Audio Coding with an Excitation Pattern Distortion Measure

Ricky Der

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

January 2005

A thesis submitted to McGill University in partial fulfillment for the degree of Master of Engineering.

Dimly felt, sounds yet faint,
now clarion,
falls the shroud
to behold the spring smile of
a mountain flower.

Abstract

A new distortion measure for audio coding is proposed, posed as a distance measure on the space of excitation patterns. We investigate the psychoacoustic properties of the measure, as well as the implementation issues that arise under a constrained-distortion coding structure. Experimental results show that the excitation distortion metric produces higher-quality coded files than the usual Noise-to-Mask ratio measure, at the same rate.

Sommaire

Une nouvelle mesure de distortion pour codage audio est proposée sous la forme d'une mesure de distance dans l'espace des formes d'excitation. Nous étudions les caractéristiques psychacoustiques de la mesure, ainsi que les facteurs reliés à l'implémentation, issus d'une structure de codage contrainte par la distortion. Les résultats expérimentaux démontrent que la métrique de la distortion d'excitation produit des fichiers codés de meilleure qualité que ceux produits par la mesure conventionelle utilisant le rapport entre le bruit et le masque, lors du codage au même débit.

Acknowledgments

My gratitude goes to Prof. P. Kabal and Prof. W.-Y. Chan for being discerning and challenging supervisors, as well as Prof. P. Depalle for his careful examination of the work. Various members of the TSP Lab also provided technical assistance, particularly Alex.

Contents

Prolegomena

“Do your own thing. Your reward will be doing it, your punishment having done it.”
—K. Potas

The problem of audio coding concerns the approximate representation of a sound signal, while utilising the fewest resources possible. We can exactify this statement by replacing the word “approximation” with a distortion function D, and the word “resources” by the rate function R. If x is a target signal and $\tilde{x}$ its representation, audio coding is nothing more than the goal of — in so far as it is possible to do — the simultaneous minimization of $D(x, \tilde{x})$ and $R(x, \tilde{x})$ — a classical problem certainly embracing applications of far greater generality than merely the class of audio functions.

The apposite questions are, of course, how to define D and R appropriately. For the latter, there is a consensus: the definition should in some way quantify the cardinality of the set of representations, whether deterministically or statistically. More difficult, on the other hand, is the specification of D. Nearly everyone will agree with the following abstract definition: a distortion measure on a set S is a function $D(x, y) : S \times S \to \mathbb{R}^+$ which quantifies the difference between two elements x, y from the set S. This prescription, naturally, contains insufficient structure to be useful even to mathematicians, who have gone lengths to find constraints on D that are intuitively appealing and lead to generalised notions of closeness and continuity (such as those of metric and topological spaces). The aim of the psychoacoustician is to find models of D leading to faithful representations of physiological and psychological auditory phenomena. An audio engineer seeks the constraints upon D which are useful (not necessarily *truthful*) to the application at hand: some approximation to the term “auditory distortion” — an ill-defined and somewhat subjective concept — which yet may be efficiently incorporated into coding structures.

Subjective as the term may be, most contemporary audio coders do, nevertheless, make use of simple rules borrowed from the science of hearing, *psychoacoustics*, in an attempt to formulate rough measures of distortion. In particular, modern coders exploit the phenomenon of *auditory masking*. This is little more than the rendering of some sound inaudible by the simultaneous presentation of a (usually stronger), masking sound. The phenomenon manifests itself in a myriad different ways in everyday life; noise from a passing train rendering conversation impossible is one oft-experienced example. Models of masking lead to the definition of a psychoacoustically motivated distortion function, known as the Noise-Mask-Ratio (NMR). That it has proved more successful than distortion functions rooted purely on mathematical concerns — squared error, for example — is a testament to the relative accuracy of the models in capturing the content of auditory difference.

This thesis presents an alternative to the ubiquitous Noise-Mask-Ratio. Indeed we will argue that conventional approaches utilising masking results are often inappropriate, and should be replaced by a more general class of distortion measures, which we term measures of "Excitation Distortion" (ED). Like NMR, ED measures will be psychoacoustically founded; they will be, however, applicable in quantitating auditory difference even in scenarios where masking is not involved, while reducing to masking results in the appropriate limits. Our main goals will be to: 1) Construct a conceptually sound (both from the mathematical and psychoacoustic viewpoints) framework for ED, 2) Examine the issues which arise when applying ED in the coding process, 3) Develop efficient allocation algorithms for minimizing ED, and 4) Perform comparative experimental tests between ED and NMR under a coding context.

The structure of the thesis naturally mimics the sequence of stated aims. We begin with a review of the concept of excitation pattern, which is so critical to both formulations of ED and NMR. Linearised models of the excitation transformation are developed, including a linear Bark-domain transformation. The relationship between the excitation pattern and absolute threshold is examined, for which purpose we introduce the concept of an internal noise excitation density. Finally, we tantalise the reader with the possibility of *complex* excitation patterns, which preserve some phase information. This idea may prove particularly germane toward a complete characterization of distortion.

Chapter II proceeds to argue the deficiencies in the NMR approach, motivating the need for other distortion measures. The notion of excitation distortion is then formally introduced, buttressed with a mathematical groundwork consistent with psychoacoustic threshold phenomena. This centers around the concepts of *gain-invariance* and *scale-invariance*. The class of all gain-invariant functions is derived, as well as a generalisation to Zwicker's 1 dB just-noticeable criterion. We proceed to design and specify a parametric family of distortion measures, which are optimal approximations to loudness difference. We then examine the properties of this class, and its ability to predict key psychoacoustic results.

The subsequent chapter considers the application of the newly-designed metrics to the audio coding problem. After an overview of the concept of equivalent distortion measures and basic material on quantization, we introduce the operational rate-distortion paradigm. Dependent versus independent quantization problems are defined in a very general sense, and we show how ED can assume either form depending upon the domain of quantization. The possibility of excitation-domain coding is briefly touched upon. A constrained-distortion framework is setup, *causal* coding is considered, but discarded. Next, non-causal (global) constrained-distortion coding is investigated, most urgently in response to the issues raised by lapped transform representation. We review what the operational rate-distortion literature has to offer in terms of solving the bit-allocation problem for dependent quantization problems in an optimal way, as well as standard greedy approaches to bit-allocation. Both avenues turn out to be rather unsatisfactory, in terms of either time-complexity or inability to achieve distortion constraints. A new time-efficient incremental bit allocation algorithm is developed to minimize ED (as well a host of other measures) in the stream coding context.

The final chapter exhibits our efforts to experimentally validate ED with respect to NMR. The techniques of the previous sections are applied to create constrained-distortion coded files, and matched-rate pairs are subjected to listening tests.

While this thesis records — as much as it was within the author's ability to do so — a somewhat coherent and unified account of the always messy and non-coherent research process, there yet remain, from time constraints, a plenitude of un-investigated paths and possibilities. The present work will perhaps thus read less as a *finished* product, and more as the reminiscences of an organized diary, collecting sundry theorems, notes, analyses and diversions that have occurred along the way. We have attempted to create a somewhat comprehensive record; a certain brevity has been correspondingly exacted. Thus, for instance, we have provided no proofs to any of the theorems. The omission allows a succinct presentation of qualitative content; at any rate, the intrepid reader will want to attempt these exercises for himself. In this way, it is hoped that the present document will serve as a useful initialization point for the tyro wishing to pursue these ideas further.

Chapter 1

Aspects of the Excitation Mapping

1.1 Basic Theory of Excitation Patterns

An excitation pattern is a function quantifying the (average) physical activity of hair cells along the basilar membrane, in response to a stimulus. One of the most successful concepts in psychoacoustics, excitation patterns provide unified explanations of such disparate phenomena as pitch discrimination, just-noticeable differences in amplitude, masking, loudness and absolute threshold, to name a few.

The distributions are not usually measured directly with probes, but indirectly by observing masked audiograms, or calculated from the signal power spectrum via a model. Such models broadly conform to the sequence of steps delineated in Figure 1.1. What follows is a compact review of each of these components, and a collection of a few standard formulae which will prove useful for calculations. The reader may consult the standard references [37] and [17] for further elaboration.

For the sequel, we use the following symbols to denote the signal at each stage: 1) P — power spectrum, 2) $\tilde{P}$ — ear-filtered power spectrum, 3) E — frequency-spread signal (excitation pattern), 4) E' — frequency spread signal plus internal noise, 5) $\tilde{E}$ — time-spread excitation pattern.

1.1.1 Fixed Ear Filter

The frequency response of the transformation from free-field to the inner ear has been measured with fair accuracy, and the data enumerated in an ISO standard. Moreover, bounds on the response from the 10th–90th percentiles are known, as well as variation with respect to age [37]. Analytical approximations to the response data usually involve a tripartition: a term for each of low, middle and high frequencies respectively. One such example is the formula contained in the

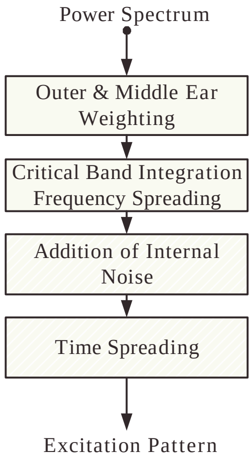

Power Spectrum
Outer & Middle Ear Weighting
Critical Band Integration Frequency Spreading
Addition of Internal Noise
Time Spreading
Excitation Pattern

Fig. 1.1 Generic Excitation Model

ITU standard, “Perceptual Evaluation of Audio Quality (PEAQ)”, valid for frequencies $f > 0$ (Hz) [14]:

$$H_{\mathrm{dB}}(\omega) = -2.184(f/1000)^{-0.8} + 6.5e^{-0.6(f/1000-3.3)^2} - 0.001(f/1000)^{3.6} \quad (\mathrm{dB}). \tag{1.1}$$

1.1.2 Critical Band Integration/Frequency Spreading

The cochlear membrane is most often modelled as a continuous bank of (possibly non-linear) auditory filters. The excitation pattern of a signal S at frequency f is then (theoretically) defined as the power at the output of each filter centered at frequency f in response to signal S. Practically, one obtains excitation data as follows. For simple stimuli, such as sinusoidal functions, the excitation pattern E is determined first by experimentally measuring the masking threshold $M(\omega)$ for a noise masker and sinusoidal maskee S with frequency ω. E is then assumed to be a scalar multiple of M, with the appropriate scaling constant set as the ratio between masker and maskee powers when ω coincides with the center frequency of the noise masker. More complex excitation patterns are built up as superpositions of these basic sinusoidal excitation vectors.

Perhaps the most coherent model of the excitation transformation, and one that summarises the description above, is given by Moore and Glasberg [18], [8]. Instead of partitioning the excitation transformation into the distinct steps of 1) spectrum integration over unit Bark-lengths, and 2) frequency spreading, as does Zwicker [37], the two are conflated by introducing auditory filters of infinite support but of finite area. This allows for compact representation of the excitation pattern $E(\Omega)$ at frequency Ω as:

$$E(\Omega) = \int_0^\infty W_P(\omega, \Omega)\tilde{P}(\omega)\, d\omega \tag{1.2}$$

where $W_P(\omega, \Omega)$ is the kernel of auditory filters and $\tilde{P}(\omega)$ is the input (ear-filtered) power spectrum. We have appended a subscript P to the kernel to indicate that the filters are level-dependent on the argument $\tilde{P}$. Otherwise, this transformation very much resembles a linear operation. Indeed, using sinusoidal inputs $\tilde{P}(\omega) = \delta(\omega - \omega')$, one sees that the basis functions are of the type $W(\omega', \Omega)$ with fixed ω'. Assuming the following parametric form for the auditory filters,

$$W(\omega, \Omega) = \left(1 + p\frac{|\omega - \Omega|}{\Omega}\right) e^{-p\frac{|\omega-\Omega|}{\Omega}}, \tag{1.3}$$

Moore and Glasberg fitted the experimental data to obtain an algorithm for the shape parameter p, as a function of $\tilde{P}$, and thus for W. We shall not reproduce it here, since the result is somewhat complicated; rather, we shall investigate a linearised version in Section 1.2.

Non-Linear Addition of Patterns

Let $\tilde{P}_1(\omega)$ and $\tilde{P}_2(\omega)$ be two power spectra with localized, disjoint supports. If the frequency separation between the two signals is large, then (1.2) gives the total excitation pattern $E(\tilde{P}_1 + \tilde{P}_2) \simeq E(\tilde{P}_1) + E(\tilde{P}_2)$ with $E(\tilde{P}_i)$ the excitation due to $\tilde{P}_i$ alone. If the shape parameter p is constant, then (1.2) is always a linear superposition of basis functions. It has been observed, however, that the masking threshold of the sum of two sinusoids is somewhat higher than that which would be predicted by a mere linear addition alone [16]. This "excess masking" has motivated some psychoacousticians to define a non-linear superposition property via $E(\tilde{P}_1 + \tilde{P}_2) = (E^q(\tilde{P}_1) + E^q(\tilde{P}_2))^{1/q}$, $0 < q \leq 1$, termed "spreading in the q-power domain". Typical values for q range from 0.2–0.4.

One may define q-power domain spreading more generally using

$$E_q(\Omega) = \left(\int_0^\infty W_P^q(\omega, \Omega)\tilde{P}^q(\omega)\, d\omega\right)^{1/q}. \tag{1.4}$$

It is an immediate consequence of Jensen's inequality that $E_q(\Omega) \geq E_{q'}(\Omega)$ whenever $q \leq q'$ — thus low values of q increase the amount of additional spreading.

Excitation and Masking

As noted above, the excitation pattern is usually experimentally determined by observing the masking pattern for a critical-band noise masker with sinusoidal maskee. That it thus provides a way to compute the masking threshold for this case is hardly surprising. All that is required is to determine the requisite scaling factor, or what is known as *masking offset* as a function of the masker frequency. Both Zwicker [37] and Kapust [32] have provided approximations to this data; the latter for $\omega > 0$ (Hz),

$$s_{\mathrm{dB}}(\omega) = -2 - 2.05 \arctan(\omega/4000) - 0.75 \arctan\left(\frac{1}{2.56}\left(\frac{\omega}{1000}\right)^2\right), \tag{1.5}$$

$$s(\omega) = 10^{s_{\mathrm{dB}}/10}. \tag{1.6}$$

The masked threshold $M(\omega)$ is then given by $s(\omega)E(\omega)$. While convenient, the reader should be reminded that this formula is only valid rigorously in the noise-masker tone-maskee scenario. It is a commonplace experience, particularly in the audio engineering literature, to speak cavalierly of *the* masking threshold for a given signal, independently of the target; this is a useful simplifying device, but only an approximation.

Bark-domain Patterns

It is well known that the bandwidths of the auditory filters increase with center frequency. This is a direct consequence of the logarithmic-like mapping between the frequency of a sinusoidal stimulus and its place of maximum excitation along the cochlear. The transformation between linear frequency and what is called *tonal* frequency is a further warping that is sometimes included in models of excitation. Varying measures of tonal frequency have been formulated, such as the Barkhausen scale [37] or "number of ERBs (Equivalent Rectangular Bandwidth)" [18]. We shall make use of the Bark frequency conversion formula of [14]:

$$\Omega = 650 \sinh(z/7), \tag{1.7}$$

where Ω is Hertz frequency and z Bark frequency. Its convenience lies in the fact that the expression is easily analytically invertible, something not shared by the conversion formula of [37], for instance. Such conversion is most usefully applied in re-scaling the frequency axis of an excitation pattern (more generally, any function of frequency), for instance by substitution of (1.7) into (1.2).

1.1.3 Internal Noise Excitation

The commonly accepted mechanism behind absolute threshold is that omnipresent internal noise masks low-level inputs [37]. This noise excitation is an additive signal possessing strong low-frequency components, rapidly decaying above 500 Hz. Most computations of excitation in fact do not include this additive term; internal noise finds application mostly in loudness models; for example that of [20]. One exception is the PEAQ standard [14], where an internal noise term is added to the power spectrum prior to spreading. This has the effect of introducing a lower-bound on the masking function, approximately accounting for absolute threshold in a way less heuristic than, for instance, Johnston's psychoacoustic model [15].

We can define $\epsilon(\omega)$ as the noise excitation level per critical band necessary to mask a just-noticeable sinusoid at frequency ω [37]. The definition gives a constructive way for its computation: if $s(\omega)$ is the threshold factor in a general critical-band-wide noise-masking-tone situation (not necessarily at absolute threshold), then

$$\epsilon(\omega) = \frac{|H(\omega)|^2 A(\omega)}{s(\omega)} \tag{1.8}$$

where $|H(\omega)|^2$ is the square magnitude response of the fixed outer-middle ear filter and $A(\omega)$ is the power of a just-noticeable sinusoid.

Since the threshold variable s is computed with respect to a unit critical-bandwidth noise masker, the form of the internal excitation given above is appropriate for models using auditory filters of unit ($l = 1$) critical-band width. There are a number of psychoacoustic models in the engineering literature which integrate along *sub*-critical band lengths, among them the MPEG-I standard ($l = 0.34$ Bark-width), as well as PEAQ ($l = 0.5$ and $l = 0.25$). While these sub-critical band models are *incapable* of predicting even the simplest masking phenomena, this does not mean they cannot be useful in applications. Our aim will be to provide a description of $\epsilon_l(\omega)$ which is consistent for a given bandwidth integration length l.

As an initial step, we define the following intermediary:

Definition 1.1. *Let $N_{[a,b]}$ be the total internal noise excitation power lying in the frequency interval $[a, b]$. Then the* internal noise excitation density *is defined as the function $\rho(z)$ satisfying $N_{[a,b]} = \int_a^b \rho(z)\, dz$, for all $a, b > 0$.*

The main idea is this: ostensibly, the internal noise power spectrum is some fixed power function. Its excitation power spectrum *density*, and not the integrated critical band power, is then the more fundamental entity; obtaining the density allows the concept of internal noise excitation to be "ported" to all excitation models. In particular, we have the relation

$$\epsilon_l(z) = \int_{b_l(z)} \rho(u)\, du \tag{1.9}$$

where $b_l(z)$ is an interval of length l centered on bark frequency z.

The derivation now proceeds with the following prerequisites: (1) data (or an equation) $A(z)$

governing the absolute threshold of hearing, (2) data $|H(z)|^2$ governing the square magnitude response of the middle-ear transfer function, and (3) data $s(z)$ governing the threshold factor for noise-masking-tone scenarios. From these elements, the excitation noise power density ρ is obtained.

Consider the power spectrum of a test sinusoid with frequency z and level just at absolute threshold $A(z)$. The power spectrum of the filtered test sinusoid is $I(z) = A(z)|H(z)|^2$. The integral of $\rho(z)$ over a region in the Bark domain gives the power of the internal noise in that frequency band. According to masking models, the ear integrates along one critical bandwidth, and hence the test sinusoid at bark frequency z is masked by the internal noise in the critical band whose center is at z. The relation between the density and the power of the filtered test tone then must be

$$\int_{z-\frac{1}{2}}^{z+\frac{1}{2}} \rho(u)\, du = \frac{I(z)}{s(z)}, \quad z \geq \frac{1}{2} \tag{1.10}$$

where $s(z)$ is the threshold factor. Equation (1.10) defines an integral equation which can be solved for $\rho(z)$. We now have our first theorem:

Theorem 1.1. *The solution to (1.10) is given by*

$$\rho(u) = \frac{I(\infty)}{s(\infty)} + \sum_{i=0}^{\infty} \frac{s(u+i+\frac{1}{2})I'(u+i+\frac{1}{2}) - I(u+i+\frac{1}{2})s'(u+i+\frac{1}{2})}{s^2(u+i+\frac{1}{2})}, \quad u \geq 0 \tag{1.11}$$

As an example, with the concrete formulae provided in previous sections, the boundary condition reads $\frac{I(\infty)}{s(\infty)} \simeq 4.36$, and $\rho(u)$ is as shown in Figure 1.2. This function can now be used in (1.9) in conjunction with the frequency-bark conversion formula (1.7) to obtain $\epsilon_l(\omega)$ for any l. The reader can prove as an exercise that with $l = 1$, $\epsilon_l(\omega)$ reduces to (1.8). We can also define new perceptual variables $E'(\omega) = E(\omega) + \epsilon_l(\omega)$, with the interpretation that E' represents the *total* physical activity along the basilar membrane, due to both external and internal contributions. The name “total excitation pattern” seems appropriate for E'.

1.1.4 Time-Spreading Models

The use of time-spreading in excitation models is directly analogous to the function of frequency spreading; where the latter models simultaneous inter-frequency masking, the former can account for temporal masking. We will content ourselves with an abstract description of the common features of these models for the sake of completeness.

Let $E'(t, \omega)$ be a set of (total) excitation patterns, indexed by a time parameter t. Then the

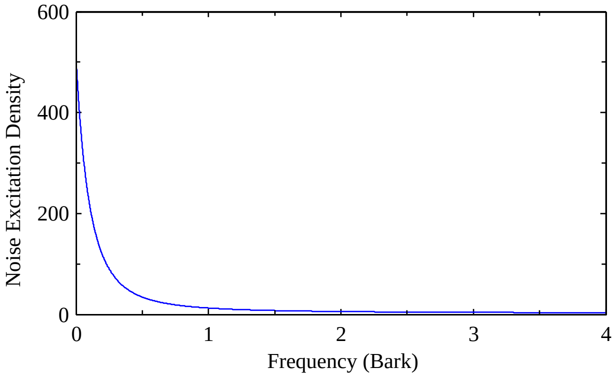


Fig. 1.2 Internal Noise Excitation Density

time-spread excitation pattern can generally be given by:

$$\tilde{E}(t,\omega) = \int_{-\infty}^{\infty} K(\lambda,\omega)E'(t-\lambda,\omega)\,d\lambda. \tag{1.12}$$

The kernel of integration will consist of a type of double-sided decaying exponential, such as:

$$K(\lambda,\omega) = \begin{cases} e^{-k_1\lambda}, & \lambda \geq 0 \\ e^{k_2\lambda}, & \lambda < 0 \end{cases} \tag{1.13}$$

The time-constants of decay $k_i > 0$ are directly related to the durations of forward and backward masking; when thus expect $k_1 > k_2$. Moreover, they may be functions of frequency (longer duration times for lower frequency), whence comes the dependence of K on ω. In some models, such as the time-varying loudness model of [9], the parameters k_i depend upon the non-spread excitation itself — for instance the presence of a signal attack or decay. In this case, the linear convolution above becomes a nonlinear mapping, just as it was with level-dependent auditory filters in (1.2). A detailed discussion of the temporal window shape K as a function of frequency and level is given in [25].

1.2 Discrete Linear Excitation Transformation

The parameter p appearing in (1.3) controls the slopes of the auditory filters. In general, p, (in addition to being dependent on filter center frequency z), will be a function of the overall power level L_z entering the auditory filter: $p = p(z, L_z)$. Moreover, an even more precise modelling [8] assumes asymmetric auditory filters parameterized by upper and lower slopes values p_u and p_l, respectively. However, for moderate sound levels, $p_u \approx p_l \equiv p$. Suppose that we fix p constant for the moderate sound level $L_z = 50$ dB SPL. Then p reduces to a univariate function of frequency, the kernel W in (1.2) is independent of the input power spectrum, and describes a linear operator on the space of power spectra.

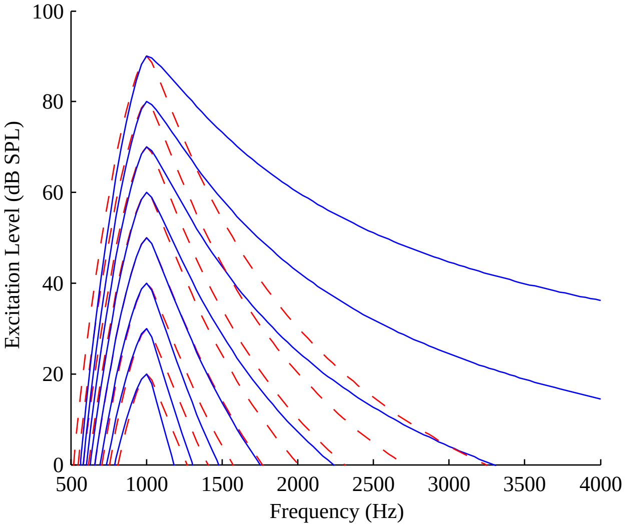


Fig. 1.3 Excitation Pattern for 1 kHz sinsusoid at varying levels as computed by the excitation transform (1.2). Solid: level-dependent transform; dashed: linear transform.

Figure 1.3 shows the predicted excitation pattern of a sinusoid at different power levels for both the full model and the linearised one. While there is a slight dependency on level in the lower slope of the patterns, the main difference lies in the so-called upward spread of masking at high power levels, which cannot, of course, be accounted for by the approximation. The linearised mapping still provides a reasonable replication of the true excitation pattern up to amplitudes in the ambit of 70 dB SPL, however.

What motivations are there for such a linearisation — and the consequent loss of modelling accuracy? Besides the nearly captious observation that we gain in computational complexity of the transform, the main advantage is that the simplification allows for the existence of a simple analytic inverse to W, mapping from excitation to power spectrum. Indeed the idea of excitation-domain coding (Section 3.3.1) is only feasible from a computational point of view when a level-independent transform is used. We shall also see that the viewpoint of excitation as a linear operator on power spectra provides additional insight into the boundaries of the excitation domain, and of so-called "negative" power spectra that may result from quantization points outside the domain.

In what follows it will be convenient to combine the steps of ear-filtering and frequency spreading into one operation, so that the mapping between a non-ear-filtered power spectrum P and its excitation pattern E is given by:

$$E(\Omega) = \int_0^\infty W_P(\omega,\Omega)|H(\omega)|^2 P(\omega)\,d\omega \equiv \int_0^\infty U_P(\omega,\Omega)P(\omega)\,d\omega \tag{1.14}$$

For use in computer algorithms, (1.14) must be frequency-discretized. Typically, one is given a sampled version $\mathbf{p}$ of the continuous power-spectrum P, as calculated, for example, in the form of a N-point DFT. It should be obvious that (1.14) can be approximated with a matrix multiplication. When the kernel U is linearised, the matrix in question is *fixed.* We then seek a matrix $\mathbf{U} : \mathbb{R}^N \to \mathbb{R}^M$ such that $\mathbf{Wp}$ is a sampled version of the continuous excitation pattern $E(\mathbf{p})$.[1]

There exist an infinite number of ways to approximate the integral (1.14) with a summation (Riemann sums, trapezoidal sums etc.), hence an infinite number of possible transformation matrices. Moreover, the matrix will depend upon the location of the desired frequencies, and its dimensions dependent on the number of desired output frequencies (it is clear that one can compute the excitation pattern to any desired resolution, in a directly analagous way that a DFT can compute to arbitrary resolution the spectrum of finite-length sequences).

Let us consider the special case $N = M$. How ought we to select the sampling locations of the excitation pattern? This will depend upon whether we seek a representation in linear frequency, or in Bark frequency. The former implies a uniform sampling in linear frequency, and the latter a uniform sampling in Bark frequency (correspondingly, non-uniform sampling in linear frequency). In practice, the power spectrum will always be delivered by a DFT; hence the sampling of the input is uniform in Hertz frequency. Let us assume then that $\mathbf{p}$ coincides with the spectrum at the frequency points $\omega_k = k\frac{F_s}{2N}$, for $k = 1,\ldots,N$, and where F_s is the (time-domain) sampling frequency.[2] The following easy theorem then provides a sensible choice for the transformation matrix:

Theorem 1.2. *Let $\omega_k = k\frac{F_s}{2N}$, for $k = 1,\ldots,N$. Let $\Omega = g(z)$ be a Bark-Hertz conversion formula. Then a matrix $\mathbf{U}$ mapping from power spectrum to excitation is given by*

1. *(Linear-frequency Excitation Transform)*

$$[\mathbf{U}]_{ij} = \gamma U(\omega_i,\omega_j), \quad 1 \le i,j \le N \tag{1.15}$$

2. *(Bark-frequency Excitation Transform)*

$$[\mathbf{U}]_{ij} = \gamma U(\omega_i,\Omega_j), \quad 1 \le i,j \le N \tag{1.16}$$

[1] *Not* $E(P)$.

[2] We ignore the DC contribution.

where $\Omega_k = g\left(\frac{k}{N} g^{-1}\left(\frac{F_s}{2}\right)\right)$*, and* γ *is an interpolation factor normally set to* $\gamma = \frac{F_s}{2N}$*.*

Remark: The factor γ is used as an interpolation constant from the sampled power spectrum to the continuous spectrum. Sometimes it is useful to change the nominal value of γ, depending upon how much information is known about the spectrum. For instance, if it is known that P is a harmonic complex, with harmonic frequencies coinciding with the sampling points ω_k, then the best selection is $\gamma = 1$. In the absence of other information, $\gamma = \frac{F_s}{2N}$ is a prudent choice.

The Excitation Pattern and Invertibility:

Having defined two concrete square matrices for the excitation transform, we may ask questions about their invertibility — or stronger — their conditioning. Recall that if κ is the condition number of a matrix $\mathbf{A}$, then $\log_{10}\kappa$ gives approximately the number of decimal digits lost in accuracy for the computation $\mathbf{A}^{-1}\mathbf{x}$. In Figure 1.4, we have plotted the condition numbers for the two matrices described in Theorem 1.2, the latter computed using the Bark formula (1.7), and the assumption of sampling frequency $F_s = 8000$ Hz, as a function of matrix dimension N.

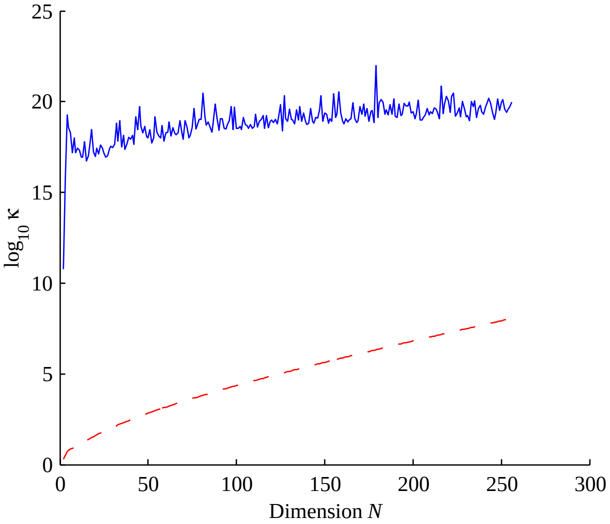


Fig. 1.4 Condition Number as a function of Transform Dimension. Solid: Bark-frequency excitation transform; Dashed: Hertz-frequency excitation transform.

If 16 decimal digits of precision are available — as is standard in 64-bit floating point representations, then we see that the linear excitation matrix is easily numerically invertible at the transform sizes considered, and invertible without the presence of audible artefacts in the power spectrum (one requires about 9 decimal digits of accuracy here), even at high dimensions. The Bark-frequency transform, however, is not invertible with any accuracy for this system — even at very small dimensions. Such ill-conditioning has to do, in fact, with the mismatch between the spacing of the given power spectrum (uniform in Hertz), and the spacing of the excitation frequencies (non-uniform in Hertz). Since the auditory filters of (1.4) depend directly on a term of the form $e^{-k|\omega-\Omega|}$, we see that in the linear-frequency transform, most entries of the matrix will tend to be near zero, except along the diagonal, where $\Omega_k = \omega_k$, because of the coincidence of excitation and power spectral frequencies, maximizing the entry. We thus obtain a large variation matrix value moving along any row/column.

For the Bark-frequency transform, on the other hand, precisely because of non-coincidence of frequencies Ω_k with ω_k, the exponential will tend to be rather small at every entry — no maximization of the exponential occurs, the matrix variation moving along any row/column is small, leading to poor conditioning.

Nevertheless, the fact remains that both transforms *are* mathematically invertible — a fact that will not surprise any psychoacoustician — but may perhaps surprise a few audio engineers, given that most well-known excitation models for coders are non-invertible; for instance Johnston's [15], the MPEG models, and PEAQ. The misunderstanding may have to do with the meaning of the term "critical-band integration" — in each of the cited models, frequency integration is performed on *disjoint* sets of critical (or sub-critical) bands. It did not help, of course, that Zwicker's influential book [37] produced a table which seemed to partition the frequency line into disjoint bands. This has led to the idea, seemingly widespread, that patterns such as excitation and loudness represent only coarse approximations to the power spectrum in terms of frequency resolution, with the transformation entailing much loss of "fine structure". The use of non-invertible models have lead to results like that of [13], in which a CELP coder produced "whispered speech" when the standard square-error criterion was replaced with a loudness difference minimization algorithm. The resulting spectrograms showed speech where the pitch had been entirely lost, due to the coarseness of the loudness pattern.

The actuality is that the excitation pattern, as well as loudness pattern, both contain precisely the same amount of information as the power spectrum. This is only true, however, when one allows the bands of integration to overlap, as is manifestly plain in (1.2), where every band of integration, no matter the center frequency, is the infinite interval $[0, \infty)$.

1.3 Complex Excitation Patterns

One complaint concerning the power-spectrum derived excitation pattern is its insensitivity to phase. It clearly then cannot be a complete characterization of an internal pattern, since the ear is capable of detecting relative phase distortion [37].

In this section we define and study the properties of a representation which is derived from a transform on the *Fourier* spectrum — as opposed to the power spectrum. The pattern will turn out to possess some strong similarities to the true excitation pattern, as well as some marked differences which allow a certain type of phase information to be preserved.

When computing the excitation, say, according to (1.2), a rectification is first performed on the signal spectrum, followed by the frequency spreading. Suppose, on the contrary that these two processes are reversed: that rectification (squaring) occurs after spreading a now *complex* spectrum. The resulting signal is still strictly positive, possessing the units of power — this we shall define as the *Complex Excitation Power Spectrum*, or CEPS for brevity. More formally,

Definition 1.2. *The CEPS of a Fourier spectrum $X(\omega)$ is defined as*

$$C(\Omega) = \left| \int_0^\infty U^{1/2}(\omega,\Omega) X(\omega)\, d\omega \right|^2 \tag{1.17}$$

where U is the kernel of (1.14).

The square root on the spreading kernel is chosen to match the units of amplitude on X. It will be convenient to work, for the remainder of this section, with its discrete version:

$$\mathbf{c} = \left| \mathbf{U}^{1/2} \mathbf{x} \right|^2 \tag{1.18}$$

in obvious notation.

In applying the spreading operator to complex signals — and not those restricted to be positive, we have performed a mathematical abstraction, with the result that the new pattern has no longer the physical meaning accorded to the excitation. The CEPS has no analogue in psychoacoustics — it may represent the activity of hair cells along the basilar membrane, in some sense or another, but probably not accurately as the true excitation pattern. Following Nietzsche, however, one ought not to confuse truth with utility. The CEPS does capture information not included in the average activity of hair cells — information encoded in a different way than magnitude. It thus may prove to be the better internal representation, in a global sense. Moreover, according to Moore in [19], the proper order of processing in the auditory system is first filtering, followed by rectification. This statement was made with reference to time-domain auditory filtering, and not a complex-valued filtering — the analogy is nonetheless suggestive.

1.3.1 Properties of the complex excitation power spectrum (CEPS)

Here we analyse a few relationships and differences between the complex excitation power spectrum and the standard excitation pattern. We will assume, for the sake of making some of the relationships clearer, that the excitation transformation matrix $\mathbf{U}$ is *fixed*, thus dealing only with level-independent operators.

Notation is fixed as follows. In the Discrete Fourier domain, we use the basis vectors $\mathbf{s}_k \equiv [0, 0, \ldots, 0, 1, 0, \ldots]^T$, where 1 appears in the k-th position. A sinusoid (complex exponential) at discrete frequency k is then written as $\lambda_k \mathbf{s}_k$ for some complex number $\lambda_k = A_k e^{j\theta_k}$, where A_k is its amplitude and θ_k its phase. A general DFT signal will be written $\mathbf{x} = \sum_k \lambda_k \mathbf{s}_k$, and its power spectrum $\mathbf{p} = \sum_k |\lambda_k|^2 \mathbf{s}_k$. We will denote by $\mathbf{u}_k$ the k-th column from the auditory filter matrix $\mathbf{U}$. It also represents the excitation pattern for a sinusoid in the k-th frequency bin. Finally, operations such as exponentiation and $|\cdot|$ on vectors are defined pointwise.

CEPS for one sinusoid

The complex excitation pattern for a single sinusoid is given, from (1.17), by:

$$\mathbf{c} = \left|\mathbf{U}^{1/2} \lambda_k \mathbf{s}_k\right|^2 \tag{1.19}$$

$$= |\lambda_k|^2 \mathbf{u}_k \tag{1.20}$$

Now compare this with the standard excitation pattern for a single sinusoid, with q-power law addition (discretized from (1.4)):

$$\mathbf{e} = \left(\mathbf{U}^q (|\lambda_k|^2 \mathbf{s}_k)^q\right)^{1/q} = |\lambda_k|^2 \mathbf{u}_k \tag{1.21}$$

Equations (1.20) and (1.21) agree, leading to the property that *the CEPS for a sinusoid of arbitrary magnitude and phase is the same as its standard excitation pattern, in any power-domain of spreading.*

CEPS for two sinusoids

The situation for two sinusoids is much more interesting. First consider the case where the two sinusoids have the same phase. We write this signal as:

$$\mathbf{x} = \lambda_{k_1} \mathbf{s}_{k_1} + \lambda_{k_2} \mathbf{s}_{k_2} \tag{1.22}$$

$$= A_{k_1} e^{j\theta} \mathbf{s}_{k_1} + A_{k_2} e^{j\theta} \mathbf{s}_{k_2} \tag{1.23}$$

The CEPS is given by

$$\mathbf{c} = \left|\mathbf{U}^{1/2} (A_{k_1} e^{j\theta} \mathbf{s}_{k_1} + A_{k_2} e^{j\theta} \mathbf{s}_{k_2})\right|^2 \tag{1.24}$$

$$= (A_{k_1} \mathbf{u}_{k_1}^{1/2} + A_{k_2} \mathbf{u}_{k_2}^{1/2})^2 \tag{1.25}$$

The last line is in fact the standard excitation pattern for two sinusoids when 0.5-power law spreading is used, leading to the fact that *the CEPS of two sinusoids with the same phase and arbitrary magnitudes is equivalent to the standard excitation pattern of the two sinusoids obtained via 0.5 power-law domain spreading.*

Now consider two sinusoids that have different phase. We can derive an explicit expression for the CEPS. Supposing we have sinusoids of amplitudes A_1 and A_2 with corresponding phases θ_1 and θ_2, it is easy to show the following:

$$\mathbf{c} = A_1^2 \mathbf{u}_{k_1} + A_2^2 \mathbf{u}_{k_2} + 2A_1 A_2 \sqrt{\mathbf{u}_{k_1}} \sqrt{\mathbf{u}_{k_2}} \cos(\theta_1 - \theta_2). \tag{1.26}$$

The above expression makes explicit the dependence of the CEPS pattern on the relative phase $\theta_1 - \theta_2$. When $\theta_1 = \theta_2$, the pattern attains its maximum value, and is equivalent to the standard excitation pattern. The CEPS then differs from the standard pattern by the introduction of a multiplicative cosine factor in the cross-term. One can see this effect in Figure 1.5.

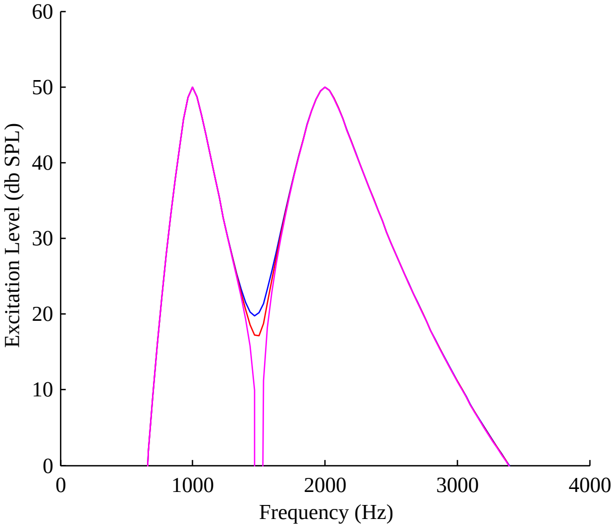


Fig. 1.5 Excitation Patterns and CEPS patterns for two sinusoids of equal magnitude under different phase conditions. Top: Standard excitation pattern = CEPS pattern, $\theta_1 - \theta_2 = 0$; Middle: CEPS pattern, $\theta_1 - \theta_2 = \pi/2$; Bottom: CEPS Pattern, $\theta_1 - \theta_2 = \pi$.

As phase difference increases, the local minimum decreases. Generally, phase malignment reduces the CEPS level relative to the standard excitation pattern. This is due to destructive interference caused by the phase difference. Equality is achieved when the two sinusoids are in phase. The patterns at the main sinusoidal frequencies (maxima) are unaffected by the phase, because the component of each sinusoid at the other frequency, via spreading, is very small. Thus relative phase information is captured at frequencies somewhere in between the sinusoids.

The final observation is that the cross-term (and thus the phase effects), are most visible at the frequencies (indices) where the excitation product $\sqrt{\mathbf{u}_{k1}}\sqrt{\mathbf{u}_{k2}}$ is largest. If these patterns are approximately symmetric, then this location occurs approximately midway between the two sinusoidal frequencies.

Some of the above observations may be easily generalised to any spectrum. We indeed have the following:

Theorem 1.3. *The CEPS pattern $C(\omega)$ derived from spectrum $X(\omega)$ satisfies the following:*

1. *$C(\omega) \leq E(\omega)$, where $E(\omega)$ is the standard excitation pattern for $|X(\omega)|^2$ in the 0.5-power law domain. Equality is achieved if the phase of $X(\omega)$ is constant.*

2. *(Rotation-invariance). $C(\omega)$ is invariant under a constant phase multiplication to $X(\omega)$. It is sensitive only to phase difference, not absolute phase.*

A consequence of the second property is that any distortion measure composed between the CEPS patterns will have the phase-factor-invariant property: $D(e^{i\theta}\mathbf{c}, \hat{\mathbf{c}}) = D(\mathbf{c}, \hat{\mathbf{c}})$. This would not be true, for instance, of a simple squared error between complex DFT spectra $D(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$.

Finally, we observe that, given the identity of CEPS to standard excitation in the case of constant phase spectrum, it appears reasonable to apply the same models for the internal noise excitation pattern (Section 1.1.3), and time-spreading (Section 1.1.4), to the CEPS. Thus, we define the *total complex excitation power spectrum* as the variable $C'(\omega) = C(\omega) + \epsilon(\omega)$, and the *time-spread complex excitation power spectrum* induced from a time-indexed family of CEPS patterns $C'(t, \omega)$ by:

$$\tilde{C}(t, \omega) = \int_{-\infty}^{\infty} K(\lambda, \omega) C'(t - \lambda, \omega) \, d\lambda \tag{1.27}$$

Chapter 2

Distortion Measures and Excitation Distortion

It is difficult to give mathematical form to the term “auditory distortion”. Variability from subject to subject, an insufficient grasp of higher-level brain processes, and the immense range of possible contexts and stimuli combine to complicate the study. Moreover, auditory difference is certainly a multi-dimensional object. The distortion incurred by corruptive white noise is qualitatively distinct from the distortion caused by frequency shifts in a harmonic complex, which is again quite different from distortion due to phase discontinuities in a sinusoid.

Nonetheless, a great deal of energy has been devoted to finding accurate models of auditory difference (in one sense or another), because significant gains, with respect to the traditional warhorses of distortion: mean-square error and signal-to-noise ratio, have been observed from the application of perceptual theory.

Perhaps the most prominent use of this theory occurs in modern audio coders. Other applications include the objective measurement of audio quality, phoneme discrimination, and residual selection algorithms in speech coders. All of the above examples make use of distortion functions rooted on perceptual features. For example, audio coders use the noise-to-mask ratio in the bit assignment for noise control; objective quality measurements typically employ some type of difference in internal patterns (e.g. excitation or loudness) to predict mean opinion scores; the partial loudness of an error difference can be used as a basis for vowel discrimination.

This chapter considers the construction of general (principally magnitude) distortion functions for perceptual coding. Our focus will be somewhat one-sided, in that we shall not consider the constraints put on us by a coder, or the implementation of these distortion functions in a coder (for that, see the next Chapter). Rather, we aim to generate a cornucopia of psychoacoustically sound distortion functions which, although initially arguing from the coding viewpoint, should prove useful for any number applications, including the ones enumerated in the previous paragraph.

We first suggest that conventional approaches which use masking results are inappropriate. Rather, we propose supra-threshold distortion measures which, in the appropriate limits, reduce to masking results. As with objective evaluators of audio quality, the measures consist of differences in internal representations, functions quantifying activity along the place dimension of the cochlear membrane.

Even once internal distributions are chosen as the fundamental variables, an appropriate metric still must be selected. It might be argued that the exact mathematical form of the distortion function is a matter falling to experimental science. While this is unreservedly true to some extent, we do posit the existence of certain criteria — consistency relations — which constrain the infinitude of possible distance measures considerably. It will in fact turn out that these constraints rule out the use of many distortion functions that have been suggested in the literature.

Much of psychoacoustics has been concerned with threshold results: masking, absolute threshold, and just-noticeable differences, to name a few, are among the most well-verified and well-examined auditory phenomena. On the other hand, there is a relative paucity of supra-threshold results. It seems natural, then, to design a framework which constrains the space of perceptual distortion functions to those which, at least, model thresholds correctly.

The constraint comes in the form of *gain invariance*, which is introduced and defined. We derive the class of all gain-invariant distortion functions; these functions are the only ones consistent with the psychoacoustics literature, in the sense that masking, and more broadly — threshold phenomena — are subsumed. Moreover, Zwicker's classical 1 dB difference limen (in excitation) can be generalised as a limen on a gain-invariant function. While each member of this family of functions gives identical and consistent threshold results, the predictions for supra-threshold distortion vary, affording extra degrees of modelling freedom. The new class of measures includes the classical logarithmic distortion function as a special case. Our main contribution is thus the introduction of a natural generalization to the first of all excitation-distortion measures: the "dB-distance" between excitation patterns.

2.1 Mis-applications of Psychoacoustics: A critique

In order to motivate the use of general auditory difference functions, we provide a sampling of applications where context-specific psychoacoustics have been applied, not always entirely appropriately, to engineering applications.

Noise-Shaping for Audio Coding

Johnston, in [15], is generally credited with the introduction of psychoacoustic models for improved audio coding. His method, subsequently generalised, has become the basis for a number of modern approaches to this problem. The idea is to compute the masking threshold $M_x(\omega)$ from the short-term power spectrum of an audio signal x. Assuming that the quantizer introduces distortion that is modelled as additive noise, i.e. $\tilde{x}(t) = x(t) + n(t)$, the noise spectrum $|N(\omega)|^2$ is shaped to lie beneath the threshold $M_x(\omega)$. Put another way, the coder searches for an allocation of bits minimizing the distortion function

$$D = \frac{|N(\omega)|^2}{M_x(\omega)} \tag{2.1}$$

The above criterion is referred to as the noise-to-mask ratio (NMR).

The traditional arguments concerning the unsuitability of the derived masking thresholds, as can be found enumerated in [34], are that (1) the thresholds in psychoacoustics typically consider only tonal targets, whereas coding noise is broadband, (2) the effect of multiple maskers is uncertain with respect to masking addition, and (3) the effect of masking multiple targets is uncertain.

The above objections are criticisms in detail as opposed to principle: they concern the divide between the simplicity of stimuli in most psychoacoustic experiments and the relative complexity of the audio coding reality. We find the concept of masked quantization noise, itself, problematic, based on the following observation: modelling coding distortions, particularly large ones, as additive uncorrelated noise is inconsistent with the decoded signal. More precisely, quantization noise is in general *not* a power-additive distortion. Indeed, for optimal minimum mean-square error quantization, the average energy of the reconstructed signal $\hat{x}$ is always smaller than the original signal x, the difference being the quantization error e,

$$\sigma_{\hat{x}}^2 = \sigma_x^2 - \sigma_e^2. \tag{2.2}$$

This result for optimal scalar quantizers attains the following form for vector quantizers, with R denoting covariance matrices for the respective variables:

$$R_{\hat{x}} = R_x - R_e \tag{2.3}$$

In particular, the constraint on the diagonal elements demonstrates that the result also holds component-by-component in the multivariate case. More generally, regardless of the particular structure of the quantizers used, coding distortion will not be power-additive, particularly at low rates where many spectral regions are coded as zero.

Note that the signal that is conventionally considered to be the maskee is actually subtractive. This is in obvious contradiction to masking experiments, where the listener is presented with a sum of two signals, the power of which is larger than the power of any one signal presented alone. In particular, all four of the tone-noise masker-maskee combinations from which modern audio coders derive their psychoacoustic rules utilise a masker signal which is *less* in power than the sum of masker and target together. The concept of signal masker and noise maskee is thus unsuited for describing the perceptual effects of quantization, except perhaps, and even then only approximately, in the low distortion regime.

We can reinforce the point by extremizing our thought-experiment. In low rate coding, many spectral regions are reproduced as zero. In this case, the distortion is equal to the signal itself, and the conventional masking argument asserts that the original signal masks a noise equivalent to itself. Of course, there is no masking because no sound reaches the ear! While an extreme position, this scenario does occur all the time in practical audio coding: rate savings can be obtained at the lowest rates only by zeroing transform coefficients. And indeed, one of the oldest methods of transform coding is precisely to zero out all coefficients below some threshold; nor is such a procedure antiquated — take threshold = masking threshold, for example. Thinking along such lines can lead to amusing paradoxes. For instance, by one reckoning, on average 50% of transform coefficients lie under the masking threshold of a frame of audio, as predicted by standard models [21]. By zeroing out these coefficients, a coder violates the very assumptions which make such a calculation permissible in the first place.

To be sure, it is of course possible to allow non power-additive distortion (such as quantization noise) to be made sufficiently small so that its addition remains undetectable. But such a phenomenon is no longer masking — it falls under the more general phenomena of just-noticeable differences, namely that of amplitude JNDs (of which masking is a special case [37]). And it is precisely those models of just-noticeable difference that we shall turn to when constructing the new distortion measures.

Masking Weights for CELP Coding

In code-excited linear-predictive speech coding, a finite-order all-pole filter is used to model the vocal tract; the output of this filter is further processed with a pitch predictor to obtain a residual. This residual is then quantized, usually with a codebook containing the reproduction points of a vector quantizer.

The codevector search can be driven by minimizing the standard square error $e_i^2 = (x_i - \tilde{x}_i)^2$, or more typically, a perceptually weighted criterion such as:

$$D = \sum_i \frac{e_i^2}{M_i} \tag{2.4}$$

where $\{M_i\}$ is the masking threshold for signal x. Such a scenario occurs, *mutatis mutandis*, in the work of [30], for example.

The objection levelled in the first example applies here as well: the error incurred in the residual quantization is not generally power-additive, and the use of masking thresholds obtained from experiments involving uncorrelated noise maskees can be misleading.

Partial Loudness for Vowel Discrimination

The work of Rao *et al* in [27] investigated the application of Moore's partial loudness model [20] as a metric for estimating perceptual effects of modifications to the spectral envelope of a harmonic sound. Partial loudness should be viewed as an outcome of partial masking — in other words, a measure of supra-threshold activity. As a distortion function, it can be used in the following way: the error $e = x - \tilde{x}$ is defined as the target in question, the background sound is equated with the original signal x, and the perceptual distortion is given by the partial loudness $L_p(e, x)$ of the target e submerged in background x.

A partial loudness model can only be applied under the assumptions with which it was formulated, however: under the conditions where target and background are power additive. Indeed, the model requires the computation of three different excitation patterns: that of the total sound, that of the background sound and that of the noise target, the last of which implicitly assumes an additive stimulus. The model cannot account for arbitrary — in particular, subtractive— distortions.

The approach to this issue taken in [27] is to redefine the total and background sound excitation patterns via $\max\{E_x, E_{\tilde{x}}\}$ and $\min\{E_x, E_{\tilde{x}}\}$ respectively. The redefinition forces the difference (corresponding to the target excitation pattern) to be positive — but at the loss of physical interpretation with respect to the masker and maskee. The artifice shows all the more plainly the conundrum that is faced when dealing with non-additive spectral distortion.

The foregoing examples serve to illustrate a recurring theme: the ubiquitous decomposition of the distortion problem as an original, uncorrupted signal in the presence of additive noise, and the search for measures to quantify the loudness, or detectability of this disturbance even when it is a fictitious entity not physically present at the receiver. In the first two examples, masked coding noise notwithstanding, the charge that equations (2.1) and (2.4) are only useful for quantitating at-threshold behavior can also be levelled. The masking threshold does not readily lend itself to measuring the amount of audible distortion (supra-masking threshold), and hence this distortion cannot be properly minimized.

2.2 General Auditory Distortion Functions

We now turn our attention to the construction of more general perceptual difference functions not requiring the assumption of spectrum-additive distortion. In what follows, we will take an abstract view of the excitation computation process, making use of the generic concepts without relying on specific formulae for their computation. The reader, if so desired, may make the translation of the term "excitation pattern" into either the term "total excitation pattern", or "time-spread excitation pattern", via any one of the concrete formula presented in the first chapter — even those of the complex excitation power spectrum. Thus, the functions derived can be tailored to meet a range of data for acoustical quantities such as absolute threshold of hearing, masking offsets, critical bandwidths and the like. Only in Section 2.3 do we make use of specific methods — by way of example — to give some feeling for model predictions.

2.2.1 Distortion Functions on Internal Representations

Excitation patterns are examples of so-called "internal representations". The specific loudness pattern, obtained by a pointwise nonlinear mapping on the excitation pattern, is another internal representation. Internal representations are functions lying in a perceptual space, and, ostensibly, distance metrics on such spaces should provide superior modelling with respect to subjective distortion than traditional metrics in the time or power-spectral domain. At the very least, they have been used to formulate models of just-noticeable differences. Zwicker's criterion [37] states that two excitation patterns $x(\omega)$ and $y(\omega)$ are indistinguishable if they differ everywhere by less than 1 dB. In symbols:

$$|10\log_{10} x(\omega) - 10\log_{10} y(\omega)| < 1, \quad \forall\omega \tag{2.5}$$

or equivalently

$$D(x;y) = \max_{\omega} |10\log_{10} x(\omega) - 10\log_{10} y(\omega)| < 1. \tag{2.6}$$

The threshold value of 1 dB has been debated, with some estimates as low as 0.1 dB [19]; the variable may also depend upon frequency, as well as context. In spite of these limitations, the criterion is still an extremely useful conceptual tool, providing a generalised way of determining perceptual equivalence.

Equation (2.6) not only defines the just-noticeable difference between two signals, but implicitly defines a distortion function $D(x;y)$ for arbitrary excitations x and y. It is thus a more general entity than the noise-mask ratio of (2.1), applying to situations even when the distortion is not power-spectrum additive.

Given two excitation distributions, (2.6) is not the only way to quantify their difference. We assert the existence of certain constraints, however, governing the form of any plausible measure. These properties are investigated in the following section.

2.2.2 Gain-invariant Distortion Functions

For simplicity of discourse, we present the theory in terms of discrete distributions. In any event, one is always presented with a discrete excitation pattern in practice. It is, however, straightforward to extend all results to the continuous case.

Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be two excitation patterns well above absolute threshold. We begin by noticing that the difference limen of (2.6) is invariant under the replacement $x \to (a_1 x_1, \ldots, a_n x_n)$, $y \to (a_1 y_1, \ldots, a_n y_n)$, for any choices of non-zero constants $a_1, \ldots, a_n$. Since the replacement corresponds to a gain transformation on both excitation patterns, we say that the distortion function $D(x; y)$ is *invariant under the group of gain transformations.* Formally:

Definition 2.1. *A distortion function $D(x; y)$ is gain-invariant (with respect to the arguments x and y), if for all non-zero $a_1, \ldots, a_n \in \mathbb{R}$,*

$$D(a_1 x_1, \ldots, a_n x_n; a_1 y_1, \ldots, a_n y_n) = D(x_1, \ldots, x_n; y_1, \ldots, y_n)$$

The following thought experiment reinforces the physical significance of this property: Let x and y be two excitation patterns within 1 dB of another. They are, by Zwicker's criterion, perceptually indistinguishable. Now if the gain transformation $T(u) = (a_1 u_1, \ldots, a_n u_n)$ is applied to both patterns, the resulting patterns $T(x)$ and $T(y)$ remain within 1 dB of one another. Hence the patterns $T(x)$ and $T(y)$ are indistinguishable. More generally, the invariance states that perceptual differences between signals do not change under gain transformations to the excitation patterns — as long as the same transformation is applied to both patterns.

Note that the criterion states nothing about the perceptual difference between x and $T(x)$ — only between $T(x)$ and $T(y)$. Since gain-invariance is a natural characteristic of the just-noticeable limen, any perceptual distortion function must possess this property to be able to predict threshold phenomena. We will take it as a fundamental hypothesis.

The next observation to make concerning the criterion of (2.6) is that it is a function of the *pointwise ratio* between its two arguments. We make the following definition:

Definition 2.2. *A distortion $D(x; y)$ is a pointwise function of the ratio of its arguments if there* *exist functions F and $D_1, \dots, D_n$ such that*

$$D(x; y) = F(D_1(x_1/y_1), \dots, D_n(x_n/y_n)) \tag{2.7}$$

The representation introduced in the definition, though not necessary, is a very useful structural decomposition. It consists of a set of local distortions $D_1, \dots, D_n$ defining a distortion pattern $\mathcal{D} = (D_1, \dots, D_n)$, and a grouping function $F(\cdot)$, which computes some overall distortion by, for example, integrating over the distortion pattern. The function F should satisfy some constraints: for instance that it be increasing in each dimension. Norms make very natural choices for F.

It is obvious that every distortion function satisfying Definition 2.7 is gain-invariant. Not so obvious is the converse:

Theorem 2.1. *If a distortion function $D(x; y)$ is gain-invariant, then it is a pointwise function of the ratio of its arguments.*

The constraint of gain-invariance is very strong, and allows one to eliminate seemingly reasonable choices for a distortion function. For example, both the correlation coefficient between two excitation patterns

$$\rho(x; y) = \frac{\sum_i x_i y_i}{(\sum_i x_i^2)^{1/2} (\sum_i y_i^2)^{1/2}} \tag{2.8}$$

and the Minkowski distance

$$D_m(x; y) = \Big(\sum_i (x_i - y_i)^p \Big)^{1/p} \tag{2.9}$$

are not gain-invariant distortions.

There is another type of invariance which is sometimes desired of a distortion function, that of an invariance with respect to the overall scales of the patterns. We will call this *scale-invariance* (also called gain-optimized in [10]); in symbols:

Definition 2.3. *A distortion function $D(x; y)$ is scale-invariant if for all non-zero $a, b \in \mathbb{R}$*

$$D(ax_1, \dots, ax_n; by_1, \dots, by_n) = D(x_1, \dots, x_n; y_1, \dots, y_n)$$

Distortion measures satisfying such a condition are immune to scaling (loudness) transformations of either of the input patterns; two signals that differ only in their loudness are considered identical. Such a property is often desired of objective evaluators of audio quality. If the distortion function is symmetric ($D(x; y) = D(y; x)$), one method of imposing scale-invariance is to define

$$D'(x; y) = \inf_{\beta > 0} D(\beta x_1, \dots, \beta x_n; y). \tag{2.10}$$

It is easy to see that $D'(x;y)$ is scale-invariant. Generally, the algorithmic implementation of a scale-invariant measure will be complicated by the fact that there rarely exist closed-form expressions for the minimization (2.10). Nevertheless, Appendix A discusses and derives a few such formulae in the cases where analysis is tractable.

Table 2.1 summarises the possibilities and gives examples for different combinations of constraints.

Table 2.1 Distortion functions satisfying various invariances

Type	Example
Gain-invariant only	$\max_i \lvert \log_{10}(x_i/y_i) \rvert$
Scale-invariant only	$\dfrac{\sum_i x_i y_i}{(\sum_i x_i^2)^{1/2}(\sum_i y_i^2)^{1/2}}$
Gain-invariant and Scale-invariant	$\min_{\beta>0} \max_i \lvert \log_{10}(\beta x_i/y_i) \rvert$
Neither	$(\sum_i (x_i - y_i)^p)^{1/p}$

2.2.3 Additional Constraints

Returning to the decomposition of (2.7), it is reasonable to impose the following pseudo-metric constraints on the local distortion functions D_i:

1. (Positivity) $D_i(x_i/y_i) \geq 0$ with equality iff $y_i = x_i$
2. (Symmetry) $D_i(x_i/y_i) = D_i(y_i/x_i)$

Note that a symmetric distortion function can still predict masking asymmetry (see Section 2.3.2). The reader is referred to Appendix B for a more detailed elucidation of the various types of symmetry properties that distortion measures may satisfy.

Assuming some type of "frequency equalization" has occurred in the transformation from power spectrum to excitation, a good first approximation is to take $D_1 = \cdots = D_n$. The positivity assumption implies $D_i(x_i/y_i) = |f(x_i/y_i)|$ for suitable f. Comparison of this form with the excitation distortion of (2.6) — which is itself reminiscent of a loudness difference — suggests that generally, f is an increasing compressive nonlinearity. The foregoing can be summarised with the following hypotheses:

1. D is gain-invariant, with
   $D(x;y) = F(|f(x_1/y_1)|, \ldots, |f(x_n/y_n)|)$
2. f is concave and increasing

3. $f(1) = 0$

4. $f(r^{-1}) = -f(r)$

As for the grouping function, taking inspiration from loudness and partial loudness models, where integration of the patterns is assumed, suggests the Minkowski grouping functions:

$$D(x; y) = \Big(\sum_i |f(x_i/y_i)|^p \Big)^{1/p}. \tag{2.11}$$

There is some recent evidence that the ear has the ability to integrate distortion across auditory filters [24] — in this case, the Minkowski functions model a range of such integration, from the global ($p = 0$), to the local ($p = \infty$). For classical predictions of threshold [37], as well as comparisons with masking models, which assume detection at the location of maximum distortion, we shall use $p = \infty$. Interestingly, the case $p = 2$ and $f = \log(\cdot)$ is the excitation-domain analogue of the well-known spectral distortion measure [23] in speech coding.

2.2.4 Generalized Thresholds of Discrimination

Having established the class of gain-invariant distortion functions with respect to excitation, it becomes straightforward to transfer the classical models of just-noticeable difference to thresholds on the new functions. Assume that Zwicker's just-noticeable criterion holds at $T > 0$ dB. Then we have the following:

Theorem 2.2. *Let $D(x; y) = \max_i(f(x_i/y_i))$ be a distortion function satisfying the hypotheses of Section 2.2.3. Then two excitation patterns x and y are perceptually equivalent if $D(x; y) < f(10^{T/10})$.*

It will be also convenient to define normalised versions of the distortion functions

$$D_f(x; y) = \frac{1}{f(10^{T/10})} \max_i |f(x_i/y_i)| \tag{2.12}$$

In this case, two excitation patterns x and y are imperceptibility different if $D_f(x; y) < 1$. Since the overall scale of the distortion function is irrelevant for the purposes of its minimization, the normalised forms define natural equivalence classes. Comparisons between two gain-invariant distortion measures will always be done between the normalised versions.

From the generalised threshold, we can define the set of frequencies for which there is *audible distortion*, via

$$S_{\text{aud}} = \{i : |f(x_i/y_i)| > f(10^{T/10})\} \tag{2.13}$$

A new distortion function quantifying the audible distortion can then be prescribed:

$$D_{\text{aud}}(x;y) = \Big(\sum_{S_{\text{aud}}} |f(x_i/y_i)|^p\Big)^{1/p} \tag{2.14}$$

This measure is used in Section 2.3 to predict supra-threshold distortion.

2.2.5 Distortion Functions on Loudness Patterns

A loudness distribution models the further nonlinear transformation of intensity to perceptive strength—a type of neural excitation. The level-transformed excitation pattern is termed the specific loudness pattern, and gives loudness as a function of (tonal) frequency. Steven's law, the general hypothesis that perceived strength is a power function of physical magnitude, predicts $L_x = cx^k$, where L_x is the loudness pattern of excitation x, for some constant $k < 1$. This expression is generally accurate for levels above threshold, but requires modification for low-level stimuli.

Since a specific loudness pattern accounts for both the non-ideal frequency selectivity and the nonlinear level compression of the ear, it would appear that distortion functions should be constructed in the loudness domain, as opposed to the excitation domain. Actually, the formulations are equivalent above threshold—with the constraint of gain-invariance—as the following shows.

Loudness is in general an invertible function of excitation: $L_x(i) = g(x_i)$, for some compressive nonlinearity g, so that any distortion function on excitation also induces a distortion function on loudness densities:

$$D(x;y) = D[g^{-1}(L_x(i)); g^{-1}(L_y(i))] \equiv D'(L_x; L_y) \tag{2.15}$$

Similarly, a distortion function $D(L_x; L_y)$ defined on the loudness space induces a distortion function on the excitation patterns, given by $D'(x;y) = D(g(x_i); g(y_i))$. The following now holds:

Theorem 2.3. *If Steven's law holds between excitation and loudness, then a distortion function $D(x;y)$ is gain-invariant with respect to excitation if and only if the induced distortion $D'(L_x; L_y)$ is gain-invariant with respect to loudness.*

Thus Steven's law provides a sufficient condition for a gain-invariant distortion function on excitation patterns to be gain-invariant with respect to loudness, and vice-versa. It follows that there is no advantage to using loudness patterns over excitation patterns, above threshold. Near threshold, there is in fact a disadvantage to using loudness patterns, since by definition, a signal below absolute threshold has zero loudness[1]. In this scenario, indeterminate forms appear in the

[1] Indeed, models of loudness make use of the internal noise excitation ϵ to obtain this result.

ratio L_x/L_y of a gain-invariant function. Distortion functions on the total excitation patterns (Section 1.1.3) thus provide the most convenient way of mediating between the two regions.

2.2.6 Design Example: Loudness Difference and Gain-Invariance

Let $L_x(\omega)$ and $L_y(\omega)$ be the loudness densities for the excitation patterns x and y (we revert to continuous frequency variables for notational convenience in this section). One can define a distortion function by

$$D(L_x; L_y) = \Big(\int |L_x(\omega) - L_y(\omega)|^p \, d\omega \Big)^{1/p} \tag{2.16}$$

This loudness difference has been popular in a number of applications, including: vowel discrimination [3], objective evaluation of speech quality [35], perceptual audio quality measure (PAQM) [2], and audio coding [4]. In particular, the results of [35] and [2] show that (2.16) correlates well with mean opinion scores.

The measure (2.16) directly utilises internal patterns, without recourse to a fictitious "noise". It is, however, not gain-invariant, and hence not consistent with limen results. Nonetheless, given its relative experimental success, it may be used as the basis for constructing a class of gain-invariant functions. More specifically, we can pose the following question: what distortion function satisfying the hypotheses of Section 2.2.3 approximates a loudness difference with minimum error? Since both the desired distortion function and (2.16) are pointwise, it suffices to minimise the error between the local distortion functions $f(x(\omega)/y(\omega))$ and $L_x(\omega) - L_y(\omega)$ for fixed ω, over some region in the (x, y) excitation plane. For levels above threshold, $L_x - L_y$ is well approximated by $c(x^k - y^k)$, and we can thus formulate the problem mathematically as follows:

Find f such that

$$\int_0^U \int_0^U \left[f\left(\frac{x}{y}\right) - c(x^k - y^k) \right]^2 dxdy \quad \text{is minimized} \tag{2.17}$$

The optimization problem thus posed is infinite-dimensional, since the unknown variables form a continuous set. The solution attains a surprisingly simple structure:

Theorem 2.4. *The solution to (2.17) is*

$$f\left(\frac{x}{y}\right) = \begin{cases} \left(1 - \left(\frac{y}{x}\right)^k\right) \frac{U^k}{c(k+1)}, & x > y \\ -\left(1 - \left(\frac{x}{y}\right)^k\right) \frac{U^k}{c(k+1)}, & x < y \end{cases} \tag{2.18}$$

The normalised form $\hat{f}$ is independent of the region of optimization, given by

$$\hat{f}\left(\frac{x}{y}\right) = \begin{cases} \left(1 - \left(\frac{y}{x}\right)^k\right) \dfrac{1}{1 - 10^{-Tk/10}}, & x > y \\ -\left(1 - \left(\frac{x}{y}\right)^k\right) \dfrac{1}{1 - 10^{-Tk/10}}, & x < y \end{cases} \tag{2.19}$$

It is easily verified that $\hat{f}$ satisfies the hypotheses of Section 2.2.3; i.e. $\hat{f}$ is concave and increasing, $\hat{f}(1) = 0$, and $\hat{f}(r^{-1}) = -\hat{f}(r)$. An equivalent form for the local distortion function is

$$\left|\hat{f}\left(\frac{x}{y}\right)\right| = \frac{1}{(1 - 10^{-Tk/10})} \frac{|x^k - y^k|}{\max\{x^k, y^k\}} \tag{2.20}$$

which is a *relative* loudness difference. In summary, the optimal normalised gain-invariant (mean-square) approximation to the distortion function of (2.16) is given in continuous excitation variables by

$$D(x; y) = \frac{1}{1 - 10^{-Tk/10}} \left(\int \frac{|x(\omega)^k - y(\omega)^k|^p}{\max\{x(\omega)^{kp}, y(\omega)^{kp}\}}\, d\omega\right)^{1/p} \tag{2.21}$$

Figure 2.1 compares the newly derived class of nonlinearities with the logarithmic nonlinearity.

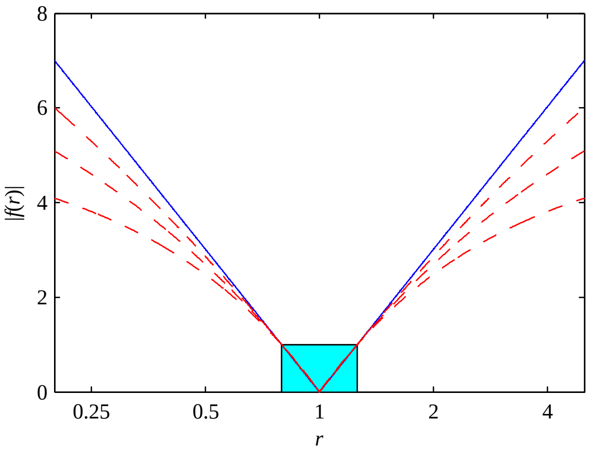


Fig. 2.1 Local distortion function $|f(r)|$. Solid: $f = 10|\log_{10}(r)|$; Dashed: Eq. (2.20), top to bottom: $k = 0.23$, $k = 0.5$, $k = 0.9$. The shaded area defines the boundaries of the generalised just-noticeable limen with $T = 1$. Two excitations points x and y are indistinguishable if their ratio $r = x/y$ lies on a curve within the region.

A rather surprising result is obtained by examining the limit of (2.19) as the loudness exponent

k approaches zero. We have

$$\lim_{k \to 0} \frac{1 - \left(\frac{y}{x}\right)^k}{1 - 10^{-Tk/10}} = \frac{1}{T} 10 \log_{10}(x/y), \quad x > y \tag{2.22}$$

and similarly

$$\lim_{k \to 0} \frac{-1 + \left(\frac{x}{y}\right)^k}{1 - 10^{-Tk/10}} = \frac{1}{T} 10 \log_{10}(x/y), \quad x < y \tag{2.23}$$

Thus

$$\lim_{k \to 0} \frac{1}{(1 - 10^{-Tk/10})} \frac{|x^k - y^k|}{\max\{x^k, y^k\}} = \frac{10|\log_{10}(x/y)|}{T} \tag{2.24}$$

which is the classical (normalised) logarithmic distortion function. The standard dB-distance between excitation patterns is then seen to be a special case of the derived class of distortion measures (2.20), in the limit as the loudness exponent approaches zero.

2.3 Masking Properties of General Auditory Distortion Functions

In this section we narrow the scope of the discussion to a single class of threshold phenomena: masking thresholds. Any distortion function seeking to replace noise-mask measures must subsume masking results in the appropriate limits. We will investigate three such conditions: 1) absolute threshold, 2) tone-masking-noise (TMN), and 3) noise-masking-tone (NMT). In the course of the analysis, a refinement to (2.12) in the form of a weighting function will be derived to account for the frequency dependence of the masking offset. Finally, we give an example of supra-threshold predictions in the case of a partially masked tone.

2.3.1 Absolute Threshold

Let x and y be two excitation patterns. For the purposes of threshold calculations, we will use the distortion function of (2.11), $p = \infty$, where the excitation pattern is taken explicitly to be the *standard* total excitation, reiterated in continuous form here, with the internal noise contribution presented explicitly:

$$D_f(x; y) = \max_{\omega} \left| f\left(\frac{x(\omega) + \epsilon(\omega)}{y(\omega) + \epsilon(\omega)} \right) \right| \tag{2.25}$$

where $\epsilon(\omega)$ is the excitation due to internal noise.

Under absolute threshold conditions, $x(\omega) = E_{\omega_c}(\omega)$, the excitation pattern of a just-noticeable sinusoid at frequency ω_c, and $y(\omega) = 0$. For the detection of a sinusoid at frequency ω_c against internal noise, only the main excitation is relevant; the peak deviation of the distortion pattern will occur at the frequency ω_c. This is because the slopes of a sinusoidal excitation decay more rapidly than the noise excitation. Hence the distortion function reduces to:

$$\begin{aligned} D_{\omega_c}(x,y) &= \max_{\omega} \left| f\left(\frac{E_{\omega_c}(\omega)+\epsilon(\omega)}{\epsilon(\omega)}\right)\right| & (2.26)\\ &= f\left(1+\frac{E_{\omega_c}(\omega_c)}{\epsilon(\omega_c)}\right) & (2.27)\\ &= f\left(1+s(\omega_c)\right) & (2.28)\end{aligned}$$

where $s(\omega_c)$ is the threshold factor of (1.5).

If the excitation distortion model is valid, then we must have $f((1+s(\omega_c)) = f(10^{T/10})$, or $T = 10\log_{10}(1+s(\omega_c))$. The factor $s(\omega_c)$ is frequency-dependent, showing that the difference limen fluctuates with frequency even in the relatively simple masking context, as is well-known. For many applications, however, it is often useful to have a single number describing the just-noticeable difference irrespective of frequency. This can be accommodated by introducing the frequency weights

$$W(\omega) = \frac{1}{f\left(1+s(\omega)\right)} \quad (2.29)$$

and redefining the distortion function (2.25) as

$$D_f(x;y) = \max_{\omega}\left|W(\omega)\cdot f\left(\frac{x(\omega)+\epsilon(\omega)}{y(\omega)+\epsilon(\omega)}\right)\right| \quad (2.30)$$

It is now obvious that

$$D_{\omega_c}(x;y) = \max_{\omega}\left|W(\omega)\cdot f\left(\frac{E_{\omega_c}(\omega)+\epsilon(\omega)}{\epsilon(\omega)}\right)\right| = 1 \quad (2.31)$$

expresses the condition of absolute threshold. In fact, the weighting function $f(1+s(\omega))$ of (2.29) is a generalisation of the threshold normalisation factor introduced in (2.12). Figure 2.2 gives a plot of the frequency weighting function, with $f = 10\log_{10}(\cdot)$ and Kapust's equation (1.5) for the Noise-Masking-Tone (NMT) threshold factor.

2.3.2 Masking

Noise-Masking-Tone

The distortion function as described in the previous section already has the capability to predict the masking threshold for general noise-masking-tone contexts, since absolute threshold itself is a special case of NMT. More formally, assume that $x(\omega) = E_{\omega_c}(\omega) + E_N(\omega)$ is the combined

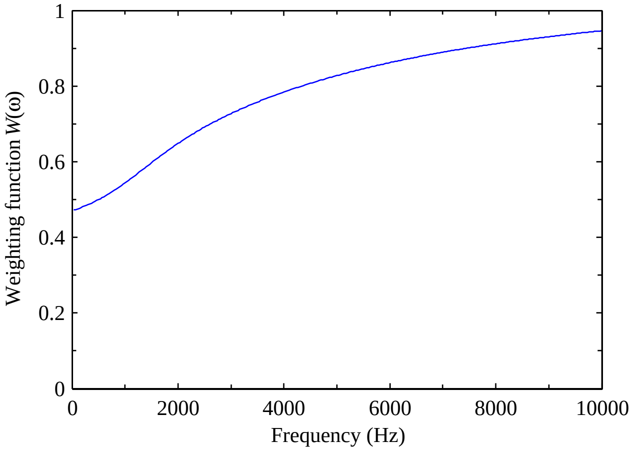


Fig. 2.2 Weighting function for $f = 10\log_{10}(\cdot)$

excitation pattern of a sine wave of frequency ω_c in noise, and $y(\omega) = E_N(\omega)$ is the excitation pattern of the noise alone. Once again, the detection of the sine wave only involves the auditory filter tuned to the frequency ω_c and the distortion function reads:

$$D_{\omega_c}(x;y) = \left| W(\omega_c) \cdot f\left(\frac{E_{\omega_c}(\omega_c) + E_N(\omega_c) + \epsilon(\omega_c)}{E_N(\omega_c) + \epsilon(\omega_c)}\right)\right| \tag{2.32}$$

$$= W(\omega_c) \cdot f\left(1 + \frac{E_{\omega_c}(\omega_c)}{E_N(\omega_c) + \epsilon(\omega_c)}\right) \tag{2.33}$$

For levels larger than absolute threshold, this becomes

$$D_{f_c}(x;y) \cong W(\omega_c) \cdot f\left(1 + \frac{E_{\omega_c}(\omega_c)}{E_N(\omega_c)}\right) \tag{2.34}$$

$$= W(\omega_c) \cdot f\left(1 + s(\omega_c)\right) = 1 \tag{2.35}$$

which verifies the model.

Tone-Masking-Noise

While the phenomenon of noise-masking-tone has been well investigated in the psychoacoustics literature, there has been comparatively little work on the tone-masking-noise case. Given the fact that excitation patterns are generally derived from masked audiograms involving NMT experiments, and given that the weighting function derived in Section 2.3.1 involved masking offsets taken from the NMT case, it seems doubtful that the distortion function of (2.30) can account for tone-masking-noise phenomena.

The key qualitative difference between these two masking contexts is that in tone-masking-noise, the detection point does *not* occur at the frequency of the masking sinusoid. To illustrate this characteristic, let us consider the distortion patterns of the two cases. Above threshold, we have

$$\mathcal{D}_{\mathrm{NMT}}(\omega) = W(\omega) \cdot f\left(1 + \frac{E_T(\omega)}{E_N(\omega)}\right) \tag{2.36}$$

$$\mathcal{D}_{\mathrm{TMN}}(\omega) = W(\omega) \cdot f\left(1 + \frac{E_N(\omega)}{E_T(\omega)}\right) \tag{2.37}$$

in obvious subscripts. These distortion patterns are plotted in Fig. 2.3 for a 1 kHz tone and critical-band wide noise centered at the same frequency, with relative powers chosen so that neither masks the other. The distortion is computed with the logarithmic measure $f = 10\log_{10}(\cdot)$, but any normalised gain-invariant metric would produce a qualitatively similar result.

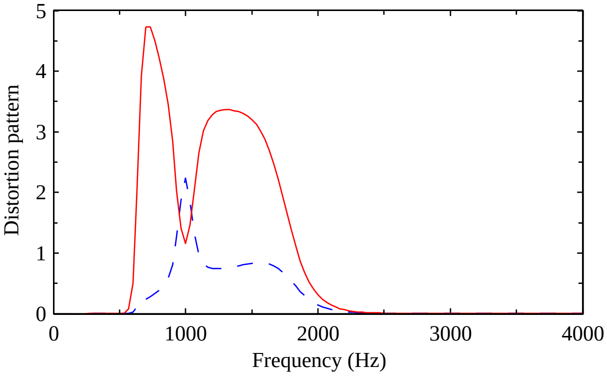


Fig. 2.3 Distortion patterns $\mathcal{D}$ for tone masker, noise probe (solid), and tone probe, noise masker (dash).

Referring to the plot, when noise is considered masker, and tone the probe, maximum distortion occurs at the frequency of greatest excitation, 1 kHz. However, when the roles are reversed, the tone must cover the larger bandwidth noise, and in general, maximum deviations occur at the auditory filters to the side of the center frequency. While there are two sidelobes, the lower frequency sidelobe tends to dominate detection because the upwards spread of the tone excitation provides stronger cover for higher frequencies. This type of off-frequency listening manifests any time the bandwidth ratio between masker and probe is less than one. Observe also that in the region of main activity, the two patterns are roughly reciprocals of one another, with the local maximum exchanged for a local minimum, in compliance with the form of (2.36) and (2.37). At frequencies farther away from the center, both excitation patterns reach low levels and hence are masked by internal noise, driving distortion to zero.

The above implies that off-frequency listening plays a role in determining the masking offset whenever the probe bandwidth is larger than the masker bandwidth, with the possibility of contributing to the well-known masking asymmetry. To explore this further, we used the model to predict masking offsets for both TMN and NMT cases; smoothed versions of these are displayed in Fig. 2.4.

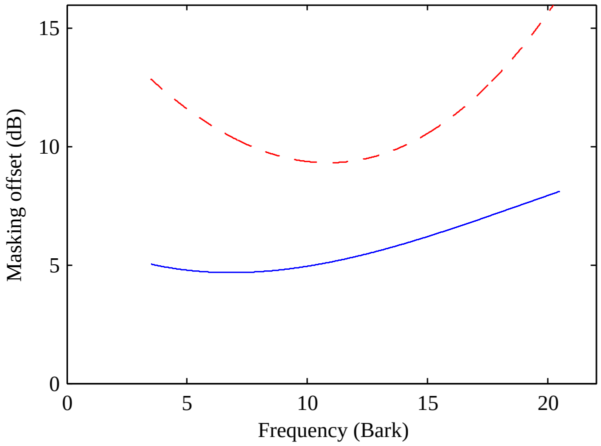


Fig. 2.4 Model predictions for 1) noise-masking-tone (solid) and 2) tone-masking-noise (dash) masking offsets

For this experiment, the noise masker/maskee is of 1 Bark bandwidth. As expected, the masking offset predictions for noise-masking-tone are excellent, in the 5 dB range, increasing slightly with frequency. Some masking asymmetry is predicted, anywhere from 5–10 dB of additional masking with a tone masker and noise probe, but the overall values are short of the more realistic 15–35 dB masking offset typical of TMN. Hall in [12] has suggested that off-frequency listening is insufficient to account for the asymmetry of masking—in particular, the qualitative change that occurs when probe bandwidth exceeds masker bandwidth. The temporal information contained in the phase of a tone may be part of that mechanism — lost as it is to the standard excitation pattern.

2.3.3 Supra-Threshold Distortion

We have repeatedly emphasized that all gain-invariant functions, by virtue of the transferability of limen discussed in Section 2.2.4, possess identical threshold properties. Thus the masking offset predictions, illustrated by example with the logarithmic distortion, are inherited by the entire class of normalised gain-invariant measures.

The same is not true for supra-threshold phenomena. Indeed, this is where the utility of the new family enters: as generalised functions consistent with threshold phenomena but offering modelling freedom for the relatively uncharted territories of audible distortion. As an example, Fig. 2.5 uses the measure of (2.14) to give distortion predictions for the case of a partially masked tone.

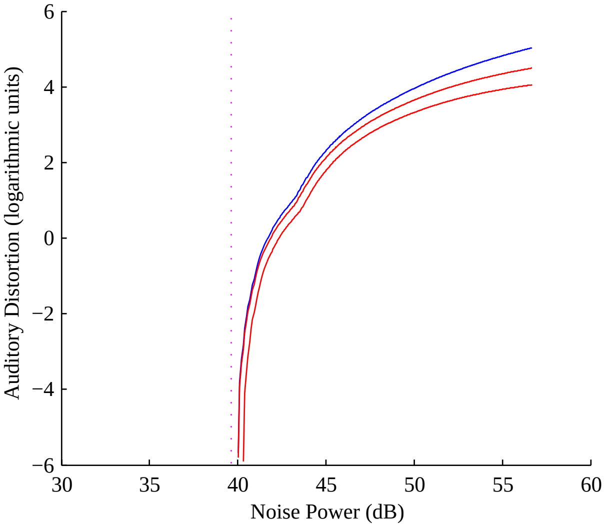


Fig. 2.5 Supra-threshold distortion predictions for tone (1 kHz, 50 dB SPL) partially masked by critical-band wide noise. Solid line: $f = 10\log_{10}(\cdot)$, dash: Eq. (2.20); top to bottom: $k = 0$, $k = 0.5$, $k = 0.95$. Dotted line: TMN masking threshold

In this scenario, a 1 kHz tone of 50 dB SPL is submerged in critical-band wide noise of varying level. The excitation patterns $x = E_T + E_N$ and $y = E_T$, with E_T and E_N the patterns of tone and noise respectively, are used as the arguments for $D(x; y)$ to quantitate the distortion to the tone. When the noise power is below masking threshold, there is no distortion to the tone. As threshold is exceeded, a rise in audible distortion, similar to the loudness recruitment of partially masked probes, occurs. Four different gain-invariant functions are displayed, each predicting an identical threshold but providing alternative extensions.

While the functions in Figure 2.5 do resemble classical loudness recruitment curves (see [37] for an example), too much emphasis should not be placed on finding parameters for a precise matching. This is because, 1) loudness recruitment is typically experimentally investigated under fairly simple conditions (sinusoid in noise, for instance), and 2) partial loudness is in itself only one particular quality, one face of the multi-faceted, multi-valued object of auditory distortion.

2.4 Chapter Summary and Caveats

This chapter analyzed the general form of perceptual distortion functions consistent with psychoacoustic threshold phenomena. We also analyzed the form of a variety of distortion measures topical in current applications and, notwithstanding their relative experimental success, observed that they either 1) tacitly relied on a decomposition of the corrupted signal into signal + uncorrelated noise, which was not usually appropriate, or 2) were not consistent with, and hence unable to predict, threshold results. To address the first issue, we formulated measures as distances in the space of excitation (standard or CEPS) representations. Using the just-noticeable difference limen, the second issue was addressed by the hypothesis of invariance with respect to certain gain transformations; proposed as a starting point, this concept lead to a generalisation of Zwicker's 1 dB rule and a class of distortion functions among which the logarithmic distortion function was a special case. The class inherited all the at-threshold properties of the logarithmic function, but gave different predictions for the important case of supra-threshold (audible) distortion.

It is worth emphasizing that we make no claims to having found the *ultimate* quantifier of auditory distortion. By looking for measures on the space of excitations, we have restricted ourselves to the information that can be obtained from just such a representation. Further restrictions are incurred, depending upon the precise types of excitation variables used. For instance, should non-time-spread excitation be used, temporal masking phenomena cannot be subsumed; if *standard* excitation is used, and not CEPS, the distortion measures are restricted further to only include quantitation of magnitude distortion, and, as such, even the noise-to-mask ratio function is not entirely subsumed by the proposed measure. A more complete characterization would include measures for not only different degrees, but also *qualitatively* distinct types of distortion. The perception of these types of differences are also modified by the contexts under which they are presented; towards this end, some type of "cognitive" modelling would undoubtedly be necessary for best performance.

Chapter 3

Excitation Distortion for Audio Coding

In the previous chapter, we derived a family of functions possessing sound psychoacoustical properties. The general form of that class was

$$D(x;y) = \frac{1}{1-10^{-k/10}} \left(\int W^p(\omega) \frac{|x^k(\omega) - y^k(\omega)|^p}{\max\{x^{kp}(\omega), y^{kp}(\omega)\}} \, d\omega \right)^{1/p} \tag{3.1}$$

with x and y total excitation patterns, $0 \leq k \leq 1$, $p > 0$, and where $W(\omega)$ is a weighting function having the structure (2.29), if desired. We used the term "Excitation Distortion" to denote distortion functions of the above type.

In this chapter, we raise the question of how to incorporate (3.1) in audio coding structures. Before we begin, however, let us review a few standard ideas about distortion measures and quantization.

3.1 Equivalent Distortion Measures

There are two notions of "equivalent" distortion measures, both introduced in [10]. The first is directly taken from the idea of equivalent metrics (or topologies) from mathematics. Given two distortion measures $d(x, y)$ and $d'(x, y)$ defined on a set S, we say that d is stronger than d', writing $d \gg d'$, if small d implies small d'. More precisely, $d \gg d'$ if, for any $\epsilon > 0$, there exists $\delta > 0$ such that $d < \delta$ implies $d' < \epsilon$. If both $d \gg d'$ and $d' \gg d$, then d and d' are *equivalent*, and we write $d \equiv d'$.

The above definition of mathematical equivalence is fairly weak. The equivalence only states that d and d' measure approximately the same effect. For the $\mathcal{L}^p$ metrics defined by

$$d_p = \left(\int |f - g|^p \, d\omega \right)^{1/p}, \tag{3.2}$$

Minkowski's inequality shows that $d_p \gg d_q$ whenever $p \geq q$. However, if the space is *finite-dimensional*, so that a finite sum replaces the integral in (3.2), then all the $\mathcal{L}^p$ metrics become *equivalent*. This is easily shown by the two inequalities $(\sum_{i=1}^n |x_i|^p)^{1/p} \leq n^{1/p} \max_{i=1}^n |x_i|$, and $\max_{i=1}^n |x_i| \leq (\sum_{i=1}^n |x_i|^p)^{1/p}$. Even more, it can be shown that *all* metrics arising from a norm on a finite-dimensional space are equivalent. Since, in any practical computer implementation, the spaces are of necessity finite-dimensional, the concept of mathematical equivalency is generally toothless.

What we are more interested in is whether two distortion measures will pick different representation points among a set of possible representations. Along these lines, the second notion of equivalence introduced in [10] is more useful, called *nearest-neighbor equivalence.* To define it, let $d(x, y)$ and $d'(x, y)$ be two distortion metrics on a set S, and let $R \subset S$ be a set (finite or infinite) of reproductions. Then d and d' are nearest-neighbor equivalent if, for any $x \in S$,

$$\arg\min_{y \in S} d(x, y) = \arg\min_{y \in S} d'(x, y) \tag{3.3}$$

This notion of equivalence is much more in line with the kind of equivalence we want in using distortion measures for quantization applications. If two distortion measures are equivalent in the nearest-neighbor sense, then they will produce identical quantization choices. A simple example is as follows: define $d' = f(d)$. Then d' and d are nearest-neighbor equivalent whenever f is an increasing positive function.

We can apply these notions of equivalency to the excitation distortion family (3.1). How different are the measures for parameters k, p and weighting function $W(\omega)$? The following theorem provides an answer.

Theorem 3.1. *The family of ED functions (3.1), restricted to the space of finite-dimensional discrete excitation patterns:*

1. *Define* equivalent *metrics over all parameters* $(k, p, W(\omega))$*,*

2. *Are* not *nearest-neighbor equivalent for different choices of parameters* $(k, p, W(\omega))$*.*

Thus, while every function in (3.1) approximately quantitates distortion in the same way, they will not in general select the same reproduction points in a quantization algorithm.

3.2 Quantization Structures

3.2.1 Basic Ideas

A quantizer on the set S is a function $q : S \to T$ where T is a finite subset of S. The values $y_i \in T$ form the set of reproduction values, and the sets $C_i = q^{-1}(\{y_i\})$ partition S into cells. Thus the complete specification of q requires only the definition of the cells C_i and the reproduction set y_i associated with each cell. If D is a distortion measure on S, then the distortion incurred by quantizing x is given by $D(x, q(x))$. Let us further suppose that there is a probability measure on S, with density $f(x)$, $x \in S$. Then one defines the average distortion incurred by q as

$$D(q) \equiv E[D] = \int_S D(x, q(x)) f(x)\, dx \tag{3.4}$$

A given quantizer will also have a rate $R(q)$ associated with it. Usually this is defined as the average length of the codewords used to enumerate T. Without actually specifying a coding scheme, one can also use more abstract definitions such as $R(q) = \log_2 |T|$ or $R(q) = \text{entropy } T$. Typically, D is small only when the number of reproduction points is large, hence one obtains a trade-off between R and D. One naturally seeks to obtain the optimal trade-off, which is given by the distortion-rate function [11]:

$$D(R) = \inf_{q : R(q) \leq R} D(q) \tag{3.5}$$

The question, of course, is how to design q so that it operates near the optimal $D(R)$ curve. One methodology is the so-called "unconstrained" approach. For instance, fixing the rate R, one can try to obtain the best partitioning $\{C_i\}$ and reproduction set $y_i \in T$ minimizing D. Classic results along this line include the Lloyd conditions for optimal quantization [7]. While general, the unconstrained approach can lead to designs which are impossible in practical systems. For instance, many of the D-R bounds can only be approached by a block-encoding where the block-size approaches infinity, implying long delay and large memory requirements.

The other approach, which is pervasive in much of audio, image and video coding, is the "operational" rate-distortion paradigm. Here, one abandons the idea of unconstrained optimality. Instead, a coding structure is first fixed, *a priori*, which satisfies the complexity and memory requirements of the system. The structure is parameterized by a number of variables, and then one searches for the best rate-distortion performance *among those parameters* [22].

A simple example will serve to illustrate the method. Define a set of N scalar uniform quantizers $\mathbf{q} = \{q_i\}$, each parameterized by step-size $\delta = \{\delta_i\}$, acting upon N variables $\mathbf{x} = \{x_i\} \in S$. By making a choice of step-size parameter δ, we fix the quantizers and thus specify a single rate-distortion point $(D(\mathbf{q}), R(\mathbf{q}))$. By sweeping δ over all possible values of step-sizes in $(\mathbb{R}^{+})^N$, we obtain the set of all *achievable* rate-distortion points. The convex hull of this set defines the *operational rate-distortion curve.* It is called operational because the points on the curve are directly achievable by the system for some step-size parameter δ.

A special case of importance occurs when the set of possible parameters is finite — for example, in the above, by restricting the step-sizes, say, to values $(1/2)^k$, for positive integer k less than some number. The set of achievable rate-distortion points is then also finite, and the problem of finding good operating points becomes one of (finite) discrete optimization.

We used a probabilistic definition for the distortion D above. This involves finding some statistical model for the sources to be coded. For many natural sources (audio, images etc.), it may be impossible to find good models which can fully characterise a given source, — particularly for non-stationary processes. In this case, one may move from a system involving fixed parameters optimizing an average distortion, to a system where the parameters dynamically change depending upon input, and where average distortion is replaced with *actual* distortion.

Returning to our scalar quantization example, suppose the variables to be quantized $\{x_i\}$ are discrete Fourier coefficients of audio, where the coefficients are computed over some time-frame. Instead of trying to design the step-sizes to obtain the optimal trade-off between average distortion over *all* possible sample points, versus rate, we can attempt, given an actual vector of coefficients $\{x_i\}$, to find the parameters δ which trade-off the actual distortion $D(\mathbf{x}, q(\mathbf{x}))$ (to rate). Since the step-sizes change with every input, an additional coding system needs to be designed for the parameters, to be sent as *side-information*, with an associated rate cost $R(\delta)$. The rate-distortion trade-off then becomes one of balancing actual distortion $D(\mathbf{x}, q(\mathbf{x}))$ on the one hand, and *total* rate due to both quantizer and side-information: $R = R(\mathbf{q}) + R(\delta)$. In this system, the rate-distortion optimization is carried out every frame.

3.2.2 Classification of R-D Optimization Problems

Having defined the set of operational rate-distortion points, we can now formulate two types of problems: rate-constrained versus distortion-constrained. Let the quantizers $\mathbf{q}$ be parameterized over a set P. They are given by:

1. (Rate-constrained Problem). Fixing a target rate R',

$$\min_{\mathbf{q} \in P} D(\mathbf{q}), \quad \text{subject to} \tag{3.6}$$

$$R(\mathbf{q}) \leq R' \tag{3.7}$$

2. (Distortion-constrained Problem). Fixing a target distortion D',

$$\min_{\mathbf{q}\in P} R(\mathbf{q}), \quad \text{subject to} \tag{3.8}$$

$$D(\mathbf{q}) \leq D' \tag{3.9}$$

Independent and Dependent Quantization

If the set of coder parameterizations P is finite, then one obvious solution exists for finding the optimal solution to either of the formulations (3.7) and (3.9): namely the brute-force approach of computing all the rate-distortion data for every operating point in P. Sometimes this is the only course of action, especially when there is no structure to either D or R, or when $|P|$ is small. More often than not, however, either D or R admits some decomposition which allows more efficient algorithms than exhaustive search. One well-researched and convenient assumption is the case when D is additively separable; i.e.

$$D(x_1, q_1(x_1), \ldots, x_n, q_n(x_n)) = d_1(x_1, q_1(x_1)) + d_2(x_2, q_2(x_2)) + \cdots + d_n(x_n, q_n(x_n)) \tag{3.10}$$

for some functions $d_1, \ldots, d_n$. Equation (3.10) is quite strong: it essentially reduces the multivariate function D into a sequence of single-variate functions d_i. Such simplification is the key to a variety of efficient algorithms, such as Lagrangian relaxation [31].

Sometimes a function may not possess the decomposition (3.10), but a weaker representation:

$$D(x_1, q_1(x_1), \ldots, x_n, q_n(x_n)) = \sum_{i=1}^{n} d_i(x_{i-a}, q_{i-a}(x_{i-a}), \ldots, x_{i+a}, q_{i+a}(x_{i+a}))^1 \tag{3.11}$$

for some functions d_i, and a positive number a. Here, each of the functions d_i depend upon a *range* of quantizers, from $i-a$ to $i+a$, instead of a single quantizer, as in (3.10). Depending upon how large a is, one obtains more or less dependency of the functions d_i. In the case each d_i is a function of *all* quantizers, one has a global dependency and the representation (3.11) is entirely trivial, of course.

Based on the preceding discussion, it should not be a surprise that if a distortion measure has the form (3.10), then the rate-distortion optimization problem is called an *independent* quantization problem, and if it should have the form (3.11), for some $a > 0$, then it is called a *dependent* quantization problem.

The above example illustrated the case where the functions were additively separable. We now provide a more general notion of independent and dependent quantization. Towards this end, let

[1] The subscript notation means that $i - a = \max\{1, i - a\}$, and $i + a = \min\{n, i + a\}$.

us make the following definition:

Definition 3.1. *A function f of the variables $\{x_i\}_{i=1}^n, \{q_i\}_{i=1}^n$ is called F-separable (or more succinctly* separable*) if there exists a function F increasing in each dimension and functions $\{f_i\}_{i=1}^n$ such that f admits the decomposition:*

$$f(x_1, q_1(x_1), \ldots, x_n, q_n(x_n)) = F[f_1(x_1, q_1(x_1)), \ldots, f_n(x_n, q_n(x_n))] \tag{3.12}$$

This definition gives a natural generalization to the notion of additive separability. In particular, additive separability occurs when $F = \sum_i |\cdot|$. If $f > 0$ is separable, there is no loss in generality in assuming that $F > 0$ and $f_i > 0$. Hence the representation (3.12) implies some type of grouping or integration of individual distortions (or rates) f_i. We now define an *independent* quantization problem as one in which *both* the rate function R and the distortion function D are separable, and a *dependent* quantization problem as one in which at least one of R and D is not separable. We shall also say that, in the case that f is not separable, that it is "F-type" if it has the *dependent* decomposition:

$$f(x_1, q_1(x_1), \ldots, x_n, q_n(x_n)) = F[f_i(x_{i-a}, q_{i-a}(x_{i-a}), \ldots, x_{i+a}, q_{i+a}(x_{i+a}))] \tag{3.13}$$

for some function F increasing in each dimension, and some positive number a. Once again, this provides a natural generalisation of (3.11), which is additive-type.

Broadly speaking, dependent quantization problems pose greater difficulties than independent problems in terms of optimal bit-allocation. For instance, certain dynamic programming algorithms increase in computational complexity as the neighborhood of dependency increases [29], and standard incremental bit allocation can fail when the quantizers are dependent [5]. We shall not try to provide any survey of the large literature on independent and dependent bit-allocation, however; instead we invoke the germane approaches when required for our specific problem (Section 3.3.5).

3.3 Quantization with Excitation Distortion

3.3.1 Quantization Domain

Let us assume an operational rate-distortion paradigm, by fixing a set of uniform scalar quantizers $\{q_i\}$ acting on variables $\{x_i\}$, parameterized by step-sizes $\{\delta_i\}$. One question which arises immediately is: what domain should the quantization variables $\{x_i\}$ be chosen in? The standard transform domains are the DFT, DCT and MDCT domains. But the fact that the excitation pattern can be viewed as an invertible linear transform on power spectra (Section 1.2) also suggests yet another possibility: quantization in the excitation domain itself.

To explore this possibility, let us turn back to the distortion measure (3.1). It is obvious that if the variables to be quantized consist of discrete total excitation variables, then D is separable, with $F = (\sum_i |\cdot|)^{1/p}$, and the single-variate distortions given by

$$d_i = W^p(\omega_i) \frac{|x^k(\omega_i) - y^k(\omega_i)|^p}{\max\{x^{kp}(\omega_i), y^{kp}(\omega_i)\}}. \tag{3.14}$$

Moreover, in the case $k = 0$, one obtains the logarithmic distortion, and if we choose to quantize logarithmic excitation variables $\log(x_i)$, then the distortion measure D is not only separable, one even converts the ratio measure D into a weighted $\mathcal{L}^p$ difference metric! Note that for no other parameter k is it possible for D to be a difference measure, because $f = \log x$ is the only (continuous) function satisfying $f(x/y) = f(x) - f(y)$, and hence the only transformation that can turn the ratio measure into a difference measure.

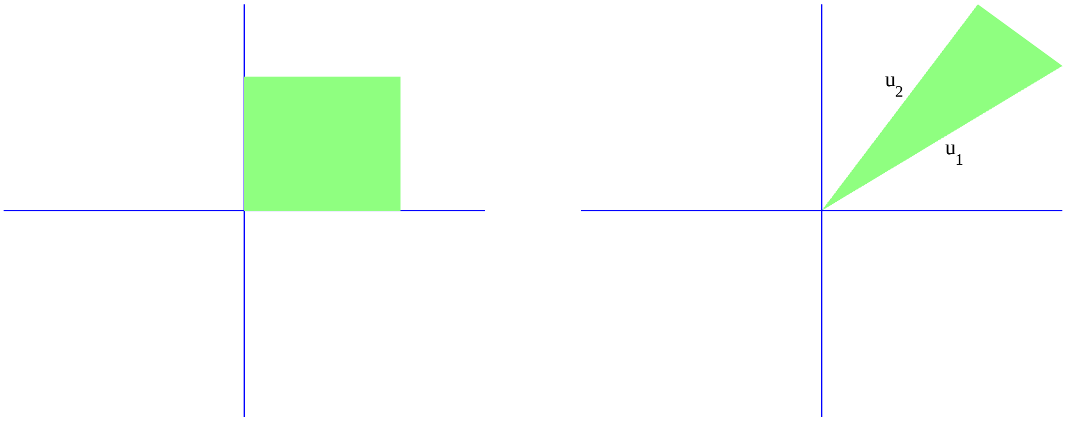
In contrast, now suppose that the quantizers act on a standard domain, such as the DFT domain. One may still admit a decomposition of the form (3.14), but due to the spreading operator U, each d_i is a function of *every* quantizer q_j, giving rise to a very *non-separable* distortion function. The upshot is then this: quantizing in a DFT (or any other non-spread) domain results in a (globally) dependent quantization problem, whereas quantizing in the excitation (or more generally spread) domain results in a simpler *independent* quantization problem.

The picture is not quite as convenient as this, however. First, for simplicity, let us confine the discussion to the standard excitation pattern, instead of the CEPS. The discrete excitation transform is then an invertible linear mapping U from $\mathbb{R}^n$ to $\mathbb{R}^n$. The space of *power spectra* however, is $S = (\mathbb{R}^+)^n \subset \mathbb{R}^n$. The space of excitation patterns C then occupies a subset of $\mathbb{R}^n$, given by the image $U(S)$, which is the convex cone spanned by the columns of the matrix representative $\mathbf{U}$.[2] Moreover, since the entries in $\mathbf{U}$ are positive, the convex cone of excitations is always a subset of $(\mathbf{R}^+)^n$.

Figure 3.1 provides an illustration of the situation with $n = 2$. Because the set of excitation spectra lie in a subset of the positive octant, the use of a rectangular lattice (such as occurs with scalar quantizers) can easily contain reproduction points lying outside the convex cone C. These points, under the action of the inverse $\mathbf{U}^{-1}$, are mapped *outside* of the positive octant in Figure 3.1(b) and hence do not form valid power spectra.

The possibility of "negative" power spectra when quantizing in the excitation domain is intimately related to the problem of masking threshold deconvolution, as raised by Johnston in [15]. In that paper, the author realised that the computed masking threshold M had to be

[2] Recall that the convex cone spanned by $v_1, \ldots, v_n$ is the set of all vectors of the form $\alpha_1 v_1 + \ldots \alpha_n v_n$, with $\alpha_i \geq 0$.



(a) Set of Power Spectra

(b) Set of Excitation Spectra

Fig. 3.1 Action of **U** on the set of power spectra. The image of the shaded area in (a) under mapping **U** is the shaded area in (b), bounded by the column vectors $[u_1, u_2] = \mathbf{U}$. $\mathbf{U} = [u_1, u_2]$

"deconvolved" into an unspread domain for comparison with the error signal. This was done by multiplying the masking threshold by the (non-singular) inverse spreading matrix. Johnston, however, noted the process "often leads to artifacts such as negative energy for a threshold, zero threshold etc.". The cause of this remained somewhat mysterious: "The unusual errors come about because the deconvolution process seeks a strictly numerical solution that disregards the physical and acoustical realities of the situation".

From the perspective of our present view, the phenomenon of negative thresholds in deconvolution is hardly unusual. It is a simple manifestation of the fact that, while the excitation pattern E is a member of the convex cone C, the masking threshold M, given by the division of E and a *non-frequency-constant* masking offset, is usually *not* a member. Hence M is mapped by the inverse transform to a point outside the positive octant. If the masking offset is *constant* in frequency, it is easy to check that $M \in C$, and the deconvolution process never produces negative thresholds. But because the convex cone of excitations occupies such a small volume out of $\mathbb{R}^n$, the process of masking offset division generally moves the excitation point outside of C, resulting in the phenomena.

The problem of reproduction vectors lying outside the space of excitation patterns is a serious one, and for that reason, we shall henceforth assume a coding structure where the quantizers q_i act in a non-spread frequency domain x_i. This does not mean that it is impossible to code in the excitation domain. In particular, the thesis [33] provides one method of overcoming the issue.

3.3.2 Framework

Our main goal will not be to design a complete coder driven by the excitation distortion measure, but merely a prototype of one sufficiently sophisticated and general to produce *both* ED-coded and NMR-coded files, thereby permitting experimental comparisons between the two metrics (Chapter 4). Consequently, we shall not be concerned with the design of an entropy coder mapping the reproduction values $q_i(x_i)$ to binary sequences, nor even the coding of the side-information δ_i. In order to attain some generality, we shall pose no restrictions on the sampling rate, number of quantizers, or even the domain of the variables x_i— other than the fact that they lie in some standard non-spread frequency domain (DFT, MDCT etc.).

Of the two types of R-D optimization problems, we will choose the constrained-distortion framework (3.8). This mode of operation is useful for a variety of reasons. First, transparent coding can be expressed as a constant-distortion criterion: $D(\mathbf{x}, \tilde{\mathbf{q}}(\mathbf{x})) = K$ for some threshold K. Second, a qualitative understanding of supra-threshold distortion incurred from using a particular psychoacoustic model can only be evaluated subjectively when every frame is coded to the same distortion. Third, a distortion-constrained scheme can form the kernel for a rate-constrained scheme. In particular, a constant-distortion engine can find a *relative* allocation of bits b_i, while an outer algorithm adjusts the *absolute* sizes of b_i to meet the rate constraint. This process occurs in the MPEG coding standard, for example, where the inner loop adjusts the relative step-sizes of the quantizers q_i to meet a distortion constraint (masking threshold), while an outer loop varies a global gain factor to meet the rate target.

3.3.3 Causal Coding versus Non-Causal Coding

The standard speech or audio coder is a *causal* coder; namely, a file of data is sectioned into overlapping time frames, transformed into a frequency representation and quantized. Each frame is processed sequentially, and in as much that there is any dependence between frames in terms of quantization decisions, only past frames $n-1, n-2, \ldots$ can influence the coding outcomes of the n-th frame.

Our distortion measure (3.1), like most other audio metrics, is posed as a measure between two spectra occurring simultaneously[3] (this holds even in the case when time-spread excitations are used), hence the standard causal coding paradigm would at first glance also seem suitable: namely, the data is divided into overlapping frames, the short-term reference spectra $X(\omega)$ computed and quantized to $\tilde{X}(\omega)$ so that the reconstructed excitation pattern $E(\tilde{X}(\omega))$ at frame n matches as closely as possible the reference excitation pattern $E(X(\omega))$ at frame n, in the sense of (3.1). The

[3] More precisely, if $E(t, \omega)$ is one set of time-indexed excitations and $F(t, \omega)$ is another set, then the distortion measure only utilises differences between $E(t_1, \omega)$ and $F(t_2, \omega)$ for $t_1 = t_2$

constrained-distortion optimization is carried out every frame; moreover, if the distortion target K is kept fixed for every frame, then the output file is distortion-constrained in the strong sense that $D < K$ for all time frames. It is true that each frame now has a different rate R_i associated to it, but there still exist reasonable notions of rate for the entire file; for instance the average $\frac{1}{N}\sum_{i=1}^{N} R_i$ over all frames.

The situation is actually much more complicated. Figure 3.2 gives a block-diagram overview of the coding process just named, illustrated in the case of 50% overlap; however, the following points will apply at *any* overlap greater than zero. Observe that there are actually *two* versions of the "reconstructed" excitation pattern that one may sensibly define. Version 1 is derived directly from the quantized spectral coefficients. Version 2 is derived from a spectral analysis of the final time-domain coded signal, *after* overlap-add.

Generally, these two versions of excitation patterns will not be the same, unless there is no overlap in frames. It is also clear that the true reconstructed excitation pattern is the one obtained only after time-domain addition, because this is the signal upon which the listener performs the perceptual processing. The comparison between reference and reconstruction thus must be an end-to-end process; the intermediary Version 1 excitation pattern is little more than a by-product of the signal decomposition.

What ramifications does this have for quantization? The most important one is that the excitation-distortion allocation problem becomes not only frequency-dependent, but also time-dependent. This is readily seen by observing that the Version 2 reconstructed excitation pattern at frame n is a function of the quantized spectral coefficients of both frame $n-1$ and frame $n+1$, in addition to the spectral coefficients of frame n.

The time-frequency dependency need not be confined to excitation distortion alone. The fact that the reconstructed spectrum after overlap does not match the quantized spectral coefficients implies that *any* measure posed in the frequency domain gives rise to a time-frequency dependent allocation problem, including NMR, Spectral Distortion (SD), as well as ED. Figure 3.3 shows the dependencies in the time-frequency plane for different types of measures and transform overlaps. The grid represents some time-frequency decomposition, and each square is associated with both a quantizer and a distortion point. The lightly shaded area denotes the neighborhood of all quantizers which have some influence over the single darkly shaded time-frequency distortion point.

An implication of the time dependency is that constrained distortion optimization in the strong sense ($D_n < K$ for all time frames n) is not generally possible with causal coding. This is because there is no guarantee that, even if the first n frames are quantized to satisfy $D_n < K$, it is possible to achieve the distortion constraint for frame $n+1$, i.e. $D_{n+1} < K$. For instance, it is very easy to find an example where the end-to-end excitation distortion in all frames k, $k < n$

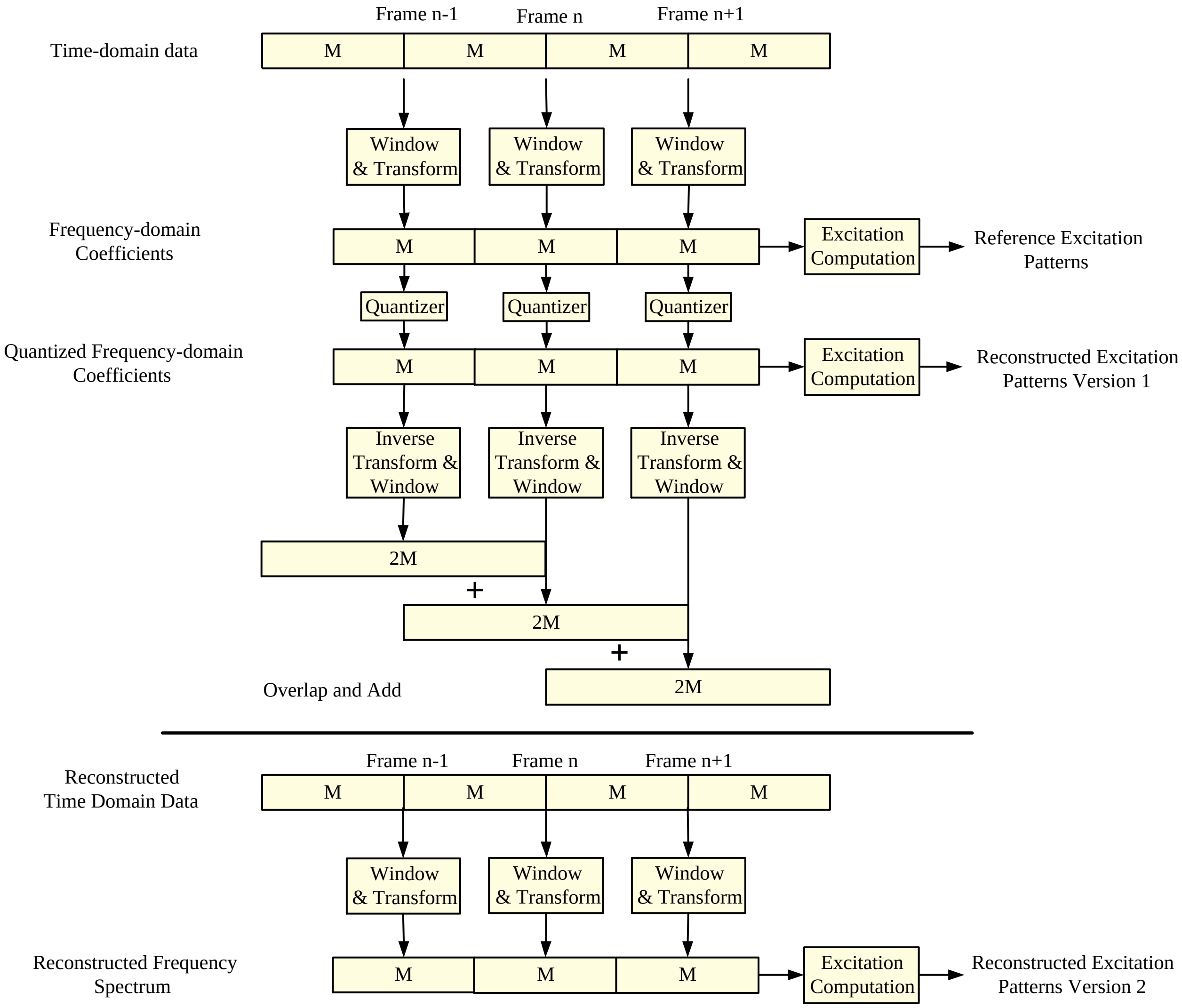
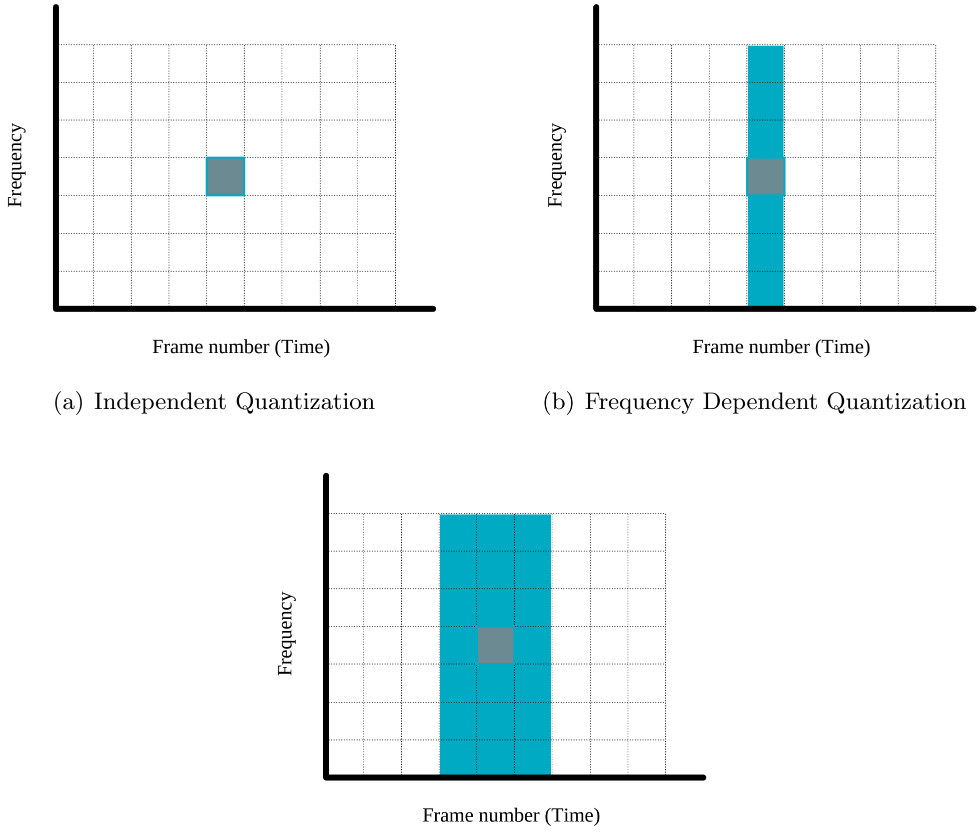

Frame n-1
Frame n
Frame n+1
Time-domain data
M
M
M
M
Window & Transform
Window & Transform
Window & Transform
Frequency-domain Coefficients
M
M
M
Excitation Computation
Reference Excitation Patterns
Quantizer
Quantizer
Quantizer
Quantized Frequency-domain Coefficients
M
M
M
Excitation Computation
Reconstructed Excitation Patterns Version 1
Inverse Transform & Window
Inverse Transform & Window
Inverse Transform & Window
2M
+
2M
+
Overlap and Add
2M
Frame n-1
Frame n
Frame n+1
Reconstructed Time Domain Data
M
M
M
M
Window & Transform
Window & Transform
Window & Transform
Reconstructed Frequency Spectrum
M
M
M
Excitation Computation
Reconstructed Excitation Patterns Version 2

Fig. 3.2 Coding Process with 50% frame overlap. Because of overlap-add reconstruction, the quantized frequency coefficients are not the same as the reconstructed frequency coefficients.



(a) Independent Quantization

(b) Frequency Dependent Quantization

(c) Frequency/Time Dependent Quantization

Fig. 3.3 Quantizer Dependencies. Diagram (a) is descriptive a measure like NMR, with no coder overlap; (b) applies for ED, with no coder overlap; (c) applies for any distance measure on frequency spectra, at coder overlap O satisfying $0\% < O \leq 50\%$.

are below 1 dB, but such that it is impossible to bring the end-to-end excitation distortion in frame n below 1 dB — even with perfect coding of the n-th frame. This is a direct consequence of frame-dependency and the sensitivity of error propagation in the overlapped coding system. Nor are such examples rare; we have found, experimentally, that it is more often the case than not that past quantization decisions made in previous frames make it impossible to achieve distortion targets in future frames.

There exist two obvious solutions to the posed dilemma. The first is to forego overlapped representations, so that one obtains only a frequency-dependent problem, making constrained-distortion causal coding feasible. There are a few problems with this approach. First, it prohibits any lapped-transform decomposition; among them the MDCT decomposition, which is perhaps the most popular and widely-used transform in audio coding. Second, the quantized coefficients must come from a non-overlapped analysis; by the Nyquist conditions for perfect reconstruction, the only time-window that may be used is a rectangular one, which has relatively poor sidelobe suppression properties. Finally, it is well-known that non-overlapped reconstruction in transform coders result in discontinuities at frame boundaries due to quantization: the result is a highly audible low-frequency clicking at all but the very highest rates.

The other possibility, and the one which we shall pursue, is to renounce causal coding in favor of non-causal coding[4]. That is, instead of processing time frames sequentially, we view the rate-distortion optimization over an entire audio file in the time-frequency plane, without constraints on which parts of the file need to be processed first. Non-causal coding increases the computational complexity of the coder, besides introducing a larger delay; on the flip-side, we obtain great generality and applicability in the retention of overlapped representations.

3.3.4 Non-Causal Coding: Formulation

Let us now setup our problem more concretely. Given an input signal S, we assume the existence of some (invertible) time-frequency transform which produces the discrete time-frequency representation $x(t_i, \omega_j) \equiv x(i, j)$ for S, with $1 \leq i \leq M$ and $1 \leq j \leq N$. For every (i, j) coefficient, we associate a set of uniform scalar quantizers $q_{i,j,b}$, parameterized by the whole number b (a quantity similar to “bits”), assigning a step-size to the quantizer through the formula:

$$\delta_{i,j,b} = K_{i,j}\Gamma^{-b}, \quad 0 < \Gamma < 1 \tag{3.15}$$

where $K_{i,j}$ is chosen large so that $q_{i,j,0}(x(i, j)) = 0$ (for instance, $K_{i,j} > 2x(i, j)$).[5] Thus the signal is quantized to zero, when 0 bits ($b = 0$) are used for each quantizer. Moreover, we will write $b = b(i, j)$ as a way of indicating the quantizer used in the location (i, j), so that $b(i, j)$ has the interpretation of being a bit-allocation. For instance, $b(i, j) = 3$ codes the state in which a scalar quantizer of step-size $\delta_{i,j,3}$ is used in the time-frequency location (i, j).

We can define the rate in a couple of natural ways, without explicitly constructing a code for the quantized coefficients. One definition is the empirical entropy of the quantization levels, thus estimating, in some sense, the rate of a coder whose quantized spectral coefficients are Huffman coded. This definition is rather intractable for the purposes of R-D optimization, however, since it can only be computed once every coefficient in the whole file is quantized.[6] Thus the rate

[4] Alternative synonyms include global or stream coding.

[5] The step-sizes of the quantizers constitute side-information, which can be further decomposed as side-information representing $K_{i,j}$ and b. One can choose here to decrease the amount of side-information by, for instance, setting $K_{i,j} = \text{const}\ \forall i, j$. Then the quantizers are completely parameterized by the integer b.

[6] *Given* a coded file, we will use empirical entropy as a measurement of rate in Chapter 4, however, since it gives

function R becomes globally dependent over the entire file! It will be more convenient to use a localised measure of rate. We shall define it as the average bit-allocation over the entire file:

$$R = \frac{1}{NM} \sum_{i=1}^{M} \sum_{j=1}^{N} b(i,j) \tag{3.16}$$

Observe that R is additively separable. It will, of course, tend to correlate with the empirical entropy, since distortion is reduced as b increases.

Let $D_i(x(i,j), \tilde{x}(i,j))$ be a distortion measure, defined between two spectra $x(i,j)$ and $\tilde{x}(i,j)$ for fixed time index i (such as ED (3.1) or NMR (2.1)). In line with our goal of constrained distortion, we define the distortion D as the maximum distortion over all time frames:

$$D = \max_{1 \le i \le n} D_i(x(i,j), q_{i,j,b}(x(i,j))) \tag{3.17}$$

Now our problem can be stated thusly: *Given a target distortion K, to find a bit allocation $b(i,j)$ such that $D < K$ with minimal rate R.*

3.3.5 Optimal and Incremental Approaches to Bit Allocation

Having formulated specifically the rate-distortion problem of interest in the previous section, we can now give a discussion of which bit allocation algorithms may be applicable. For the purposes of an algorithmic complexity analysis, let us define the following variables: B for the number of quantizers associated with each time-frequency coefficient, M the number of time-frames, N the number of frequency coefficients per frame (so that in total there are NM time-frequency locations), and finally n the size of the neighborhood of dependence (see Figure 3.3).

We can also give reasonable values for each of these parameters. For instance, $N = 20$ frequency subbands; audio files a few seconds long so that $M = 100$ time frames. The neighborhood of dependency, for overlapped ($\le 50\%$) representations, is then $n = 3N = 60$. To compute B, we need to estimate the maximum number of bits required for transparent coding. Assuming a step-size reduction factor $\Gamma = 0.9$ in (3.15), we have $\delta = 2|x|(0.9)^b$, where x is the transform coefficient. The maximum error incurred by a uniform quantizer of step-size δ is $\delta/2$, hence the ratio between reconstructed and reference coefficients is approximately:

$$20 \log_{10} \left(\frac{|x \pm \delta/2|}{|x|} \right) \approx 20 \log_{10}(1 + (0.9)^b) \tag{3.18}$$

Setting the right-hand side to 1 dB (just-noticeable threshold) and solving for b gives $b \approx 20$; for

a more realistic emulation of rate in a real coder.

Moore's more conservative estimate of 0.1 dB we obtain $b \approx 40$. Thus anywhere between 20 to 40 bits (for $\Gamma = 0.9$) may be required per coefficient for transparent coding. Let us use the optimistic $B = 20$ in the sequel. In all of the following methods, the dominating factor in time-complexity is the number of evaluations of the distortion function. We shall thus write complexity in the form $O(\phi)$, where ϕ is the number of distortion function evaluations.

Optimal methods

Brute Force: A brute-force evaluation of the distortion for all possible combinations of bit allocation is of complexity $O(B^{NM}) \sim 10^{2000}$ which is of course unfeasible.

Dynamic Programming/Lagrangian Relaxation: A classical approach to the problem of constrained optimization is the theorem of Everett [6], which shows, roughly stated, that a solution to the unconstrained minimization of the Lagrangian

$$L = R + \lambda D \tag{3.19}$$

is a solution to the constrained-distortion problem (3.9). What is remarkable about this theorem is its generality: essentially no restrictions are made on the rate or distortion measures, other that they be real functions on quite arbitrary sets. While the unconstrained optimization problem may seem easier than the constrained one, computational savings are unfortunately obtained only in the case that the rate and distortion measures are additively separable. Under such conditions, efficient algorithms are available, such as the one of Shoham, [31], the algorithm of Westerink [36], or the algorithm of Riskin [28]. Unfortunately, the distortion measure under consideration (3.17) is neither separable, nor even of additive-type. These algorithms thus do not apply.

It is, however, a distortion measure of the "max-type" admitting the weaker, dependent decomposition (3.13), with $F = \max_i(\cdot)$. The so-called "min-max" distortion has received less attention than the "min-average" problem, perhaps because it is not amenable to Lagrangian relaxation. Nevertheless, Schuster and Katsaggelos have formulated a trellis-search approach to both separable (independent) and non-separable (dependent) variants of the max-type optimization problem. Aggarwal [1] essentially lifted this procedure and applied it to the case of optimal selection of parameters in minimizing maximum NMR in an MPEG-IV coder. The method is *not* restricted to finding points on the rate-distortion convex hull, since no relaxation is used. The search will find *the* optimal rate-distortion parameters.

It is possible to apply this algorithm to our problem. First, we build a one-to-one association between quantizers and distortion, and hence a one-to-one association between distortion and rate. This can be accomplished by defining the distortion pattern in the (i, j) location as $D_{ij} = D_i$, for all j, with D_i as in (3.17). Let $D = K$ be our distortion target. The trick now is to redefine the rate function as:

$$b'(i,j) = \begin{cases} \infty, & D_{i,j} > K \\ b(i,j), & D_{i,j} \leq K \end{cases} \tag{3.20}$$

with the induced re-specification of the overall rate

$$R = \frac{1}{MN} \sum_{i,j} b'(i,j). \tag{3.21}$$

This cost function is additive — however, it is *not* separable because the redefinition (3.20) transfers the dependencies implicit with D into dependencies with b'! Nevertheless, trellis search remains applicable for non-separable functions, as long as they are of *additive*-type. For instance, the algorithm for additive dependent quantization of Ramchandran [26] can now be applied. The constraint on the maximum distortion is carried in the new rate-definition b'; any branch leading to a trellis point which produces distortion larger than the constraint is associated infinite cost; hence the minimizing path cannot pass through any such a state, as long as there exist parameters for which the distortion target can be satisfied.

Unfortunately, there does exists a downside: the computational complexity of such a procedure is $O(NM(B)^n)$ — exponential in the size of the neighborhood of dependency n [29]. Using our previous estimates, this comes out to approximately 10^{81} distortion evaluations — better than brute force — but still entirely unfeasible.

Sub-Optimal Incremental Algorithms

Given the enormous complexity of the algorithms presented in the previous section, we must forego rate-distortion optimality and make use of more heuristic procedures. One class of such methods are called "greedy" algorithms. They are useful when there is a way to order the quantizers in terms of increasing rate and quality, such as a sequence $q_1, q_2, q_3 \dots$ satisfying $\lim_{b\to\infty} q_b(x) = x$. In our case, the natural sequence is $q_{i,j,0}, q_{i,j,1}, q_{i,j,2}, \dots$ where $q_{i,j,b}$ is defined in Section 3.3.4. When this ordering is available, the parameter search is empowered with some sense of "direction", whereas for completely general definitions of rate R and distortion D there is no such structure.

Greedy Search: The standard greedy algorithm approaches the rate-distortion problem in the following way: beginning with an initial bit-allocation $b(i,j) = 0$ for all i, j, the algorithm finds the (i,j) location for which the bit increase $b(i,j) = b(i,j)+1$ results in a maximal decrease in distortion D. Computational complexity for this algorithm is upper-bounded by $O(B(NM)^2) \equiv 10^7$, which is large but feasible.

Unfortunately, while performing well for independent quantization problems, the greedy algorithm can fail to halt for dependent quantization problems [5]. In particular, allocation distributions can arise such that for *no* time-frequency location does the bit increase $b(i,j) = b(i,j) + v$ result in a decrease in distortion, for *any* positive integer v. This phenomenon is essentially the consequence of trying to minimize an irreducibly multivariate, non-separable distortion function by checking changes in the objective function only along the axial directions (1,0,0,...), (0,1,0,0,...), (0,0,1,0,0,...) etc. Indeed the very notion of separability implies a function that is "naturally" oriented along axial directions[7]. This explains why search directions parallel to the axes (such as occur in "single-coefficient" updates) work well for these types of functions. It is also why the greedy search tends to perform poorly or not at all for non-separable functions.

Reverse Allocation

The idea of a "reverse" greedy algorithm may have first been introduced by the present author in [5]. It was formulated in that paper under restricted conditions, for an excitation distortion measure of a specific type. However, the experimental results therein showed that reverse allocation could meet distortion targets at approximately 50% the rate of a suitably defined multi-coefficient (forward) greedy algorithm. We now provide a general formulation for reverse allocation, applicable for the bit-allocation of a wide range of distortion measures.

The basic idea is extremely simple and consists of the following: we first obtain, by any method, a bit allocation satisfying the distortion constraint without necessarily worrying about rate-optimality; this bit allocation is used as an initialization to a de-allocation algorithm, which successively removes bits until the distortion constraint is breached. More specifically, at each iteration a bit is removed from the location (i,j) which results in the smallest updated distortion. The process continues until the constraint $D < K$ is first breached; the last allocation for which the target is achieved is retained.

It is important to observe that while a greedy search driven by a non-separable distortion function is used for the *removal* of bits, it does not suffer from the same issues as a forward greedy algorithm. For instance, the halting problem does not occur here since a decrease in distortion in de-allocation—though generally unexpected—is a *positive* result, whereas a generally unexpected increase in distortion in the forward algorithm is a negative outcome. Indeed, the reversal of priorities in the inverse algorithm can transform the weaknesses of the forward greedy search into strengths in the reverse case.

[7] For instance, consider the additively separable function $f(x,y) = ax^2 + cy^2$ as opposed to the non-separable function $g(x,y) = ax^2 + bxy + cy^2$. The level sections of g are ellipses rotated with respect to the x and y axes

There are a variety of ways of obtaining the initializing bit distribution. In [5] an initializing bit allocation was found by utilising the fact that, under certain conditions, there exists a simpler non-spread measure that can over-bound the excitation distortion. A standard forward greedy algorithm minimizing the non-spread measure obtained the desired initialization. With general distortion measures, one does not always have the luxury of such a structure. There always exists one crude estimate, however: simply set $b(i,j) = B, \forall i,j$ with B sufficiently large. A more intelligent design is to use some type of forward multi-allocation algorithm:

Algorithm 3.1 (Initialization)
Given a distortion target K,

1. Set $b(i,j) = 0, \quad \forall i,j$
2. Compute the distortion pattern D_{ij} using bit allocation $b(i,j)$ for all i,j.
3. If $\max_{i,j} D_{ij} < K$ stop. If not, locate $(i^*,j^*) = \arg\max_{i,j} D_{ij}$.
4. Find the set $\mathcal{P}$ of all indices (i,j) such that $D_{i^*j^*}$ is a function of $q_{i,j}$.
5. Set $b(i,j) = b(i,j) + 1 \quad \forall (i,j) \in \mathcal{P}$.
6. Go to Step 2.

Step 4 of the algorithm involves finding the "dependency" neighborhood of the distortion point $D_{i^*j^*}$ (see Figure 3.3). When the coding process involves an overlapped representation with overlap O such that $0\% < O < 50\%$, this neighborhood is always given by $\mathcal{P} = \{(i,j) : i^* - 1 \leq i \leq i^* + 1\}$. The above initialization algorithm is guaranteed to converge as long as the distortion function D_i of (3.17) satisfies the continuity requirement

$$\lim_{x \to y} D_i(x,y) = 0 \tag{3.22}$$

and the quantizers are "well-designed" in the sense that $\lim_{b\to\infty} q_{i,j,b}(x) = x$.

We now give a formal description of the de-allocation process. In the process we shall introduce a complexity-scaling parameter that allows the user to trade-off time-complexity for rate-distortion optimality.

Algorithm 3.2 (De-allocation)
We assume that some initialization process has already arrived at a bit allocation $b(i,j)$ satisfying the distortion constraint K. Begin by partitioning the set of time-frequency locations into disjoint sets $A_k, 1 \leq k \leq L$. Let us define the function $T_b(i,j)$ as the updated distortion D that

occurs when the bit allocation $b(i,j)$ is replaced with $b(i,j)-1$ (only in the (i,j) location) when $b(i,j) \geq 1$ *and* when the updated distortion satisfies $D \leq K$. We define $T_b(i,j) = \infty$ if either $b(i,j) = 0$, or if the updated distortion $D > K$.

1. Compute $T_b(i,j)$ for all $(i,j) \in A_1$.
2. Find $(i^*, j^*) = \arg\min_{i,j} T_b(i,j)$.
3. Set $b(i^*, j^*) = b(i^*, j^*) - 1$ if $T_b(i^*, j^*) < \infty$.
4. Go to Step 1, and repeat Steps 1–3 for each of $A_2, A_3, A_k, \ldots, A_L$.
5. If at least one bit was removed through Steps 1–4, repeat Steps 1–5. If not, the algorithm terminates.

We shall call the concatenation of the Initialization Algorithm and De-allocation Algorithm by the title "Reverse Allocation Algorithm". It is a rather general procedure, which can be applied to any distortion function, separable or not, which satisfies the continuity property (3.22). The algorithm is guaranteed to halt in a finite number of iterations, for any distortion target K.

The partitioning of the time-frequency locations into sets A_k provides a way of trading off computational effort and rate-distortion optimality. The main point to the partitioning is in reducing the number of distortion function evaluations required before removing one bit. For instance, if $L = 1$ and one takes A_1 the entire set of indices, the algorithm must test every location in the file before removing a single bit. On the other hand, by choosing a fine partition so that $A = \max_k |A_k|$ is small, the algorithm is required at most to evaluate the distortion A times before attempting to remove a bit. Thus the latter algorithm can generally run up to NM/A times faster than the former. However, because the search-field in the latter is far more restricted than in the former, it will tend to remove relatively fewer bits before exceeding the distortion constraint.

The complexity of the Initialization Algorithm is no more than $O(B(NM))$, while the complexity for De-allocation is no more than $O(B(NM)^2)$. The combined complexity of Reverse Allocation is therefore no more than that $O(B(NM)^2)$ — the same as the standard greedy algorithm. Depending on how well localised the neighborhood of dependence is, a careful implementation of the partitioning $\{A_k\}$ can obtain $O(ABNM)$, so that it is even possible to have complexity $O(BNM)$ in some cases by using the finest partition A_k possible: each A_k a singleton.

The following table summarises the complexity of the allocation algorithms discussed.

Table 3.1 Complexity of Various Allocation Algorithms

Algorithm	Complexity	Typical No. of Evaluations
Brute-Force	$O(B^{NM})$	$\sim 10^{2000}$
Dynamic Programming	$O(NMB^{n})$	$\sim 10^{81}$
Greedy Algorithm	$O(B(NM)^{2})$	$\sim 10^{7}$
Reverse Allocation	$O(BNM)$ to $O(B(NM)^{2})$	$\sim 10^{5} - 10^{7}$

Chapter 4

Experimental Results

The investigations of Chapter 2 suggested that, conceptually at least, ED should quantitate distortion in a way closer to human audition than does NMR. Chapter 3 looked at the issues that arose when attempting to implement the new measure in transform coders, and in particular developed a new bit allocation algorithm for dependent, non-causal coding. In this final chapter, we integrate all of the findings of the past chapters to build a non-causal constrained-distortion ED and NMR coder, with the rate-distortion optimization driven by Reverse Allocation. The resulting coders are not complete, of course, but do provide the basis for a fair experimental comparison between the measures.

4.1 Phase and Magnitude Quantization

Time permits us only to investigate the performance of excitation distortion when the standard excitation pattern is used, as opposed to the Complex Excitation Pattern (Section 1.3). The resulting ED distortion measure is thus phase-blind. As a consequence, we will choose to design the quantizers in a magnitude-domain, with the assumption of perfect phase coding. More specifically, we define our domain of quantization to be the space of square-root power spectra $|X(\omega)|$, where $X(\omega)$ is the Fourier transform of the signal. This domain has the advantage that NMR retains the squared-error form it assumes in the Fourier domain; i.e. if $re^{i\theta}$ and $\tilde{r}e^{i\theta}$ are the unquantized and quantized magnitude spectra respectively, then:

$$\frac{N(\omega)}{M(\omega)} = \frac{|re^{i\theta} - \tilde{r}e^{i\theta}|^2}{M(\omega)} = \frac{|r - \tilde{r}|^2}{M(\omega)} \tag{4.1}$$

which is a structure not obtained, for instance, if quantization were to occur in the power spectral domain.

The fact that we quantize only in the magnitude domain does *not* mean that there is no phase distortion in the coded signal, however. This is due, once more, to the overlap-add coding process. No phase distortion exists when comparing a reference spectra to the intermediary quantized spectra of Fig. 3.2, but certainly phase distortion will exist in an end-to-end comparison between reference spectra and the final spectra *after* overlap-add. We simply use distortion measures that ignore phase effects.

4.2 Specifics of the Coding Structure

We shall work with audio sampled at 8000 Hz. The time-frequency decomposition is achieved using 30 ms time frames (240 samples/frame) overlapped by 50%, windowed by a square-root Hanning function, and transformed via DFT into a 121 point magnitude spectrum representation $x(i, j)$. For every frame, these coefficients are divided into 18 subbands of unit critical-bandwidth, which we index by k. Each group G_k of coefficients is given a product scalar quantizer parameterized by a single step-size δ_k. The step-sizes are allowed to attain values on the discrete set defined by (3.15), with $\Gamma = 0.9$, $K_{ij} = 2.01 \max_{G_k : i \in G_k} |x_{ij}|$.

Model of Excitation, Masking Threshold, Noise

Given a power spectrum, we calculate the total excitation pattern $E'(\omega_i)$ by first applying the excitation transform $\mathbf{U}$ (1.14), and then adding the internal noise $\epsilon(\omega_i)$ according to (1.7). The masking threshold associated with that power spectrum is given by $M(\omega_i) = s(\omega_i)E(\omega_i)$, with threshold factor $s(\omega_i)$ as in (1.5). We thus take care to use the *same* excitation transformation model for *both* NMR and ED. An important observation is that both the NMR and ED patterns are of *high-resolution* — they are computed on the same frequency resolution as the DFT, and not the resolution of the subbands. Finally, we define the noise variable $N(\omega_i)$ as the squared error $|x(i) - \tilde{x}(i)|^2$, passed through the middle outer ear filter of (1.1).

Rate-Distortion Optimization

The Reverse Allocation algorithm of Section 3.3.5 is applied to find, given a target K, the step-size parameters necessary to drive $D < K$. We define the overall rate of a coded file as the average empirical entropy of the quantization levels. This emulates the rate of a coder in which the quantization levels are Huffman coded. More specifically, the integer quantization level l of a coefficient x using step-size δ is given by:

$$l = \text{round}(x/\delta). \tag{4.2}$$

The specification of the step-sizes δ then define a time-frequency matrix of integer quantization levels. The empirical entropy $H(i)$ of each frequency bin i is computed according to:

$$H(i) = -\sum_{l} r(l,i)\log_2 r(l,i) \tag{4.3}$$

where $r(l,i)$ is the relative frequency of level l among all levels in bin i, over all time. The empirical rate R of a coded file is then defined as the average entropy over frequency:

$$R = \frac{1}{18}\sum_{i=1}^{18} H(i) \tag{4.4}$$

The ED distortion function of (3.1) is used with $p = \infty$, hence minimizing *maximum* excitation distortion (over frequency). Similarly, per frame, we define the NMR distortion function D_{NMR} as the maximum NMR over frequency:

$$D_{\mathrm{NMR}} = \max_{i} \frac{N(\omega_i)}{M(\omega_i)} \tag{4.5}$$

Given the definition of the *global* file distortion in (3.17) as the maximum distortion over all *time* frames, the coded files will be such that the NMR and ED patterns are no greater than the target K, in both time and frequency.

4.3 Matched-Rate Comparison

To compare coded files, the distortion targets are tuned until the files have the same empirical entropy. Our first test consisted of coding a male speech file. The ED function (3.1) was tested with a number of different parameters. As a notational convenience, let us use the following symbols: 1) ED — logarithmic excitation distortion, weighting function $W(\omega) = 1$, 2) ED_{W} — logarithmic distortion, weighting function $W(\omega)$ as in (2.29), 3) ED_f — excitation distortion with f the optimal approximation to loudness difference, $W(\omega) = 1$.

Four rate settings were fixed, distributed from approximately 0.65 bits/coefficient to 1.4 bits/coefficient. The latter files are in the high-quality/near-transparent regime, while the file at the lowest rate has significant distortion. In every case, an informal cadre of trained listeners agreed with the following quality ranking:

$$\mathrm{ED}_f = \mathrm{ED} > \mathrm{ED}_{\mathrm{W}} > \mathrm{NMR} \tag{4.6}$$

Our second test file consisted of a vocal quartet — qualitatively quite different from the male speech. Once again, we have found that ED_f produces a subjectively higher quality coded file than that of NMR, at all rates.

4.4 Conclusion and Future Work

The above experimental results, though small in number, are sufficiently encouraging to suggest that the class of excitation distortion measures holds promise in improving the quality of existing audio coders.

There exist many research projects, of which only the shallowest of forays have been undertaken in this thesis, that may now be fruitfully pursued. Among them are:

1. Perform a comprehensive set of listening tests between ED and NMR coded files, for a range of parameters k, p and weighting function $W(\omega)$.

2. Test ED when the Complex Excitation Power Spectrum is used, in lieu of the standard excitation. In this case, phase distortion can be incorporated and quantization performed in the DFT domain.

3. Investigate the performance of ED when an MDCT representation is used in conjunction with some binary phase coding algorithm.

4. Create a framework which allows quantization in the excitation domain *and* takes into account lapped transform representation (quantized excitation is not the same as reconstructed excitation).

5. Experimentally quantify the trade-off between computational complexity and rate-distortion optimality in the choice of partition A_k for Reverse Allocation.

6. Create a framework for dealing with *time-spread* excitation distortion, as well as time-spread CEPS, which takes into account lapped transforms.

Appendix A

Closed-form Expressions for Scale-invariant Measures

Given any symmetric distortion $D(\mathbf{x}, \mathbf{y})$ on vectors $\mathbf{x}$ and $\mathbf{y}$ one can define a scale-invariant version of the measure via

$$D'(\mathbf{x}, \mathbf{y}) \equiv \min_{\beta>0} D(\beta\mathbf{x}, \mathbf{y}) \tag{A.1}$$

Such measures are often useful for quality measurement schemes. The minimum of the above equation can be difficult to evaluate, however, depending on the nature of the original distortion function D. An iterative search for the minimum could be computationally expensive especially in analysis-by-synthesis processes where the distortion function must be called often. It would be useful to find closed-form solutions for the scale-invariant measure D'. We look at some particularly tractable cases when D is the NMR or ED function.

A.1 Scale-invariant NMR Distortion

Consider the following NMR function, ignoring phase, as a function only of the root power spectra $\mathbf{x}$ and $\mathbf{y}$ given by:

$$D(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} \left(\frac{(x_i - y_i)^2}{M_i} \right)^p \right)^{1/p} \tag{A.2}$$

The case $p = 1$ quantifies average NMR, $p = 2$ mean-square NMR, and $p = \infty$ maximum NMR. As discussed in Appendix B, the NMR function is not metrically symmetric in its arguments (since the masking threshold of the reference signal $\mathbf{x}$ is singled out). The scale-invariant version then must be defined carefully with the scale-factor applied to the reproduced signal $\mathbf{y}$, so that

$$D' = \min_{\beta>0} \left(\sum_{i=1}^{n} \left(\frac{(x_i - \beta y_i)^2}{M_i} \right)^p \right)^{1/p} \tag{A.3}$$

Now we attempt to find a closed-form solution for D'. This will consist in finding a closed-form solution for the minimizing scale-factor β as a function of the other parameters.

It suffices to minimize

$$J = \sum_{i=1}^{n} \frac{(x_i - \beta y_i)^{2p}}{M_i^p} \tag{A.4}$$

where the search is conducted over real $\beta > 0$

Taking the derivative, we obtain:

$$\frac{dJ}{d\beta} = \sum_{i=1}^{n} \frac{-2py_i(x_i - \beta y_i)^{2p-1}}{M_i^p} = 0 \tag{A.5}$$

It is evident if p is a positive integer, solving this equation is tantamount to finding the (real and positive) root(s) of a $2p - 1$ order polynomial. Closed-form solutions only exist for $p = 1$ (linear) and $p = 2$ (cubic polynomial). No closed-form solutions exist for the roots of quintic and higher polynomials. Let us give the solution explicitly for $p = 1$. In this case, the cost function is quadratic and the minimum is unique. With a little algebra, we have

$$\beta = \frac{\sum_{i=1}^{n} \frac{x_i y_i}{M_i}}{\sum_{i=1}^{n} \frac{y_i^2}{M_i}} \tag{A.6}$$

Since $\mathbf{x}$ and $\mathbf{y}$ are root power spectra, they are positive and $\beta > 0$. Therefore,

Theorem A.1. *The scale-invariant function minimizing average NMR is given by*

$$D'(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \frac{(x_i - \beta y_i)^2}{M_i(\boldsymbol{x})} \tag{A.7}$$

where

$$\beta = \frac{\sum_{i=1}^{n} \frac{x_i y_i}{M_i(\boldsymbol{x})}}{\sum_{i=1}^{n} \frac{y_i^2}{M_i(\boldsymbol{x})}}. \tag{A.8}$$

A.2 Scale-invariant ED with weighting function

The distortion function to be considered here is

$$D(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} w_i |\ln(x_i/y_i)|^p \right)^{1/p} \tag{A.9}$$

where $\mathbf{x} > 0$ and $\mathbf{y} > 0$ are the total excitation patterns.

Let us consider the case where p is a positive even integer. Then it suffices to minimize

$$J = \sum_{i=1}^{n} w_i (\ln(\beta x_i/y_i))^p \tag{A.10}$$

Now the problem of finding an optimal β proceeds identically to that of the NMR problem. We shall provide the final result, leaving the proof to the reader.

Theorem A.2.

1. *Closed-form expressions for the minimizing scale factor β in (A.10) are possible only with $p = 2$ and $p = 4$.*

2. *The scale-invariant function minimizing weighted mean-square ED ($p = 2$) is*

$$D'(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} w_i \left(\ln(\beta x_i/y_i)\right)^2 \tag{A.11}$$

where

$$\beta = \left[\prod_{i=1}^{n} \left(\frac{y_i}{x_i} \right)^{w_i} \right]^{1/\sum_i w_i} \tag{A.12}$$

Appendix B

Symmetry Properties of Distortion Measures

This appendix defines and investigates a number of symmetry properties of distortion functions. We consider distortion functions D defined between two power spectra $x(\omega)$ and $y(\omega)$. We will omit the frequency variable whenever convenient.

Typically a distortion function can be defined by introducing a distortion pattern

$$\mathcal{D}(\omega) = \mathbf{g}(x, y) \tag{B.1}$$

where $\mathbf{g}$ is a two-variable operator and then defining the distortion as some norm of the distortion pattern

$$D(x, y) = \|\mathcal{D}\| \tag{B.2}$$

Some examples of the distortion functions which we will study, falling into the above class are listed below:

1. $\mathcal{D} = (x - y)^2$ (squared-error)

2. $\mathcal{D} = \dfrac{(x - y)^2}{M(x)}$ (Noise-to-Mask Ratio)

   where $M(x)$ is the masking threshold of spectrum x.

3. $\mathcal{D} = |L(x) - L(y)|$ (Loudness difference)

where $L(u)$ is the loudness distribution of the spectrum u.

4. $\mathcal{D} = \dfrac{1}{1 - 10^{-k/10}} \dfrac{|[T(x)]^k - [T(y)]^k|}{\max\{[T(x)]^k, [T(y)]^k\}}$ (Relative loudness difference)

   where T is the transformation from spectrum to excitation pattern. This is the gain-invariant distortion function introduced by finding optimal gain-invariant approximations to loudness difference (Section 2.2.6).

5. $\mathcal{D} = \dfrac{T(x)}{T(y)} - \ln\left(\dfrac{T(x)}{T(y)}\right) - 1$ (Itakura-Saito Distortion on Excitations)

In each case, a distortion pattern can be defined by using $D(x, y) = \|\mathcal{D}\|$. A notable example for which the distortion is not defined this way is the audible distortion criterion (2.14), defined by

$$D(x, y) = \sum_{\mathrm{S_{aud}}} |\mathcal{D}| \tag{B.3}$$

where $\mathcal{D}$ is the gain-invariant distortion pattern fourth in the list above, and $\mathrm{S_{aud}}$ is the set of frequencies for which there is audible distortion (defined as regions where $\mathcal{D}$ exceeds the generalised JND).

Instead of using a norm, we can also use a general grouping function F, so that $D = F(\mathcal{D})$. We then require that F be increasing in each dimension.

B.1 Symmetry Properties

There are many notions of symmetry/asymmetry available. We discuss the following types: 1) metrical symmetry, 2) additive/subtractive distortion symmetry, 3) masking symmetry/asymmetry for tone-noise and noise-tone masking. The first two notions are available for all distortion functions, whereas the last property only makes sense in the context of distortion measures that are psychoacoustically motivated: i.e. take into account excitation or masking elements of human audition. Any distortion measure formulated with a distortion pattern in the excitation domain can potentially exhibit masking asymmetry. The noise-to-mask ratio criterion can potentially exhibit asymmetries in masking situations by adjusting the masking threshold $M(x)$ appropriately. From here we consider only the first two types of asymmetry, which make sense for all distortion measures.

B.1.1 Metrical Symmetry

A distortion measure is metrically symmetric if

$$D(x, y) = D(y, x) \tag{B.4}$$

The term “metric” is used because this property is the second property of metric spaces (the first being positivity and the third being the triangle inequality). The conceptual point of this property is that there is no a priori reference pattern, both x and y are on equal terms. Ultimately, it means that the distance from x to y is the same as the distance from y to x. The perceptual significance of this formal attribute is clear in many of the simple psychoacoustic experiments, for instance, evaluating JNDs in amplitude or frequency changes of sinusoids. Given two sinusoids s_1 and s_2, the distance between them shouldn’t depend upon which of them is used as the first variable, or in other words, which of them is used as the “reference”.

When realistic spectra are used, taken from speech or audio, it is usually possible for the human to identify the “undistorted”, reference signal, thus there may be some motivation to use asymmetric distortion functions, though it is not really clear if it is necessary.

The squared-error distortion function is metrically symmetric, as are the loudness difference and the relative loudness difference measures.

On the other hand, both the NMR and Itakura-Saito distortion measures are metrically asymmetric. They both single out a “reference” pattern x; NMR uses the masking threshold of x and not y.

For gain-invariant functions like the Itakura-Saito distortion and relative loudness difference, the measures are completely parameterized by a function $f(r)$, where r is the ratio $r = x/y$. In this case, the symmetry condition requires $f(r) = f(r^{-1})$, or that the functions exhibit symmetry about a vertical axis when the x-axis is plotted logarithmically. It is easy to see from the following figure that the Itakura-Saito distortion is metrically asymmetric and that the relative loudness difference is metrically symmetric.

B.1.2 Additive/Subtractive Distortion Symmetry

A distortion function can sometimes penalize additive distortion more than subtractive distortion, or vice versa. By the terms “additive” and “subtractive” distortion we mean power-spectrum (or excitation-spectrum) additive and subtractive. Hence n is an additive distortion if the distorted spectrum is $y = x + n$ and subtractive if $y = x - n$.

A distortion function penalizes additive distortion more than subtractive distortion if

$$D(x; x + n) > D(x; x - n), \quad x > n \tag{B.5}$$

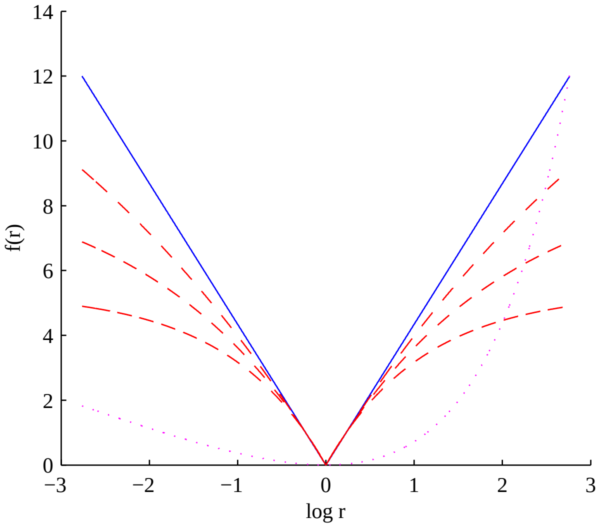


Fig. B.1 Solid: logarithmic distortion function; Dashed: Relative loudness difference for different values of k; Dotted: Itakura-Saito distortion

and vice versa if the inequality sign is reversed. It penalizes additive and subtractive distortion symmetrically if

$$D(x; x+n) = D(x; x-n), \quad x > n \tag{B.6}$$

The Noise-to-Mask Ratio (and squared error, a special case) is an example of a distortion function which treats additive and subtractive distortion equally, since

$$\frac{(x-(x-n))^2}{M(x)} = \frac{n^2}{M(x)} = \frac{(x-(x+n))^2}{M(x)} \tag{B.7}$$

The Loudness Difference and Relative Loudness Difference both penalize additive distortion less than subtractive distortion. To see this, it is easiest to recast the distortion measures in the excitation domain with corresponding excitation patterns u and v for power spectra x and y, where the distortion patterns can be distilled into the simpler (and for the second measure, more general) forms:

1. $\mathcal{D}(u, v) = |h(u) - h(v)|, \qquad$ (Loudness difference)

   where h is an increasing compressive nonlinearity

2. $\mathcal{D}(u, v) = f(u/v), \qquad$ (General gain-invariant symmetric distortion)

where $f(r)$ is increasing when $r > 1$ and also the metrical symmetry condition $f(r) = f(r^{-1})$ is satisfied.

Since these distortion measures are naturally formulated in the excitation domain, we then ask whether a distortion penalises excitation-additive or excitation-subtractive distortion (which is generally the same concept as power-additive and power-subtractive distortion, and exactly so if the excitation transform is linear). Thus we say that a distortion function defined in the excitation domain penalizes excitation-additive distortion less than excitation-subtractive distortion if $D(u, u+e) < D(u, u-e)$ and so on.

Now consider the case of Loudness Difference. Since h is a compressive nonlinearity,

$$h(u+e) - h(u) < h(u) - h(u-e). \tag{B.8}$$

Both sides of this equation are positive, since h is an increasing function, and so

$$|h(u+e) - h(u)| < |h(u) - h(u-e)| \tag{B.9}$$

$$\mathcal{D}(u, u+e) < \mathcal{D}(u, u-e) \tag{B.10}$$

Since one distortion pattern is less than the other everywhere, taking the operator F (which increases in each dimension) on each side preserves the inequality sign, and thus Loudness difference penalizes additive distortion *less* than subtractive distortion.

The general symmetric gain-invariant function, of which Relative Loudness difference is a special case, also penalizes additive distortion less than subtractive distortion. To show this, begin with the inequality

$$\frac{u}{u-e} > \frac{u+e}{u}, \quad u > e \tag{B.11}$$

which is easily verified by cross-multiplying. Now, both sides of the inequality are larger than 1, and since f is increasing when its argument is larger than 1, then

$$f\left(\frac{u}{u-e}\right) > f\left(\frac{u+e}{u}\right), \quad u > e \tag{B.12}$$

which implies, from the symmetry condition $f(r) = f(r^{-1})$ that

$$f\left(\frac{u}{u-e}\right) > f\left(\frac{u}{u+e}\right) \tag{B.13}$$

$$\mathcal{D}(u, u-e) > \mathcal{D}(u, u+e) \tag{B.14}$$

Thus the Relative Loudness Difference, and also the dB-distance, penalize excitation-additive distortion less than excitation-subtractive distortion.

The Itakura-Saito Distortion also penalizes additive distortion *less* than subtractive distortion. The proof runs as follows: we need to show that $D(u, u+e) < D(u, u-e)$, or that

$$\frac{u}{u+e} - \ln\left(\frac{u}{u+e}\right) - 1 < \frac{u}{u-e} - \ln\left(\frac{u}{u-e}\right) - 1, \quad u > e \tag{B.15}$$

After some algebra, this is so if and only if

$$f(u) = \frac{u}{u-e} - \frac{u}{u+e} + \ln(u-e) - \ln(u+e) > 0, \quad u > e \tag{B.16}$$

Checking the derivative, we see that:

$$f'(u) = -\frac{2e}{(u-e)^2} + \frac{2e}{(u-e)(u+e)} = -\frac{4e^2}{(u-e)^2(u+e)} < 0 \tag{B.17}$$

Thus f is always decreasing. It is easily checked that $f(u) \to 0$ as $u \to \infty$, which together with the fact that f is continuous and decreasing on $u \in (e, \infty)$ implies that $f > 0$ on that interval, proving the statement.

B.2 Summary

1. Squared-error distortion measure is *metrically symmetric*, and penalizes both power-additive and power-subtractive distortion *equally*

2. Noise-to-Mask Ratio distortion measure is *metrically asymmetric*, and penalizes both power-additive and power-subtractive distortion *equally*

3. Loudness Difference measure is *metrically symmetric*, and penalizes excitation-additive distortion *less* than excitation-subtractive distortion

4. Relative Loudness Difference measure is *metrically symmetric*, and penalizes excitation-additive distortion *less* than excitation-subtractive distortion

5. Itakura-Saito Distortion measure on Excitations is *metrically asymmetric*, and penalizes excitation-additive distortion *less* than excitation-subtractive distortion

Finally, if the excitation transformation T is *linear*, then items 3, 4 and 5 hold also with excitation-additive and excitation-subtractive replaced with power-additive and power-subtractive distortion. If the transformation is not linear, then the situation is much more complicated. For instance, if we allow level-dependence of the auditory filters, increasing the amplitude of a sine wave generally causes a more significant increase in the slopes of the upper excitation pattern than a power decrease of the same magnitude in the sine wave. Because of the nonlinear level-dependence of the excitation transformation, it is possible in this case for the additive power distortion to significantly change the excitation pattern in the upper slopes in this case, so much so that the *overall* distortion is higher for power-additive distortion than power-subtractive distortion.

References

[1] A. Aggarwal. “Towards Weighted Mean-Square Error Optimality of Scalable Audio Coding”. Ph.D. thesis, University of California: Santa Barbara, 2002.

[2] J. Beerends, J. Stemerdink, “A Perceptual Audio Quality measure Based on a Psychoacoustic Sound Representation”, *J. Audio Engineering Soc.*, Vol. 40, No. 12, pp. 963–978, 1992.

[3] R. Blandon, B. Lindblom, “Modeling the judgement of vowel quality differences”, *J. Acoust. Soc. Am,*, 69, pp. 1414–1422, 1981.

[4] R. Der, P. Kabal, W.-Y. Chan, “Towards a New Perceptual Coding Paradigm for Audio Signals”, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing* (Hong Kong), I-824–I-827, April 2003.

[5] R. Der, P. Kabal, W.-Y. Chan. “Bit Allocation for Frequency and Time Spread Perceptual Coding”. *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing.* (Montreal, QC), pp. IV-201-IV-204, May 2004.

[6] H. Everett. “Generalised Lagrange Multiplier Method for Solving Problems of Optimal Allocation of Resources”. *Operations Research.* Vol. 11, Issue 3, pp. 399–417, 1963.

[7] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Press, 1992.

[8] B. R. Glasberg, B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data”, *Hearing Research*, 47, pp. 103–138, 1990.

[9] B. R. Glasberg, B. C. J. Moore. “A Model of Loudness Applicable to Time-Varying Sounds”, J. Audio Eng. Soc., vol. 50, pp. 331-342, May 2002.

[10] R. Gray, A. Buzo, A. Gray, W. Matsuyama. “Distortion Measures for Speech Processing”. *IEEE Trans. on Acoustics, Speech and Signal processing.* Vol. 28, pp. 367–376, Aug. 1980.

[11] R. Gray, D. Neuhoff. “Quantization”. *IEEE Trans. on Information Theory.* Vol. 44, pp. 2325–2383, Oct. 1998.

[12] J. L. Hall, “Asymmetry of masking revisited: Generalization of masker and probe bandwidth”, *J. Acoust. Soc. Am.*, 101 (2), pp. 1023–1033, Feb. 1997.

[13] M. Hauenstein, N. Gortz. “On the Application of a Psychoacoustically Motivated Speech-Quality Measure in CELP Speech-Coding”. *Proc. European Signal Processing Conf.* (Rhodes, Greece), pp. 1421-1424, Sept. 1998.

[14] ITU-R, Geneva, “Recommendation BS.1387-1, Methods for Objective Measurements of Perceived Audio Quality”, Nov. 2001.

[15] J. D Johnston, “Transform Coding of Audio Signals Using Perceptual Noise Criteria”, *IEEE J. Selected Areas Commun.*, Vol. 6, No. 2, pp 314–323, Feb. 1988.

[16] R. Lutfi. “Additivity of simultaneous masking”. *J. Acoust. Soc. Am.*, 73, pp. 262–267, 1983.

[17] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Macmillan, 1977.

[18] B. C. J. Moore, B. R. Glasberg, “Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns”, *Hearing Research*, 28, pp. 209–225, 1987.

[19] B. C. J. Moore, “Masking in the Human Auditory System”, *Collected Papers on Digital Audio Bit-Rate Reduction*, Audio Engineering Society, 1996.

[20] B. C. J. Moore, “A Model for the Prediction of Thresholds, Loudness, and Partial Loudness”, *J. Audio Engineering Soc.*, Vol. 45, No. 4, pp. 224–239, April 1997.

[21] H. Najafzadeh. *Perceptual Coding of Narrow-band Audio Signals.* Ph.D. Thesis, Dept. Electrical and Computer Engineering, McGill University, 2000.

[22] A. Ortega, K. Ramchandran. “Rate-Distortion Methods for Image and Video Compression”. *IEEE Signal Proc. Magazine*, pp. 23–50, Nov. 1998.

[23] K. Paliwal, B. Atal, “Efficient Vector Quantization of LPC parameters at 24 bits/frame”, *IEEE Trans. Speech, Audio Processing*, Vol. 1, No. 1, pp. 3–14, 1993.

[24] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens. “A New Psychoacoustical Masking Model for Audio Coding Applications”. *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing.* (Orlando, FL), pp. 1806-1808, May 2002.

[25] C. J. Plack, B. C. J. Moore. “Temporal Window Shape as a Function of Frequency and Level”, J. Acoust. Soc. Am., vol. 87, pp. 2178–2187 (1990).

[26] K. Ramchandran, A. Ortega, and M. Vetterli, “Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders,” *IEEE Trans. Image Processing*, vol. 3, pp. 533-545, Sept. 1994.

[27] P. Rao, R. van Dinther, R. Veldhuis, A. Kohlrausch, “A measure of predicting audibility discrimination thresholds for spectral envelope distortions in vowel sounds”, *J. Acoust. Soc. Am.*, 109 (5), pp. 2085–2097, May 2001.

[28] E. Riskin. “Optimal Bit Allocation via the Generalised BFOS Algorithm”. *IEEE Trans. on Information Theory.* Vol. 37, pp. 400–402, Mar. 1991.

[29] G. M. Schuster, G. Melnikov, A. K. Katsaggelos. “A review of the minimum maximum criterion for optimal bit allocation among dependent quantizers”. *IEEE Trans. Multimedia*, Vol. 1, pp. 3–17, Mar. 1999.

[30] D. Sen, W. H. Holmes, “Perceptual enhancement of CELP speech coders”, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing* (Adelaide), pp. II-105–II-108, April 1994.

[31] Y. Shoham, A. Gersho. “Efficient Bit Allocation for an Arbitrary Set of Quantizers,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 36, pp. 1445–1453.

[32] T. Thiede, “Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank”, Ph.D. thesis, Fachbereich Electrotechnik, Technical University Berlin, 1999.

[33] S. Vakil. *Gaussian Mixture Model Based Coding of Speech and Audio.* Masters Thesis, Dept. Electrical and Computer Engineering, McGill University, 2004.

[34] R. Veldhuis, “Bit Rates in Audio Source Coding”, *IEEE J. Selected Areas Commun.*, Vol. 10, No. 1, pp. 86–96, Jan. 1992.

[35] S. Wang, A. Sekey, A. Gersho, “An Objective Measure for Predicting Subjective Quality of Speech Coders”, *IEEE J. Selected Areas Commun.*, Vol. 10, No. 5, pp. 819–829, June 1992.

[36] P.H. Westerink, J. Biemond, D. E. Boekee. “An optimal bit allocation algorithm for sub-band coding” *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, I-757–I-760, April 1988.

[37] E. Zwicker and H. Fastl, *Psychoacoutics: Facts and Models*, Springer-Verlag, second edition, 1999.