions valid for the applications of interest in this article (e.g., coding, watermarking). This leads to a more tractable measure in the sense that analytic expressions now exist for masking thresholds. Furthermore, it will be shown that the proposed methods predicts similar masking thresholds compared to an advanced spectro-temporal model with a large reduction in complexity.

II. Preliminaries

Let x and y denote two finite length discrete-time signals of length N, representing the original and degraded audio signal, respectively. The degraded signal will be written as $y = x + \varepsilon$, where ε can be interpreted as the introduced degradation by the system of interest (e.g., quantization noise). The N-point discrete Fourier transform (DFT) of x, say $\hat{x}$, is defined as

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N}, \quad k = 0, \ldots, N-1 \qquad (1)$$

where k represents the DFT-bin index, j the imaginary unit and n the time index. Similar definitions hold for $\hat{y}$ and $\hat{\varepsilon}$. Furthermore, circular convolution will be denoted by $x \circledast y$. The ℓ_p-norm of x is defined as

$$||x||_p = \left(\sum_n |x(n)|^p \right)^{1/p}. \qquad (2)$$

In this work, we assume that all time-domain signals and filters are real valued.

III. Proposed Spectro-Temporal Distortion Measure

Fig. 1 shows the structure of the proposed method. First, an auditory model, which mimics certain stages of the auditory periphery, is applied to the clean and degraded signal in order to obtain their corresponding internal representations, denoted by $I_{x,i}$ and $I_{y,i}$, respectively, where i denotes the auditory channel. A perceptual difference is then defined by applying a distance measure between the internal representations denoted by "perceptual distance" in the figure. Note that this approach of modeling stages of the auditory periphery and comparing these signals in a spectro-temporal auditory domain is typically used by more advanced perceptual models, e.g., [16], [17], [20], and [21], and not by short-time models used in online optimization algorithms (like the Par-model) due to complexity reasons. However, we will show that under certain assumptions the complexity of such an advanced auditory modeling approach can be greatly reduced.

In Section III-A, more details will be given about the auditory model we use, followed by defining a perceptual distance measure between these internal representations in Section III-B. Then, under certain assumptions, the model will be simplified

in order to reduce its complexity in Section III-C, followed by some implementational details in Section III-D.

A. Auditory Model

The auditory model consists of a filter representing the frequency characteristics of the outer and middle ear, followed by an auditory filter bank resembling the properties of the basilar membrane in the cochlea. An envelope extraction stage is used to simulate the properties of the hair-cell transduction. Subsequently, a constant is added to represent physiological internal noise (caused by muscle activity, blood streams, etc.) in order to introduce an absolute hearing threshold. Finally, a log transform is applied to resemble the compressive behavior due to the outer hair-cells.

For the outer-middle ear filter a magnitude spectrum equal to the inverse of the threshold in quiet is used to let the model correctly predict the absolute hearing threshold. This threshold describes the playback level of a sinusoid, such that it is just not perceived by an average listener. A mathematical expression approximating the threshold in quiet can be found in [1]. For the auditory filter bank the same gammatone-based approach as in [7] is used. In total 64 filters are used where the center frequencies are linearly spaced on an ERB-scale between 0 and $f_s/2$ Hz, where f_s denotes the sample rate.

Let h_i denote the joint impulse response of the outer middle ear filter and the ith auditory filter where x filtered by h_i is denoted by $x_i = x*h_i$. Similarly we have $y_i = y*h_i$. Per channel, the envelope extraction stage is included by taking the absolute squared value followed by a low-pass filter, say h_s. With this, a mathematical description of the internal representation of x in the ith auditory filter can then be written as

$$I_{x,i} = \log\left(|x_i|^2 * h_s + c_1\right) \tag{3}$$

where c_1 denotes the constant representing internal noise. Similarly, the internal representation of y can be defined as

$$I_{y,i} = \left(|y_i|^2 * h_s + c_1\right). \tag{4}$$

B. Perceptual Distance between Internal Representations

In order to define a perceptual difference between x and y, their corresponding internal representations $I_{x,i}$ and $I_{y,i}$ should be compared somehow. One procedure is to apply an ℓ_p-norm on the difference between the internal representations of the clean and degraded audio signal, where increasing p will give more importance to high-energy regions in the eventual distance measure, e.g., spectral peaks in vowels. In this paper, we choose $p = 1$. As we will show (see Section V), for this choice of p the measure can be simplified into a mathematical tractable distortion measure while predicting results with sufficient accuracy are obtained compared to psychoacoustic listening experiments.

Applying an ℓ_1 norm to the difference between the internal representations gives a within-channel detectability defined by

$$d_i(x,y) = \|I_{y,i} - I_{x,i}\|_1. \tag{5}$$

These within-channel detectabilities are then combined by means of a summation in order to include the spectral integration properties of the auditory system

$$\begin{aligned} d(x,y) &= c_2 \sum_i d_i(x,y) \\ &= c_2 \sum_i \|I_{y,i} - I_{x,i}\|_1 \\ &= c_2 \sum_i \left\| \log\left(\frac{|y_i|^2 * h_s + c_1}{|x_i|^2 * h_s + c_1} \right) \right\|_1 \end{aligned} \tag{6}$$

where an additional calibration constant c_2 is included in order to set the sensitivity of the model (see Section III-D).

C. Low-Complexity Approximation

Equation (6) can be approximated by a simpler form which leads to an analytical expression for the masking threshold as we will show in Section IV. We assume that x and ε are uncorrelated, i.e., $E(X\mathcal{E}) = 0$, which gives the possibility to discard certain cross-terms in the within-channel temporal envelope of y. This assumption is typically valid for quantization noise in audio coders but also in data-hiding applications like watermarking. The within-channel temporal envelope of y can be expressed as

$$|y_i|^2 * h_s = |x_i + \varepsilon_i|^2 * h_s = |x_i|^2 * h_s + |\varepsilon_i|^2 * h_s + 2(x_i\varepsilon_i) * h_s. \tag{7}$$

As a consequence of the averaging properties of the smoothing low-pass filter h_s and the assumption that x and ε are uncorrelated, it holds that

$$2(x_i\varepsilon_i) * h_s \approx 2E(X_i\mathcal{E}_i) = 0. \tag{8}$$

Motivated by this the following approximation is used:

$$|y_i|^2 * h_s \approx \left(|x_i|^2 + |\varepsilon_i|^2\right) * h_s. \tag{9}$$

By combining (9) and (6) we get

$$d(x,y) \approx c_2 \sum_i \left\| \log\left(1 + \frac{|\varepsilon_i|^2 * h_s}{|x_i|^2 * h_s + c_1} \right) \right\|_1. \tag{10}$$

Next, we assume that only small errors are introduced to the clean signal which is typically the case in masking situations. Therefore, a good approximation of each element in the summation of (10) can be obtained by only taking into account the first term of the Maclaurin series expansion of $\log(1+z) \approx z$. That gives us the final expression for the new simplified measure, which will be denoted by D. That is,

$$d(x,y) \approx D(x,\varepsilon) \triangleq c_2 \sum_i \left\| \frac{|\varepsilon_i|^2 * h_s}{|x_i|^2 * h_s + c_1} \right\|_1. \tag{11}$$

For high playback level, i.e., $|x_i|^2 * h_s \gg c_1$, the measure reduces to a spectro-temporal, noise-to-signal ratio per auditory band. For very low playback levels, i.e., $|x_i|^2 * h_s \ll c_1$, it can be observed that the constant c_1 will dominate the denominator and therefore an absolute threshold in quiet is introduced.

D. Implementation Details

The parameters c_1 and c_2 are calibrated such that the model correctly predicts the threshold in quiet at 1 kHz and the 1 dB just noticeable level difference for a 70-dB SPL, 1-kHz tone (see also [7]). It is assumed that an additive distortion ε is just not detectable when $D = 1$. For this procedure the playback level of the audio signals must be known where we assume that the maximum playback level is 96-dB SPL.

For complexity reasons, the outer-middle ear filter, the auditory filter bank and the smoothing low-pass filter are all applied by means of a point-wise multiplication in the DFT-domain, where we assume that all filters have a real-valued, even-symmetric frequency response, i.e., $\hat{h}(k) = \hat{h}(-k)$. This particular choice will lead to time-domain aliasing due to circular convolution; however, proper windowing is used to minimize the effect of these unwanted artifacts. For the smoothing low-pass filter h_s the magnitude response of a one-pole filter is used with cutoff frequency $f_c = 1000$ Hz. The cutoff frequency controls the sensitivity of the model towards the temporal structure of the clean and degraded signals. The particular choice of $f_c = 1000$ roughly simulates the transduction properties of the inner hair cells [17]. Let $a = -e^{-2\pi f_c/f_s}$. The frequency response of h_s is then given by

$$\hat{h}_s(k) = \frac{(1+a)}{\sqrt{1+a^2+2a\cos(2\pi k/N)}}. \tag{12}$$

In order to save computational power the denominator in (11), i.e., $|x_i|^2 * h_s + c_1$, can be precalculated independent of ε. The measure can then be evaluated for any introduced error by just calculating the spectro-temporal envelope of ε divided by this precalculated term. In fact, the following gain-function can be precalculated independent of ε:

$$g_i^2 = \frac{c_2}{|x_i|^2 \circledast h_s + c_1} \circledast h_s \tag{13}$$

where the measure can then be expressed as follows (see Appendix A):

$$D(x,\varepsilon) = \sum_i \|\varepsilon_i g_i\|_2^2. \tag{14}$$

The measure can now be evaluated for any arbitrary error just by applying the DFT-based filter bank followed by a spectro-temporal gain function.

IV. Masking

A. Masking Threshold

Many applications are interested in a masking threshold of ε given x, i.e., the maximum level of ε such that it is just not detectable in the presence of x. This threshold can be found by solving $d(x, x+\alpha\varepsilon) = 1$ for α, where α is a scalar controlling the level of the introduced error. Notice that with the distance measure as defined in (6) it is not straightforward to determine a masking threshold. Instead of an analytical solution, a typical approach is to use adaptive procedures similarly to what is done with real listening experiments [19]. However, many iterations may be needed to determine an estimate of the masking threshold which may be computationally demanding. In addition, depending on the application the procedure has to be repeated for many different error signals ε. Nevertheless, due to the introduced simplifications for the proposed model, as explained in Section III-C, we now have the relation $D(x, \alpha\varepsilon) = \alpha^2 D(x,\varepsilon)$. This gives the following solution for the masking threshold:

$$\alpha = \frac{1}{\sqrt{D(x,\varepsilon)}}. \tag{15}$$

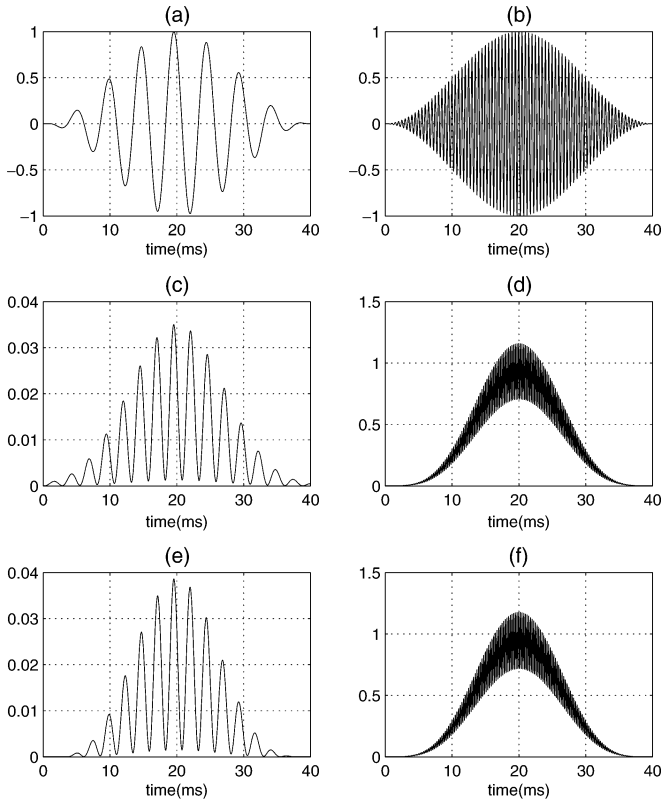


Fig. 2. (a) Windowed sinusoid of 200 Hz with (c) corresponding temporal envelope as defined in (16) and (e) approximated temporal envelope as explained in Section IV-B. Similar plots are shown in (b), (d), and (e) for a 2000-Hz sinusoid. Only the auditory filter is shown where its center frequency is closest to the frequency of the sinusoid.

B. Masking Curve

In applications like [7] and [22], knowledge of the masking curve is required which describes the masking threshold for a (windowed) sinusoid as a function of frequency. This masking curve will provide information on how to shape the spectrum of an introduced error such that perceptual impact of the error is minimized.

Unfortunately, evaluating (15) for all frequencies of interest (from 0 to $f_s/2$) may be computationally demanding. However, due to the introduced simplifications of the model as explained in the previous section an efficient DFT-based expression for the masking curve can be obtained. Let a windowed sinusoid (e.g., Hann window) be denoted by $\varepsilon_k(n) = w(n)\cos(2\pi kn/N)$,

where N is the DFT-size and k/N the normalized frequency of the sinusoid. For slowly time-varying windows the output of the auditory filter bank can be approximated as

$$\varepsilon_k \circledast h_i \approx \hat{h}_i(k)\varepsilon_k. \tag{16}$$

Note, that the auditory filters were defined such that they have a real-valued spectrum. Hence, no phase shifts and group delays have to be taken into account. Fig. 2 shows an example where the actual within channel temporal envelope, i.e., $|\varepsilon_k \circledast h_i|^2 \circledast h_s)$, and the estimated within channel temporal envelopes based on (16) are plotted for a 200- and 2000-Hz sinusoid. The plot only shows the auditory filter where its center frequency is closest to the frequency of the sinusoid. The figure reveals that a good approximation is obtained of the actual within channel temporal envelope for both frequencies.

In order to define a masking curve we have to solve $D(x, \alpha(k)\varepsilon_k) = 1$ for $\alpha(k)$. By using the approximation in (16) this gives

$$\frac{1}{\alpha^2(k)} = \sum_i \hat{h}_i^2(k)\|\varepsilon_k g_i\|_2^2 \tag{17}$$

which can be rewritten in the following form:

$$\begin{aligned}\frac{1}{\alpha^2(k)} &= \sum_i \hat{h}_i^2(k) \sum_n |(wg_i)(n)|^2 \left(\frac{1}{2} + \frac{1}{2}\cos\left(\frac{4\pi kn}{N}\right)\right) \\ &= \frac{1}{2}\sum_i \hat{h}_i^2(k) \\ &\quad \times \left(\|wg_i\|^2 + \sum_n |(wg_i)(n)|^2 \cos\left(\frac{4\pi kn}{N}\right)\right). \end{aligned} \tag{18}$$

Equation (18) can be expressed in terms of the DFT of the gain function for each auditory band multiplied with the squared window function, i.e., $|\widehat{wg_i}|^2$. That is,

$$\frac{1}{\alpha^2(k)} = \sum_i \hat{h}_i^2(k)\left(\frac{1}{2}|\widehat{wg_i}|^2(0) + \mathrm{Re}\left\{|\widehat{wg_i}|^2(2k)\right\}\right) \tag{19}$$

where $\mathrm{Re}\{\cdot\}$ denotes the real part of any arbitrary complex number. From this equation we can conclude that a complete masking curve can now be obtained by exploiting the (Fast) Fourier transform for $|\widehat{wg_i}|^2$ for each auditory band. Note, that this is a significant reduce in complexity compared to evaluating (15) for each sinusoid individually with frequency $k = 0, 1, \ldots, N/2$.

V. Model Evaluation and Comparison

To evaluate the proposed method, comparisons will be made with a sophisticated spectro-temporal model as proposed by Dau *et al.* [17], [18] and a simpler spectral-only model by van de Par *et al.* [7]. We will demonstrate that the proposed method shares some of the benefits of the complex Dau-model with respect to predicting masking thresholds for nonstationary signals, while it has a similar mathematical tractable form like the Par-model. First both reference models are explained after which comparisons are made by means of predicting masking curves and computational complexity.

A. Reference Models

1) Par-Model: The Par-model is based on the energy detection model from the field of signal detection theory as proposed by Green and Swets [23], where the task is to detect a probe (e.g., sinusoid) in the presence of some masker (e.g., white noise). For this model it is assumed that at the output of an auditory filter, the signal is absolute squared followed by a temporal integration procedure (note that this model is of a simpler form than the one which is used in the proposed method from Fig. 1). As a consequence, the listener observes the stimulus power at the output of an auditory band which is considered to be stochastic (e.g., due to internal noise). Under the assumption that the stochastic processes are independent and Identically distributed (i.i.d.) Gaussian and that the auditory system uses an optimal detector to detect the probe in presence of the masker it can be shown that the ratio between the increase in probe power and the standard deviation of the masker is defined as the sensitivity index d' [23]. The sensitivity index (i.e., distortion detectability) is monotonically increasing related to the probability of correctly detecting the probe in presence of the masker (i.e., a higher d' implies a higher probability of correctly detecting the probe in presence of the masker).

Van de Par *et al.* [7] suggested to combine the within-channel sensitivity indices over all auditory bands by means of an additive operation in order to mimic the spectral integration properties of the auditory system (see, e.g., [6] and [24]). Temporal integration is included by multiplying this summation with a factor N. As a consequence, increasing the playback length of a signal will result in a higher predicted detectability, which is in accordance with a human observer up until lengths of approximately 300 ms [25]. Similar as with the proposed method the auditory filters are implemented by means of a point-wise multiplication in the DFT-domain, hence, a circular convolution in the time-domain. This leads to the following perceptual distortion measure:

$$D_{par}(x, \varepsilon) = Nc_2 \sum_i \frac{\frac{1}{N}\|\varepsilon_i\|_2^2}{\frac{1}{N}\|x_i\|_2^2 + c_1} \tag{20}$$

where c_1 is included in order to introduce a threshold in quiet and c_2 is used to modify the sensitivity of the model. Both parameters are calibrated such that the model correctly predicts the masking threshold of a 1-kHz tone in silence and the 1 dB just noticeable level difference for a 70-dB SPL, 1-kHz tone. The model is calibrated such that $D_{par} = 1$ corresponds to a distortion at the threshold of detection of ε [7].

Note, that the Par-model also has an efficient implementation, where a gain function only depending on x can be precalculated [similarly as in (14)]. By using Parseval's theorem, i.e., $\|x\|^2 = (1/N)\|\hat{x}\|^2$, the following spectral weighting function can be used:

$$\hat{g}_{par}^2(k) = \sum_i \frac{h_i^2(k)c_2}{\frac{1}{N}\|\hat{x}_i\|^2 + Nc_1} \tag{21}$$

to express the Par-model as an efficient frequency weighted ℓ_2 norm [7]

$$D_{par}(x, \varepsilon) = \|\hat{\varepsilon}\hat{g}_{par}\|_2^2. \tag{22}$$

Van de Par *et al.* have shown that the masking curve for the Par-model can be directly related to the inverse of this spectral weighting function $\hat{g}_{par}$ [7]. However, the masking curve in [7] is based on rectangular-windowed, normalized complex exponentials rather than sinusoids. By introducing a normalization factor $\sqrt{2}/N$ a full masking curve for rectangular windowed sinusoids is given as follows (an efficient expression for the masking curve for other types of windows is not defined in [7]):

$$\alpha(k) = \frac{\sqrt{2}}{\hat{g}_{par}(k)N}. \tag{23}$$

2) Dau-Model: The Dau-model acts as an artificial observer and is originally used for accurately predicting masking thresholds for various masking conditions [17], [18]. It has a similar approach as the proposed method in the sense that it compares internal spectro-temporal representations. In order to obtain an internal representation, a 64-channel auditory filterbank is first applied, where the haircell transduction process is modeled by half-wave rectification followed by a 1-kHz low-pass filter. To introduce an absolute threshold, the hair cell output is limited to a minimum value. The auditory model is more advanced in the sense that it also models the nonlinear properties of the auditory system due to neural adaptation. This is incorporated by means of the so-called adaptation loops, which will put more emphasis on strong temporal fluctuations, e.g., transients, while more stationary sounds are converted approximately logarithmically [17]. Temporal integration of the auditory system is included by means of a 8-Hz low-pass filter per auditory band, followed by addition of internal noise simulated by Gaussian i.i.d. white noise. To let the model correctly predict the threshold in quiet, an outer-middle ear filter is applied before the auditory filterbank, similarly as with the proposed and Par-model.

In [17], the perceptual distance between two signals is determined by a correlation based comparison. Due to the addition of internal noise, the internal representations are stochastic and therefore this perceptual distance is also stochastic (similarly as with a real listener). Since we are interested in the average behavior of the model we use the approach from [26] and [27], where it has been shown that the average detectability can be described by summing the squared ℓ_2 norms between the internal representations, per auditory band. Let $\Psi_{x,i}$ and $\Psi_{y,i}$ denote the time-domain signals of the internal representations for the ith auditory band of the clean and degraded signal, respectively. In line with [26] its perceptual distance is then defined by

$$D_{dau}(x,y) = \frac{1}{\sigma}\sqrt{\sum_i \|\psi_{x,i} - \psi_{y,i}\|_2^2} \tag{24}$$

where σ represents the standard deviation of the internal noise. The calibration of σ and the used minimum value to limit the haircell output is done similarly as with the proposed method and the Par-model.

Note that for the Dau-model no analytic expression exists to obtain a masking threshold, in contrast to the Par-model and the proposed model. Instead, we use the bisection method to estimate the masking thresholds. The iterative procedure was stopped when the error was smaller than 0.1 dB. In order to obtain a masking curve, the masking threshold is determined for a limited set of 30 sinusoids, with frequencies logarithmically spaced between 100 and 10000 Hz. We found that 10–20 iterations was typically sufficient to obtain an estimate of the masking threshold.

B. Prediction of Masking Curves

To illustrate the correspondences and the differences between the two reference models and the proposed model several masking curves will be predicted. For all models a sample-rate of 44.1 kHz is used.

Masking curves are predicted for a 50-dB SPL, 1 kHz tonal masker with a length of 200 ms including 10-ms ramps. However, in this case three different time segments are analyzed as shown in Fig. 3, where masking curves are predicted before, during and after the onset of the tonal masker, denoted by Frame I, II, and III, respectively, in the figure. The first frame contains only silence, the second frame partly silence followed by a part of the sinusoid and the last frame is the complete windowed sinusoid. The three plots on the right show the predicted masking curves for all models. The bottom-right plot also contains results from psychoacoustic listening tests [28] to evaluate the model predictions.

For the first frame it can be observed that the predictions for all three models are in correspondence, where they correctly predict the masking curve to be equal to the threshold in quiet. However, for the second frame a clear difference is observed for the Par-model. While the proposed method and the Dau-model both predict a masking curve close to the threshold in quiet, the Par-model discards the preceding silence of the masker which leads to a significantly higher masking curve. Since backward masking (see, e.g., [12] and [13]) is only present from a few milliseconds before the onset of the masker, the masking curve for the first frame should be close to the threshold in quiet. This is in correspondence with the results predicted by the proposed method and the Dau-model. For the third frame, the sinusoidal masker is present in the complete frame; therefore, the predicted masking curves for all models are similar. In the bottom right plot, results from psychoacoustic listening experiments are shown [28] on top of the predicted masking curves, which are in accordance with the predictions for all models.

A similar example is illustrated in Figs. 4 and 5, which show a short-time segment of speech for a transient and a vowel region, respectively. In both figures the spectrum is downscaled for visual clarity. For the transient region one can clearly see that the masking curve is much higher for the Par-model compared to the proposed method and the Dau-model. Hence, the proposed method detects the sensitivity towards an introduced error before the onset of the transient similarly as the advanced Dau-model. Employing this property in an audio-coding context will lead to, e.g., less pre-echoes or more intelligible consonants. All three models are more in correspondence for the predicted masking curves for the vowel region as is shown in Fig. 5. This is due to the fact that the within-temporal envelopes of the vowel have more or less the same temporal structure as the windowed sinusoids which determine the masking curve.

Notice that the masking curves for the Dau-model are slightly lower for lower frequencies compared to the proposed model in Figs. 4 and 5. A possible cause for this could be the sensitivity of the adaptation loops towards the preserved phase structure

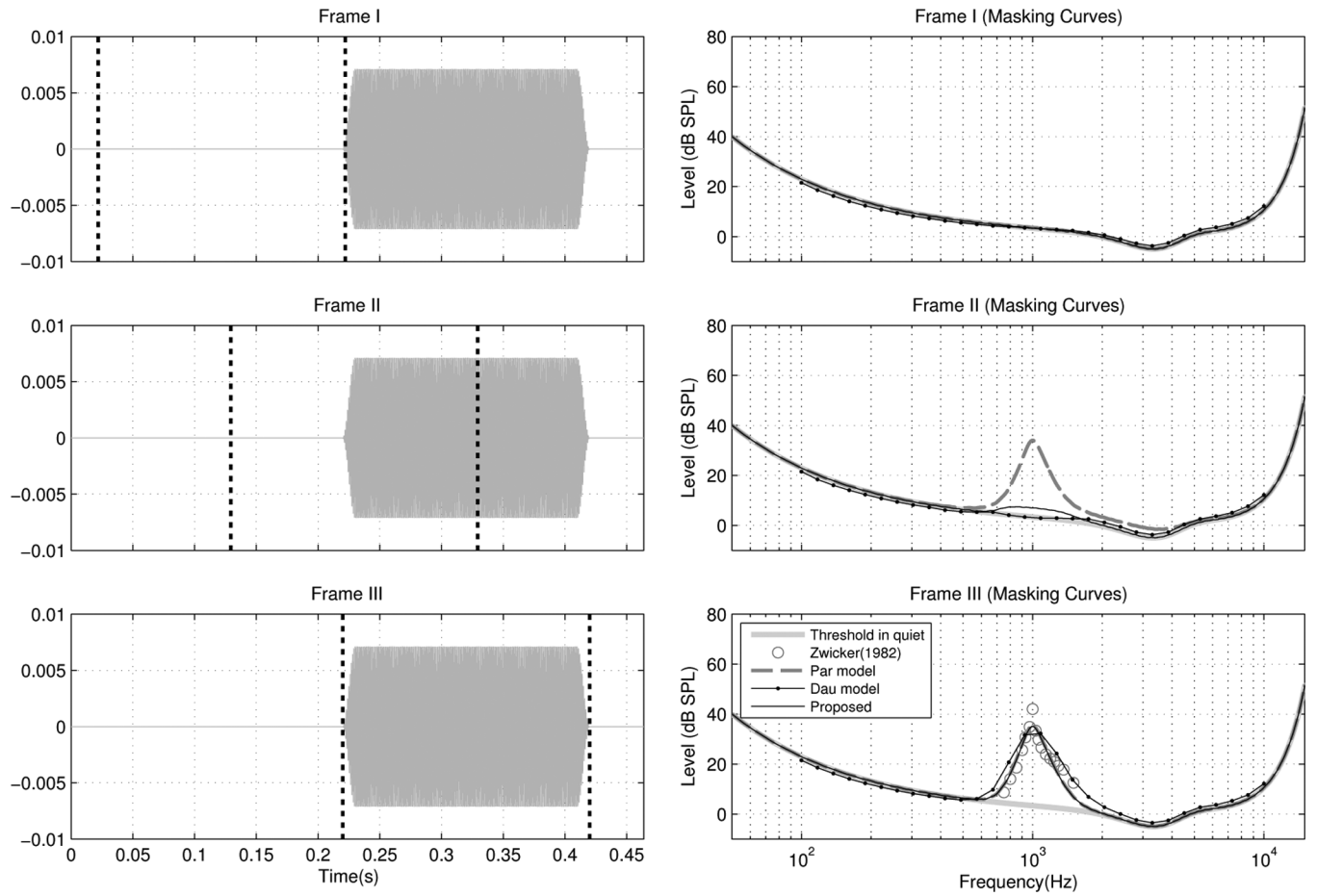


Fig. 3. Example to illustrate the difference between the proposed method, the spectro-temporal Dau-model and the Par-model [7] which is only based on spectral information. Masking curves are predicted by all models before (Frame I), during (Frame II) and after (Frame III) the onset of a 50-dB SPL, 1 kHz tonal masker with a length of 200 ms (subplots at the left). Their corresponding predicted masking curves are show in the right column plots, where the open circles in the bottom-right plot denote results from psychoacoustic listening experiments [28].

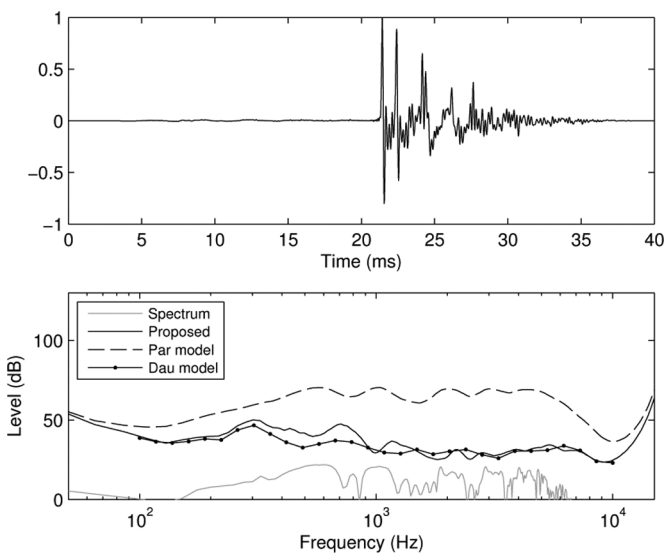


Fig. 4. Short-time (40-ms) transient region of speech (top plot) with predicted masking curves for the proposed method, the Par-model [7] and the Dau-model [17] (bottom plot). The spectrum is down-scaled for visual clarity.

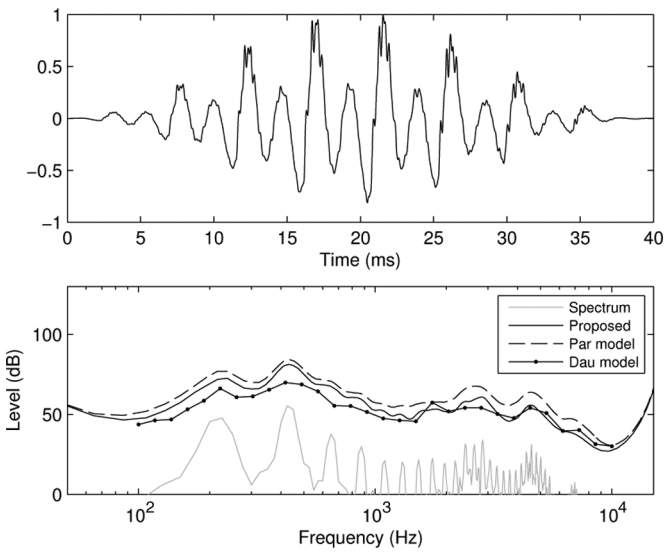


Fig. 5. Short-time (40-ms) vowel region of speech (top plot) with predicted masking curves (mc) for the proposed method, the Par-model [7] and the Dau-model [17] (bottom plot). The spectrum is down-scaled for visual clarity.

at lower auditory bands. However, the difference between the proposed model and the Dau-model is much smaller compared to the masking curve overestimation for the Par-model for the transient signal. We also would like to add that the Dau-model can also predict masking effects due to neural adaptation, i.e., forward and backward masking [12], [13]. This property is not present with the proposed method. However, we believe that for

TABLE I
NORMALIZED PROCESSING-TIME

Frame length N	128	256	512	1024	2048
	1) Model evaluation				
Par-model, Eq. (20)	0.55	0.69	1.00	1.64	3.99
Proposed, Eq. (11)	1.30	2.02	3.22	7.95	16.35
Dau-model, Eq. (24)	140.38	142.12	148.22	159.50	184.24
	2) Model evaluation with fixed x				
Par-model, Eq. (22)	0.26	0.26	0.28	0.29	0.35
Proposed, Eq. (14)	0.51	0.70	1.04	1.94	4.05
Dau-model, Eq. (24)	71.00	72.30	76.81	79.83	92.35
	3) Masking curve prediction				
Par-model, Eq. (23)	0.15	0.22	0.39	0.91	1.63
Proposed, Eq. (19)	0.86	1.34	2.46	9.07	21.48
Dau-model	No analytic expression available				

the applications of interest in this work, these masking effects are less important compared to the difference between a spectral-only and a spectro-temporal model.

C. Complexity

To give an impression of the computational power needed for the proposed method in relation to the two reference models, the computation time is measured for several frame lengths and conditions. All three models are implemented in Matlab. For the Dau-model the IIR-based auditory filterbank in [29] is used and the complex adaptation loops are implemented in a C++—based MEX file for computational efficiency. The experiments are performed on a laptop with an Intel Core2 Duo CPU T7700 running at 2.4 GHz. In total three different processing conditions are considered.

1) Evaluation of the perceptual distance for a given x and ε. This refers to (11), (20) and (24) for the proposed, Par and Dau-model, respectively.
2) Evaluation of the perceptual distance for a given ε when x is fixed. This is a relevant situation for, e.g., a rate-distortion loop in a coder. This refers to (14) and (22) for the proposed and Par-model, respectively. For the Dau-model (24) is used where $\psi_{x,i}$ is precalculated once and stored.
3) Evaluation of a complete masking curve given x. This refers to (19) and (23) for the proposed and Par-model, respectively. Note that the Dau-model is not included in this test since no analytic expression exists for a complete masking curve. A masking curve is typically used in data-hiding and coding applications to spectrally shape the introduced error in order to perceptually "hide" the introduced error more efficiently.

For each condition and model, Gaussian i.i.d. vectors of x and ε are generated[1] for $N \in \{128, 256, 512, 1024, 2048\}$. These are typical frame lengths relevant for digital audio and speech processing applications. The performance for each model, condition and frame length N is obtained by taking an average computation time over 100 evaluations. The results are shown in Table I where the processing times are normalized with respect to the first condition for the Par-model where $N = 512$. Notice that the numbers given in Table I are rough estimates that are meant as an indication. In general they depend on implementational details.

[1]A more realistic scenario would be to use speech or music for x; however, this will not affect its processing time.

From the table it is revealed that the proposed method is a factor 10–100 times faster than the Dau-model, depending on the frame length and type of test. The main reason for this difference in performance is most likely the use of a log-transform instead of the sophisticated adaptation loops and the use of an FFT-based filterbank instead of the IIR-based gammatone filters. Despite the fact that the Dau-model has no analytic expression for the masking curve available, an estimation of this curve could be obtained by means of an adaptive procedure per sinusoid (as explained in Section V-A2). However, this means that we have to evaluate the Dau-model for each of the $(N/2 + 1)$ sinusoids, multiplied with the number of iterations needed in order to obtain a masking threshold for one sinusoid (10–20 in the experiments from the previous section). Given that the evaluation of a complete masking curve for the proposed model is already much faster than evaluating the Dau-model only *once* (see Table I), one can imagine the large reduction in complexity with the proposed method when one is interested in a masking curve.

Taking into account short-time temporal information comes with a computational cost compared to spectral-only models like the Par-model. This is also what can be concluded from the table where the Par-model is, in general, 3–15 times faster than the proposed model depending on the frame-size and type of test. However, this difference is much smaller than the difference in performance between the proposed model and the Dau-model. Other ways to reduce the computational complexity of the proposed model can be considered by, e.g., reducing the amount of auditory filters.

VI. EXPERIMENTAL RESULTS

In this section, we demonstrate the properties of the proposed model by means of experimental results and make a comparison with the Par-model. The Dau-model is not included in this comparison since it does not provide the analytical expressions for masking thresholds and masking curves needed in order to generate the signals in the experiment, as will become clear in the remainder of this section.

To illustrate the properties of the proposed model, several audio signals are generated with degradations that are typical for audio and speech processing applications where auditory masking is exploited. A common approach is to spectrally shape the introduced errors according the masking curve in order to perceptually "hide" the introduced error efficiently. For these applications there is typically a constraint involved which influences the amount of added noise. For example, the total number of bits in an audio coder or the amount of information and robustness of an embedded watermark. For demonstration purposes, these errors are artificially introduced to several clean signals based on the proposed model and the Par-model after which their results are compared.

Clean signals are degraded by i.i.d. Gaussian noise where the noise-only signal is first segmented into short-time (32 ms), 50% overlapping windowed frames and filtered with the predicted masking curve belonging to the corresponding short-time frame of the clean signal. This filtering operation is applied by means of a point-wise multiplication in the DFT-domain, where a square root Hann analysis and synthesis window is used. The

total amount of noise that is added to the clean signal is controlled by a constraint on the segmental SNR. The level of the masking-curve filtered noise is adjusted per short-time frame, such that the summation of all individual frame-distortions for the model under consideration is minimized. With this approach it is expected that the proposed method will put less noise in transient regions and add more noise in more stationary frames, in contrast to the Par-model.

Let m denote the frame-index, M the total number of frames, r the segmental SNR constraint in dBs and $\alpha_m \varepsilon_m$ the masking-curve filtered noise for the mth frame. Here α_m is a scalar which controls the level of the noise in that particular frame. The globally optimal distribution of all noise-levels (i.e., α_m for $m = 1, \ldots, M$) is then given by finding the minimum of the following constrained cost function:

$$J(\alpha_{1,\ldots,M}, \lambda) = \sum_m D\left(x_m, \alpha_m \varepsilon'_m\right) + \lambda \left(\frac{1}{M} \sum_m 20 \log_{10} \left(\frac{\|x_m\|_2}{\|\alpha_m \varepsilon'_m\|_2} \right) - r \right) \quad (25)$$

where $\varepsilon'_m = \varepsilon_m \|x_m\|_2 / \|\varepsilon_m\|_2$ denotes a normalized version of ε_m, which implies $\|x_m\|_2 = \|\varepsilon'_m\|_2$. As a consequence of this normalization and using the relation $D(x_m, \alpha_m \varepsilon_m) = \alpha_m^2 D(x_m, \varepsilon_m)$ of (11), the cost function can be expressed as follows:

$$J(\alpha_{1,\ldots,M}, \lambda) = \sum_m D\left(x_m, \varepsilon'_m\right) \alpha_m^2 + \lambda' \left(\sum_m \log\left(\alpha_m^2\right) - r' \right). \quad (26)$$

where

$$r' = \frac{-M \log(10) r}{10}. \quad (27)$$

In order to find the optimal distribution of the noise over the frames, given the segmental SNR constraint, the minimum of (26) is found by setting the derivative of the cost function to zero with respect to $\alpha_{1,\ldots,M}$ and λ, that is,

$$\frac{\partial J(\alpha_{1,\ldots,M}, \lambda')}{\partial \alpha_m} = 2D\left(x_m, \varepsilon'_m\right) \alpha_m + \frac{2\lambda'}{\alpha_m} = 0$$
$$\frac{\partial J(\alpha_{1,\ldots,M}, \lambda')}{\partial \lambda'} = \sum_m \log\left(\alpha_m^2\right) - r' = 0. \quad (28)$$

Solving this gives

$$\alpha_l^2 = \frac{\left(e^{r'} \prod_m D\left(x_m, \varepsilon'_m\right) \right)^{1/M}}{D\left(x_l, \varepsilon'_l\right)} \quad (29)$$

where l is used to denote the frame-index of interest. Note, that due to the similarity between the proposed model and the Par-model the derivations for the Par-model in order to distribute the noise is identical. For the proposed model the cutoff frequency of the low-pass filter h_s was lowered to 125 Hz, which resulted in a better noise distribution between transient and stationary frames.

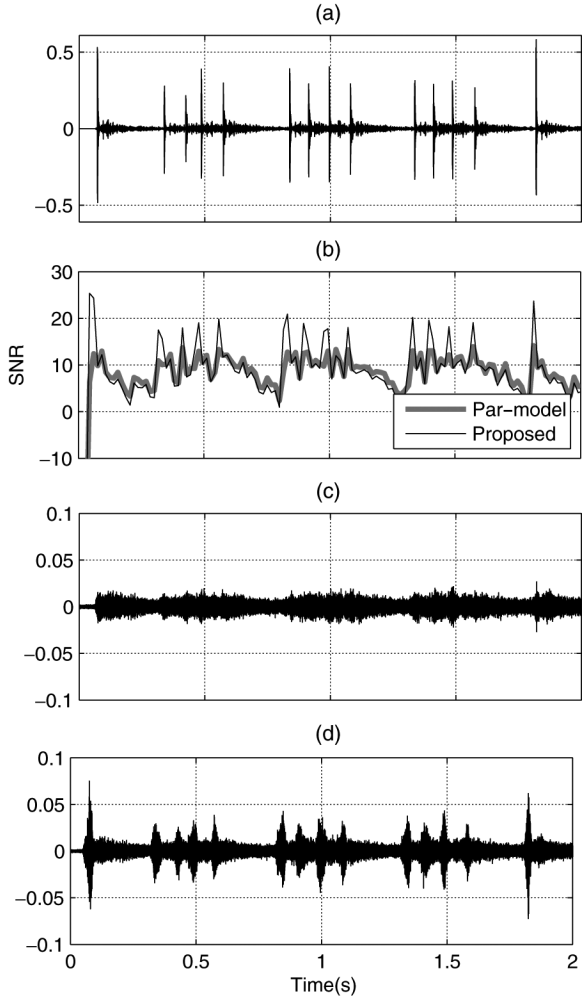


Fig. 6. Illustration of the noise distribution for the proposed model and the Par-model for the castagnettes excerpt. Subplot (a) shows the clean reference signal, where the distribution of the SNRs per frame for both models is shown in (b). Plots (c) and (d) show the added noise for both models. Notice that the proposed model detects the temporal structure within a short-time frame and puts less noise within transient-frames in contrast to the Par-model.

A. Example

To illustrate the differences in noise distribution between the proposed model and the Par-model, Fig. 6 shows the results for the castagnettes excerpt. Here, the segmental SNR was set to 10-dB SNR. In subplot (b), the SNR is plotted per frame, where it can be clearly observed that the proposed method increases the SNR in the frames when a transient is encountered (i.e., the proposed method adds less noise in these frames). The bottom two plots in Fig. 6 clearly show that the Par-model adds a lot of noise in the transient regions. The proposed method on the other hand adds more noise in the more stationary regions in order to fulfill the constraint. As will follow from the listening test (see the next section), adding more noise in the transient regions is perceptually more disturbing than the small increment of noise in the non-transient regions.

B. Listening Test

The proposed method and the Par-model are compared by means of an informal subjective listening test. Several excerpts

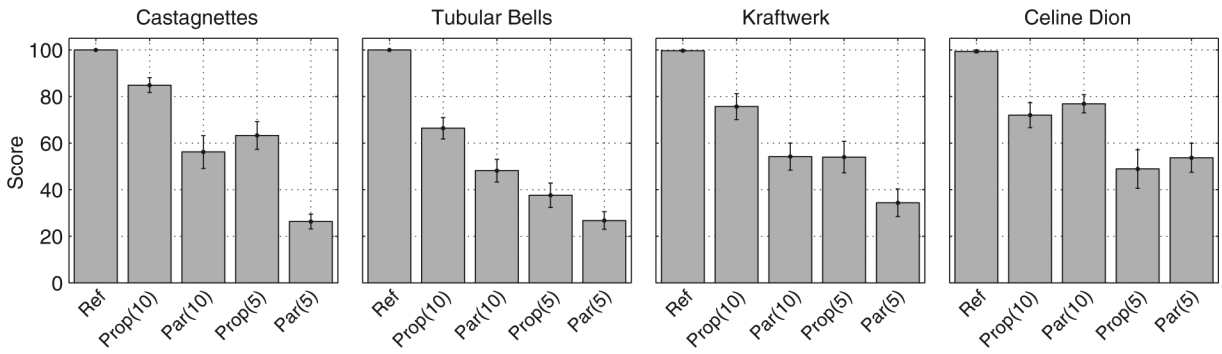


Fig. 7. Average results and standard errors across all subjects for all of the four excerpts. Noise was added to the reference signals at two different segmental SNRs (0 and 5 dB) for the proposed model (Prop) and the Par-model (Par). Higher scores imply better quality.

are degraded with the noise-distribution procedure as explained in the previous section. A sample rate of 44.1 kHz is used. The excerpts consist of castagnettes, tubular bells, Kraftwerk and Celine Dion which have a length of 7, 12, 12, and 13 seconds, respectively. Here the first three signals have strong transient regions, for which it is expected that the proposed model will show different performance than the Par-model. The Celine Dion fragment contains less transient regions and therefore more similar performance is expected between the two models for this excerpt. The constraints are set to 5- and 10-dB segmental SNR. In total, ten subjects participated in the listening test, which is similar to a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test [30]. The signals were presented via headphones, where the subjects were able to adjust their volume control to a comfortable level. In total, five different versions for each excerpt had to be ranked on a scale between 0–100 where a higher score denotes better quality. The five signals consist of four degraded versions of the excerpt (2 SNRS for each model) and a hidden reference. The subjects were instructed that a hidden reference was included and were asked to grade this signal with a score of 100. Furthermore, the subjects had access to the clean reference signal for comparison. The participants consisted of employees of Delft University of Technology and have performed in similar listening tests before. They were not connected in any way to this project.

The average scores of the listening test for all subjects are shown in Fig. 7 for each excerpt separately. From the results we can conclude that given a segmental SNR, the subjects preferred the proposed method over the Par-model for all signals, except for the Celine Dion excerpt. For Castagnettes and Kraftwerk, the proposed model has even similar performance at 5-dB SNR compared to the Par-model at 10-dB SNR. Statistical analysis is performed to verify whether these differences are significant by means of a statistical significance paired t-test for two dependent samples [31]. The null hypothesis is that both means are equal, while the alternative hypothesis corresponds to the situation that the mean score of the proposed model is higher than the score from the Par-model. Table II shows the p-values of the likelihood that the null hypothesis is true. The alternative hypothesis is accepted at a significance level of $\alpha = 0.05$. From this analysis it can be concluded that the proposed method shows statistically significant better performance for all excerpts, except Celine Dion. For the Celine Dion fragment, the difference between the Par-model and the proposed model was not statistically significant, as was hypothesized.

TABLE II
DETAILS ON THE PERFORMED T-TESTS FOR THE ALTERNATIVE HYPOTHESIS THAT THE SUBJECTIVE SCORE FOR THE PROPOSED MODEL IS HIGHER THAN THE PAR-MODEL

	Segmental SNR = 10 dB		Segmental SNR = 5 dB	
	Significant?	p-value	Significant?	p-value
Castagnettes	Yes	0.0011	Yes	0.0007
Tubular Bells	Yes	0.0048	Yes	0.0470
Kraftwerk	Yes	0.0106	Yes	0.0227
Celine Dion	No	0.7700	No	0.8785

VII. Relation Between Proposed Model and the Par-Model

In the previous experiments it was shown that the proposed method is more sensitive to transient regions compared to the Par-model. Notice that this sensitivity of the model towards the temporal structure of the signal can be controlled with the cutoff frequency f_c of the smoothing filter h_s. Here, a lower cutoff frequency implies a lower sensitivity towards the temporal structure and hence the model behaves more like a purely spectral distortion measure. In fact, it can be shown that the proposed model and the Par-model are identical when the cutoff frequency f_c of the smoothing low-pass filter h_s is set to 0 Hz in (11). Inspection of (12) shows that for a cutoff frequency of 0 Hz, we get the following magnitude response of h_s:

$$\hat{h}_s(k) = \begin{cases} 1, & k = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

Recall that the smoothing low-pass filter was implemented as a point-wise multiplication in the DFT-domain. Therefore, the output of the within-channel temporal envelope is now equal to its mean squared value:

$$\begin{aligned} \left(|x_i|^2 \circledast h_s\right)(n) &= \frac{1}{N} \sum_k \widehat{|x_i|^2}(k) \hat{h}_s(k) e^{j2\pi nk/N} \\ &= \frac{1}{N} \widehat{|x_i|^2}(0) \hat{h}_s(0) \\ &= \frac{1}{N} \|x_i\|_2^2. \end{aligned} \tag{31}$$

Note that the within-channel temporal envelope of x is now a constant value independent of time n. If we follow the same

procedure for obtaining the within-channel temporal envelope of the error ε, the distortion measure from (11) can then be expressed as

$$D(x,\varepsilon) = c_2 \sum_i \left\| \frac{\frac{1}{N}\|\varepsilon_i\|_2^2 u}{\frac{1}{N}\|x_i\|_2^2 u + c_1} \right\|_1 \tag{32}$$

where $u(n) = 1$ for $n = 0, \ldots, N-1$. The argument of the ℓ_1 norm is now a constant positive signal, independent of n. Therefore, the summation over n in this norm can be replaced by a multiplication with the total signal length N, which, in fact, gives the expression for the Par-model

$$D(x,\varepsilon) = Nc_2 \sum_i \frac{\frac{1}{N}\|\varepsilon_i\|_2^2}{\frac{1}{N}\|x_i\|_2^2 + c_1} = D_{par}(x,\varepsilon). \tag{33}$$

Note that the underlying auditory model of the Par-model is of a simpler form than the auditory model of the proposed spectro-temporal distortion measure (as explained in Section III-A). For example, a hair-cell model and a log-transform are not taken into account. With (33) we can conclude that the Par-model can actually be derived from a more complex auditory model if and only if $f_c = 0$. Also of interest is the multiplication with N in (33), which follows directly from the derivations. In the Par-model this multiplication was artificially introduced in order to include the temporal integration properties of the auditory system [7].

VIII. Conclusion

A new perceptual distortion measure is presented based on a sophisticated spectro-temporal auditory model, which is simplified under certain assumptions valid for auditory masking applications like coding or watermarking. This led to a more tractable distortion measure in the sense that analytic expressions now exist for masking thresholds. This is typically not the case for more advanced spectro-temporal models, which need computationally demanding adaptive procedures to estimate masking thresholds. Furthermore, the distortion measure is of a simpler form since it can be evaluated for any arbitrary error just by applying a DFT-based auditory filter bank, followed by a multiplication with a spectro-temporal gain function. This gain function is only dependent on the clean signal and denotes the sensitivity to errors over time and frequency and can be reused for any arbitrary error. The proposed method gave similar masking predictions as the advanced spectro-temporal Dau-model with only a fraction of its computational power.

It has been shown that the proposed model can be interpreted as an extended version of the Par-model: a perceptual model based on spectral integration which ignores time-information. The benefits of the proposed method compared to the Par-model are made clear in several experiments, from which it can be concluded that for nonstationary frames (e.g., transients) the Par-model underestimates the audibility of introduced errors and therefore overestimates the masking curve. As a consequence, the system of interest incorrectly assumes that errors are masked in a particular frame which may lead to audible artifacts like pre-echoes. This was not the case with the proposed method which correctly detects the errors made in the temporal structure of the signal.

Appendix

Derivation of spectro-temporal gain function g_i In this appendix, it will be shown how to rewrite (11)to (14). Recall that the distortion measure was defined as follows:

$$D(x,\varepsilon) = c_2 \sum_i \left\| \frac{|\varepsilon_i|^2 \circledast h_s}{|x_i|^2 \circledast h_s + c_1} \right\|_1. \tag{34}$$

Next we use the fact that the argument of the ℓ_1 norm in (34) is positive and the property $\|z\|_1 = \|z^{1/2}\|_2^2$ when $z \geq 0$. By defining the signal,

$$b_i = \frac{c_2}{|x_i|^2 \circledast h_s + c_1} \tag{35}$$

the distortion measure can now be expressed in terms of an inner product:

$$\begin{aligned} D(x,\varepsilon) &= c_2 \sum_i \left\| \left(\left(|\varepsilon_i|^2 \circledast h_s \right) b_i \right)^{\frac{1}{2}} \right\|_2 \\ &= c_2 \sum_i \left\langle \left(\left(|\varepsilon_i|^2 \circledast h_s \right) b_i \right)^{\frac{1}{2}}, \left(\left(|\varepsilon_i|^2 \circledast h_s \right) b_i \right)^{\frac{1}{2}} \right\rangle \\ &\quad \times c_2 \sum_i \left\langle |\varepsilon_i|^2 \circledast h_s, b_i \right\rangle. \end{aligned} \tag{36}$$

By applying Parseval's theorem we get the following expression in the frequency domain:

$$D(x,\varepsilon) = \frac{1}{N} \sum_i \left\langle (\widehat{|\varepsilon_i|^2 \circledast h_s}), \hat{b}_i \right\rangle. \tag{37}$$

By using the duality of a circular convolution in the time-domain and a point-wise multiplication in the frequency domain we have

$$D(x,\varepsilon) = \frac{1}{N} \sum_i \left\langle (\widehat{|\varepsilon_i|^2}) \hat{h}_s, \hat{b}_i \right\rangle = \frac{1}{N} \sum_i \left\langle (\widehat{|\varepsilon_i|^2}), \hat{b}_i \hat{h}_s^* \right\rangle. \tag{38}$$

Since $\hat{h}_s$ was defined real (see Section III-D) we have that $\hat{h}_s = \hat{h}_s^*$. Therefore, by applying Parseval's theorem again the following measure in the time-domain is obtained:

$$D(x,\varepsilon) = \sum_n \left\langle |\varepsilon_i|^2, b_i \circledast h_s \right\rangle. \tag{39}$$

Now let

$$g_i^2 = b_i \circledast h_s = \frac{c_2}{|x_i|^2 \circledast h_s + c_1} \circledast h_s \tag{40}$$

be defined as a spectro-temporal varying gain function. Due to the fact that $g_i \geq 0$, the proposed method can now be written as a summation of weighted ℓ_2 norms per channel:

$$D(x,\varepsilon) = \sum_{n,i} |\varepsilon_i(n) g_i(n)|^2 = \sum_i \|\varepsilon_i g_i\|_2^2 \tag{41}$$

REFERENCES

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

[2] M. Swanson, B. Zhu, A. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Process.*, vol. 66, no. 3, pp. 337–355, 1998.

[3] *Coding of Moving Pictures and Associated Audio for Storage at Up to About 1.5 mbit/s, Part 3: Audio*, ISO/IEC 11172-3, I. Committee, 1993.

[4] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2, no. 2, pp. 60–74, 1995.

[5] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 700–708, Jul. 2004.

[6] S. Buus, E. Schorer, M. Florentine, and E. Zwicker, "Decision rules in detection of simple and complex tones," *J. Acoust. Soc. Amer.*, vol. 80, no. 6, pp. 1646–1657, 1986.

[7] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 9, pp. 1292–1304, 2005.

[8] R. Heusdens and S. Van De Par, "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. 1809–1812.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 4, pp. 189–192.

[10] A. H. Gray, Jr and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 380–391, Oct. 1976.

[11] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988, pp. 1–377.

[12] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*. New York: Springer-Verlag, 1990.

[13] B. Moore, *An Introduction to the Psychology of Hearing*. Bingley , U.K.: Emerald Group, 2003.

[14] J. Herre, "Temporal noise shaping, quantization and coding methods in perceptual audio coding: A tutorial introduction," in *Proc. Audio Eng. Soc. Conv. 17*, 1999, pp. 312–325.

[15] R. Vafin, R. Heusden, and W. Kleijn, "Modifying transients for efficient coding of audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 5, pp. 3285–3288.

[16] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, ASSP-1982, vol. 7, pp. 1282–1285.

[17] 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. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[18] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3623–3631, 1996.

[19] H. Levitt, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 467–477, 1971.

[20] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part I-time-delay compensation," *J. Audio Eng. Soc*, vol. 50, no. 10, pp. 755–764, 2002.

[21] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, and K. Brandenburg *et al.*, "PEAQ—The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc*, vol. 48, pp. 3–29, 2000.

[22] R. Heusdens, J. Jensen, W. B. Kleijn, V. Kot, O. A. Niamut, S. van der Par, N. H. van Schijndel, and R. Vafin, "Bit-rate scalable intraframe sinusoidal audio coding based on rate-distortion optimization," *J. Audio Eng. Soc*, vol. 54, no. 3, pp. 167–188, 2006.

[23] D. Green and J. Swets, *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

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

[25] G. van den Brink, "Detection of tone pulse of various durations in noise of various bandwidths," *J. Acoust. Soc. Amer.*, vol. 36, no. 6, pp. 1206–1211, 1964.

[26] A. Kohlrausch, J. Koppens, W. Oomen, and S. van de Par, "A new perceptual model for audio coding based on spectro-temporal masking," in *Proc. Audio Eng. Soc. Conv. 124*, 2008, 5.

[27] J. H. Plasberg and W. B. Kleijn, "The sensitivity matrix: Using advanced auditory models in speech and audio processing," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 310–319, Jan. 2007.

[28] E. Zwicker and A. Jaroszewski, "Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels," *J. Acoust. Soc. Amer.*, vol. 71, no. 6, pp. 1508–1512, 1982.

[29] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Proc. Auditory Physiol. and Percept.—Proc. 9th Int. Symp. Hear.*, 1992, vol. 83, pp. 429–446.

[30] *Method for the Subjective Assessment of Intermediate Quality Level of coding Systems*, ITU-R BS. 1534-1, ITU, 2001.

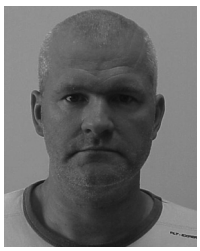
[31] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures, Third Edition*. Boca Raton, FL: Chapman & Hall/CRC, 2004.

Cees H. Taal received the B.S. and M.A. degrees in arts and technology from the Utrecht School of Arts, Utrecht, The Netherlands, in 2004 and the M.Sc. degree in media and knowledge engineering from the Delft University of Technology (DUT), Delft, The Netherlands, in 2007.

From 2008 to 2012, he was a Ph.D. Researcher in the Multimedia Signal Processing Group, DUT, under the supervision of R. Heusdens and R. Hendriks in collaboration with Oticon A/S. Currently, he is a Postdoctoral Researcher at the Sound and Image Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden. His main research interests are in the field of digital signal processing in audiology, including auditory modeling, speech enhancement, and intelligibility improvement.

Richard C. Hendriks received the B.Sc., M.Sc. (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from the Delft University of Technology (DUT), Delft, The Netherlands, in 2001, 2003, and 2008, respectively.

From 2003 to 2007, he was a Ph.D. Researcher at DUT, and from 2007 to 2010, he was a Postdoctoral Researcher at DUT. Since 2010, he has been an Assistant Professor in the Multimedia Signal Processing Group of the Faculty of Electrical Engineering, Mathematics and Computer Science, DUT. In the autumn of 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. From March 2008 to March 2009, he was a visiting researcher at Oticon A/S, Copenhagen, Denmark. His main research interests are digital speech and audio processing, including single-channel and multi-channel acoustical noise reduction, speech enhancement, and intelligibility improvement.



Richard Heusdens received the M.Sc. and Ph.D. degrees from the Delft University of Technology, Delft (DUT), The Netherlands, in 1992 and 1997, respectively.

Since 2002, he has been an Associate Professor in the Department of Mediamatics, DUT. In the spring of 1992, he joined the Digital Signal Processing Group at Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group, DUT, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio and speech processing activities within the ICT Group. He held visiting positions at KTH (Royal Institute of Technology, Sweden) in 2002 and 2008. He is involved in research projects that cover subjects such as audio and speech coding, speech enhancement, signal processing for digital hearing aids, distributed signal processing, and sensor networks.