Proc IEEE Int. Conf. Acoustics, Speech, Signal Processing (Hong Kong), pp. V-457-V-460, April 2003

Towards a New Perceptual Coding Paradigm for Audio Signals

Ricky Der[1] *Peter Kabal*[1] *Wai-Yip Chan*[2]

[1] Electrical & Computer Engineering
McGill University
Montreal, Quebec H3A 2A7

[2] Electrical & Computer Engineering
Queen's University
Kingston, Ontario K7L 3N6

Abstract

A new frequency domain approach to coding audio signals is introduced. The bit assignment strategy is aimed at reducing the perceived loudness difference between the original signal and the coded signal. As such it uses perceptual effects (spread excitation patterns), but does not directly invoke masking results. At low bit rates, examples coded with the new approach sound better than a more traditional bit allocation based on noise-to-mask ratio.

1 Introduction

This paper introduces a new approach to coding general audio signals. The new coding paradigm uses features of human sound perception, but does not directly utilise masking results. Indeed, it is argued that conventional approaches which use masking results to code audio signals are inappropriate, based on the observations that (1) modelling coding distortions, particularly large ones, as additive (uncorrelated) noise is inconsistent with the decoded signal, (2) the listener hears the decoded (and distorted) signal so that exploiting masking effects based on the original signal could be misleading, and (3) the masking threshold does not readily lend itself to measuring the amount of audible distortion (which some view as supra-masking threshold distortion) and therefore the distortion cannot be properly minimized.

2 Masking and Loudness Patterns

Perceptual masking is classically used to describe the inaudibility of weaker sounds (the maskee) which are nearby louder sounds (the masker). Masking is measured by determining the largest level of the maskee that can be *added* to the masker before the maskee becomes audible. In the coding context, it is argued that the original signal is the masker and the coding distortion is the maskee. However, the coding distortion is neither additive nor uncorrelated with the signal. In fact, 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,

$$\sigma_{\hat{x}}^2 = \sigma_x^2 - \sigma_e^2. \qquad (1)$$

This result for optimal quantizers also applies component by component for vector quantizers. 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 independent signals. It is also inconsistent with physical theories of masking, where the stronger signal masks a target due to neural swamping. In reality, quantization noise evokes no neural activity. To make the point clear, let us extremize the thought experiment to very low rates, where many spectral regions are reproduced as zero. Here, the distortion is equal to the signal itself, and conventional wisdom claims that the original signal masks a noise equivalent to itself. Of course, there is no masking because no sound reaches the ear. The concept of masker and noise maskee is unsuited for describing the perceptual effects in these regions.

Our approach to the problem begins with the just noticeable variation definition of Zwicker [1]. He stated that excitation patterns (putative patterns of physical activity along the basilar membrane) are perceptually indistinguishable when differing by less than 1 dB. In this paper, we focus on two excitation patterns: that of the original signal and that of the coded signal. We strive to minimize the loudness difference of the patterns. This approach subsumes conventional masking concepts (if the distortion is actually an additive value), and also the more general coding problem (when the distortion is not additive). By minimizing the difference between excitations (actually the difference in loudness), we have a mechanism which tells us how best to allocate resources (bits in a coding context).

This new viewpoint will entail a form of analysis-by-synthesis. For each candidate bit allocation, the excitation pattern of the coded signal is calculated ("synthesized") and compared to the target excitation pattern. From a set of candidate allocations, the one corresponding to the excitation most closely matching the target is selected.

2.1 Models of Loudness

A variable of great interest, loudness has been at the fulcrum of much psychoacoustic investigation dating back to the work of Fletcher in the 1930's. Whereas an excitation distribution is designed to predict the physical activity of hair cells along the basilar membrane, a loudness distribution models the further nonlinear transformation of intensity to perceptive strength: a type of "neural excitation". Such a neural excitation is presumed to be directly proportional to perceived strength. The level-transformed excitation is termed the *specific loudness pattern*, and gives loudness as a function of (tonal) frequency, in the unit of sones/Bark. A specific loudness pattern accounts, then, for both the non-ideal frequency selectivity of the ear, as well as the compressive nonlinear relationship between level and psychoacoustic intensity [1].

In this work, we use Zwicker's well-known model of loudness. While not a full model (it does not include the effects of modulation and beating), it has been applied successfully to the prediction of parameters in objective evaluators of audio quality such as PEAQ (Perceptual Evaluation of Audio Quality) or PAQM (Perceptual Audio Quality Measure). Zwicker did not provide analytic formulae for computing excitations: instead we use a standard (modern) procedure for evaluating these intermediary distributions, based on the FFT model of PEAQ [2].

The steps (Fig. 1) involved are (1) window the data, (2) DFT of the windowed data, (3) filter with the outer/inner ear response, (4) group frequencies into (partial) critical bands, (5) apply frequency and level dependent spreading, and (6) apply time smearing.

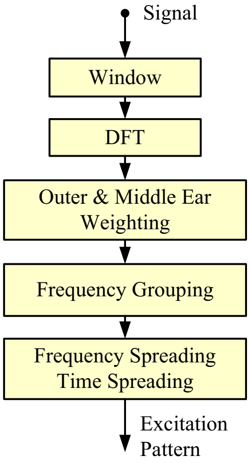
Signal
Window
DFT
Outer & Middle Ear Weighting
Frequency Grouping
Frequency Spreading Time Spreading
Excitation Pattern

Fig. 1 Processing to obtain excitation patterns

The final step from excitation to loudness invokes a general hypothesis in psychology known as Steven's law. The observation conjectures that perceived intensity is a power function of physical magnitude. Combining this with the boundary condition of absolute hearing threshold, Zwicker's equation for computing loudness density reads [1]:

$$L(f) = L_0\Big(\frac{E_{TQ}(f)}{s(f)E_0}\Big)^k \left[\Big(1-s(f)+\frac{s(f)E(f)}{E_{TQ}(f)}\Big)^k - 1\right]. \quad (2)$$

The exponent k has a nominal value of 0.23, $L_0 = 0.068$, E_0 is a reference excitation, $E(f)$ the excitation pattern, $E_{TQ}(f)$ the excitation due to a sinusoid at hearing threshold, and $s(f)$ a threshold factor [3]. In line with recent work, we ascribe part of the traditional absolute threshold curve to internal noise and part to middle-outer ear loss.

2.2 Loudness Distortion

The above gives a technique for computing the absolute loudness, as a function of tonal frequency, for any power spectrum. If a specific loudness distribution is the ultimate indicator of perceptual magnitude, then it is natural to introduce the linear loudness difference pattern:

$$L_{\text{diff}}(f) = L_S(f) - L_R(f), \quad (3)$$

where $L_S(f)$ is the loudness of the original signal and $L_R(f)$ is the loudness of the reconstructed signal. It is also natural to introduce the family of L^p norms on the difference space as measures of total distortion:

$$D_T(S,R) = \left(\int |L_S(f) - L_R(f)|^p df\right)^{1/p}. \quad (4)$$

An alternative strategy is to minimize the maximum loudness difference:

$$D_M(S,R) = \max_f |L_S(f) - L_R(f)|. \quad (5)$$

It should be emphasized that while the above equations do not explicitly make reference to masking, our distortion criteria do make an implicit use of the concept; indeed the very construction of excitation patterns relies on masking effects. Spectral components of a signal may mask one another: we eschew only the concept of masked quantization noise, not masking itself.

3 Perceptual Coding

A transform coding framework is used to compare the new distortion criterion with a traditional noise-to-mask ratio metric. The input signal is audio sampled at 8 kHz, and windowed (square-root Hanning window of length 30 ms) with 50% overlap. Coding is performed in the (complex) Discrete Fourier Transform (DFT) domain, with bands uniformly spaced on a Bark scale. The DFT bins are grouped into 24 bands, each of approximately 0.75 Bark width. The number of components per band (VQ dimensionality where a complex coefficient counts as two components in the real-imaginary representation) ranges from 4 (low bands) to 22 (high bands). Shape vector quantizers were trained off-line on a variety of audio signals, and designed to minimize the mean-square quantization error for a particular band at a given codebook size. The gains are left unquantized in this experiment. Quantized coefficients are then converted back to the time domain and the quantized signal reconstructed via overlap-add techniques (using a square-root Hanning window).

3.1 Coding Based on Perceptual Loudness Difference

A schematic diagram illustrating the perceptual loudness difference (PLD) coding structure can be found in Fig. 2. Bit assignment and quantization of the DFT coefficients are performed within an analysis-by-synthesis loop, without recourse to any rate-distortion performance model. Bits are allocated incrementally to minimize the loudness distortion as calculated from actual quantization results.

We remark in passing that there are in effect two distortion measures: a "local", within-band mean-square error measure used in the codebook search, and a global, across-frequency loudness (magnitude) measure for codebook selection. Both the total (Eq. 4) and maximum (Eq. 5) loudness difference criteria were tested in our simulations. At each iteration, the former allocates one bit to the quantizer which results in the largest decrease in overall loudness distortion; in the latter case, a bit is allocated to the band

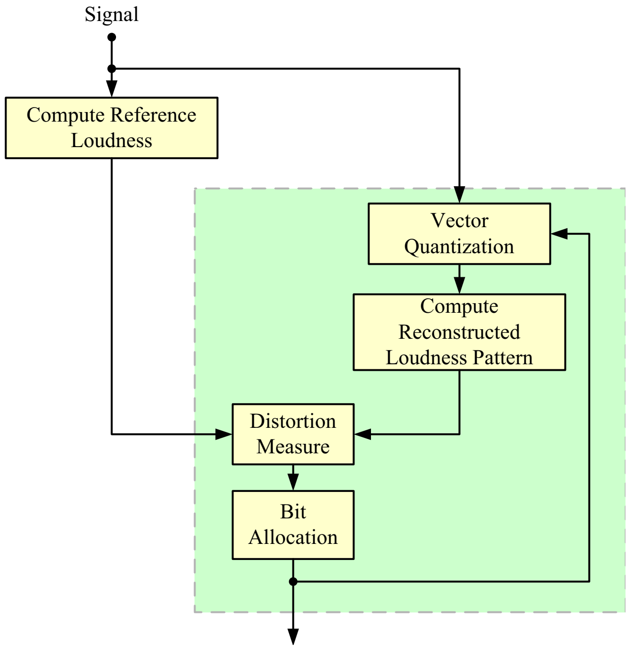
Signal
Compute Reference Loudness
Vector Quantization
Compute Reconstructed Loudness Pattern
Distortion Measure
Bit Allocation

Fig. 2 Analysis-by-synthesis loop

with the largest loudness difference. The process continues until a fixed bit quota is reached.

Overall, the incremental bit allocation strategy is “greedy” and suboptimal. One could alternatively adopt an operational rate-distortion optimization strategy, via dynamic programming, or approximated with a Lagrangian search [5].

3.2 Coding Based on Noise-to-Mask Ratio

The structure here is essentially identical to the framework in Fig. 2, with reference loudness replaced with a masking threshold, and the noise pattern in lieu of a reconstructed loudness pattern. Noise is the standard squared coding error passed through the middle/outer ear filter and frequency grouped.

The computation of the masking threshold begins with the excitation patterns as delineated above. Some differences apply: the internal noise threshold is applied before frequency spreading, and an additional normalisation step, suggested by Johnston [4], is applied prior to time-smearing. This last step is an attempt to deconvolve the excitation pattern back to an unspread power spectrum domain. We shall see that the poor approximation of normalisation to deconvolution is a major deficiency in the approach.

From normalized excitation distributions, a frequency-dependent masking offset is subtracted to obtain the masking threshold. In some models the masking offset is made a function of signal characteristics (for instance, tonality); in PEAQ, this distinction is not made — the decomposition into fractional critical bands tends to compensate for the difference.

Codebook selection is performed by observing a band’s noise-to-mask ratio (NMR). This is the approach that has been used by Johnston [4] and others. In one scenario, the allocation process employs tabulated average distortion functions; the masking level does not track the reproduced signal spectrum and it is implicitly assumed that the reproduced signal spectrum is close to the original. We use the better approach of measuring the actual distortion, dispensing with rate-distortion models. Bits are assigned one at a time to the band with the largest NMR.

4 Experimental Results

As a prelude, we first include an illustrative example of how NMR can fail to properly allocate bits. Consider a pair of tones at 941 Hz and 1633 Hz. The loudness curves for the original signal and the coded signal using Perceptual Loudness Difference and NMR methods are displayed in Fig. 3. With a ration of 16 bits, PLD correctly allocates 8 bits to each sinusoid. The NMR approach, however, wastes 3 bits in bands 5 and 6 (approximately 450 Hz) — regions with relatively little energy, allocating only 5 bits to the lower sinusoid.

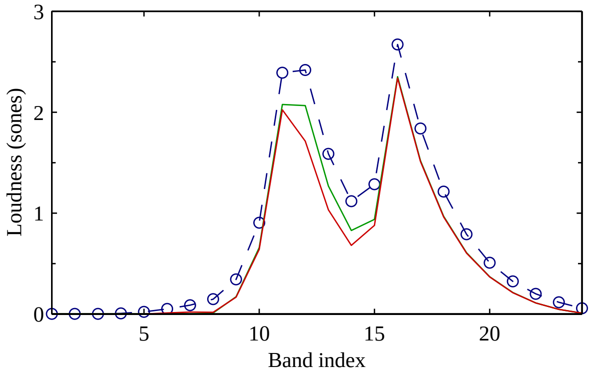

Fig. 3 Loudness curves for two tone signal using PLD and NMR bit assigments. Top curve: original signal, middle: PLD, bottom: NMR

The discrepancy is in fact caused by the lack of proper deconvolution: a masking threshold is computed in the spread domain, the noise in an unspread one, resulting in a cross-domain comparison. The main lobe of the windowed power spectrum is thinner than the corresponding lobe of the spread masking threshold, creating spurious hills at low and high frequencies of an NMR graph. In contrast, comparison is explicitly performed in a spread loudness domain with PLD: the deconvolution problem does not exist. This illustrates the danger associated with any method not operating entirely within a spread regime.

More extensive testing was performed with audio signals. The bit ration per frame was kept deliberately low to introduce audible distortion. The qualitative characteristics of NMR and PLD coded signals are quite different: in the former, the main distortion consists in musical noise, whereas a form of frequency smearing occurs in the latter. The authors found the latter disturbance less annoying. A typical example is displayed in the spectrograms of Fig. 4, consisting of a segment of unaccompanied female singing. Though not completely enlightening in a perceptual sense, a spectrogram can at least give an indication of bit allocation patterns. The musical noise of NMR is visually apparent from the tell-tale dominance of spectral collusions in the graph. In contrast, a more consistent texture of sound is produced with the *maximum* PLD method, though at the expense of a degradation in harmonic structure. Generally speaking, however, the result is a more natural reconstruction. The *integrated* PLD method (with $p = 1$) coded the signal at quality very similar to the maximum PLD criterion: once more without musical noise but with a slightly finer reproduction of low-frequency sinusoidal components.

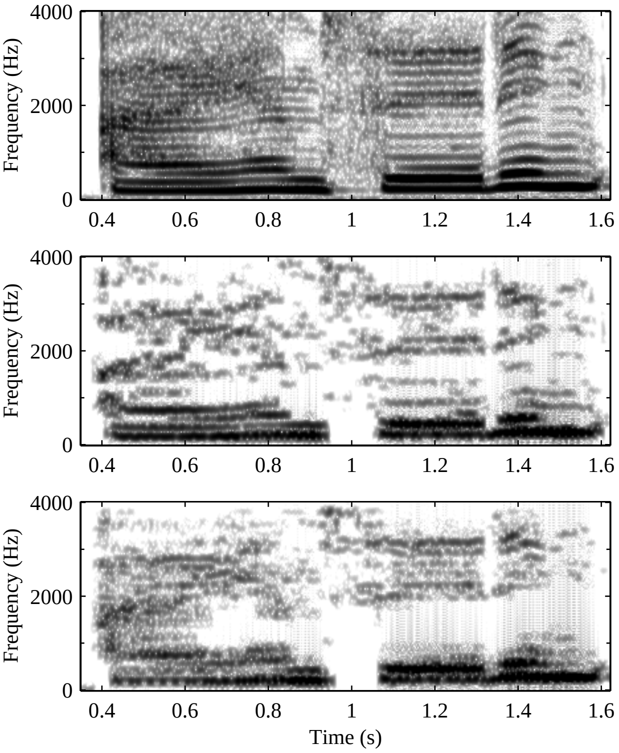

Fig. 4 Spectrograms. Top: original uncoded; Middle: NMR bit assignment (36 bits/frame); Bottom: maximum PLD bit assignment (36 bits/frame, $k = 0.23$)

We have found that the bit allocation is critically dependent on the exponent of Eq. (2). Zwicker tuned this exponent to predict the loudness of uniform exciting noise, with the result that harmonic components tend to be underestimated. Indeed, the exponent is crucial to controlling the excitation compression factor: $k = 0$, gives a uniform bit allocation across all frequencies, whereas an exponent of 1 is equivalent to minimising the difference between excitation values. Increasing the exponent tends to improve the integrity of harmonic components, reducing frequency smearing, but also introduces some musical noise. Nevertheless, we found a value of $k = 0.5$ highly satisfying for this particular example; the result was inarguably superior to the NMR-coded file.

5 Discussion and Future Work

As mentioned above, the NMR coding model described results in a cross-domain comparison. This can be avoided by omitting the (approximate) deconvolution step and instead spreading the noise to obtain a "(noise excitation)-to-masking" ratio. While still dealing with the illusory concept of quantization noise, the scheme is an improvement on the standard method (see [6] for details).

The uncertainty surrounding the loudness exponent value can be recast as an uncertainty regarding the correct form of the distortion function in Eq. (4). We remark here that the criterion is not gain-invariant, in the sense that the function changes if both signal and reconstruction are multiplied by the same constant. A loudness ratio is the obvious solution but does not possess intuitively appealing properties at low levels.

The role of phase has been ignored thus far: loudness is a measure of perceptual magnitude, although the local mean-square error distortion ensures that the phase will be (approximately) preserved. A complete characterization of tonal complexes in general requires some notion of relative phase, at least for frequencies under 5 kHz. Even with a phase distortion model, however, there are difficult questions as to its relative contribution, in conjunction with loudness, to overall distortion.

The complexity of an analysis-by-synthesis search can pose problems for real-time applications. Moreover, there is a rate cost incurred in sending the bit allocation as side-information. Both these issues can be alleviated by using rate-distortion curves for the bit allocation. There is, however, a quality degradation from using average statistics; even more, it may be difficult or impossible to produce a set of independent rate-distortion functions due to the non-linear and non-local character of spread loudness.

We have argued that masking effects as conventionally employed in a noise-to-mask distortion criterion are inappropriate, especially at low rates. An analysis-by-synthesis coding structure aiming to minimize the difference in specific loudness patterns was introduced, allowing for the computation of supra-threshold magnitude distortion. Testing produced distinctly different, and superior, results compared to traditional techniques, and suggests the method merits further thought and investigation.

References

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

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

[3] T. Thiede, *Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank*, Ph.D. thesis, Technical University Berlin, 1999, (`http://www.nue.TU-Berlin.DE`).

[4] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal Selected Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.

[5] 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.

[6] C. Cave, *Perceptual Modelling for Low-Rate Audio Coding*, M.Eng. thesis, McGill University, 2002, (`http://www.TSP.ECE.McGill.CA`).