Proceedings of the 1995 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, October 15-18, Mohonk Mountain House, New Paltz, New York

OBSERVATIONS ON AUDITORY EXCITATION AND MASKING PATTERNS

Stephen Voran

Institute for Telecommunication Sciences, U.S. Department of Commerce
NTIA/ITS.N3, 325 Broadway, Boulder, Colorado 80303, USA
sv@bldrdoc.gov

ABSTRACT

Excitation patterns and masking patterns are used extensively in perceptual audio coders and quality assessment algorithms. Numerous algorithms for calculating these patterns have been proposed. This paper provides comparisons among the patterns generated by several of these algorithms. The comparisons are based on audio program material, rather than tones and noise. Explored areas include synthesis functions, spreading functions, masking indices, tonality measures, and the treatment of the absolute threshold of hearing. Several mathematical relations are provided to characterize observations in these areas. Patterns from simpler algorithms are considered as approximations to patterns from more complex algorithms, and the approximation error is characterized. Results may be useful to those who apply auditory excitation or masking patterns in their work.

1. INTRODUCTION

Auditory excitation patterns seek to describe the auditory stimulation caused by an audio signal. Auditory masking patterns attempt to describe the hearing threshold for a subject exposed to a given audio signal. Each pattern is a scalar function of a frequency variable. The importance and utility of these patterns are clear from their extensive applications in audio coding and quality assessment, some of which can be found in [1-8]. Numerous algorithms that calculate approximations to these patterns have been proposed, and some of these are found in [1-11]. The complexities of these algorithms cover a wide range.

The algorithms are generally based on detection or "masking" experiments that characterize subjects' ability to detect a secondary auditory stimulus (target) in the presence of a primary auditory stimulus (masker). The maskers and targets are usually tones or bands of noise. The measured tone and noise masking relationships are typically applied individually to the analyzed components of an audio signal, and the results are synthesized to generate an excitation or masking pattern for that signal. Different types of masking experiments, differing interpretations of results, and differing analysis and synthesis techniques have contributed to the range of proposed excitation and masking algorithms.

When tone or noise signals are considered, the differences between these algorithms are fairly obvious. For general audio signals, the differences are less clear. This paper provides comparisons among the patterns generated by several excitation and masking algorithms operating on audio program material. Patterns from simpler algorithms are considered as approximations to patterns from more complex algorithms, and the approximation error is characterized. Conclusions regarding the appropriateness or correctness of the algorithms are outside the scope of this work. However, understanding where and why differences exist is an important first step in any search for optimality. The next section of this paper defines the masking and excitation algorithms under consideration. Section 3 presents a framework for the comparisons, and in Section 4 we summarize and explore the observed results.

2. EXCITATION AND MASKING PATTERNS

Figure 1 describes a generalized masking or excitation algorithm. Many of the operations are most conveniently described on a critical band scale, where frequency is measured in Bark rather than Hz. We used the relation $b = 6\sinh^{-1}(f/600)$ to transform a Hz scale power spectral density X(f), to critical band scale power spectral density X(b), which can serve as input to the algorithms[9].

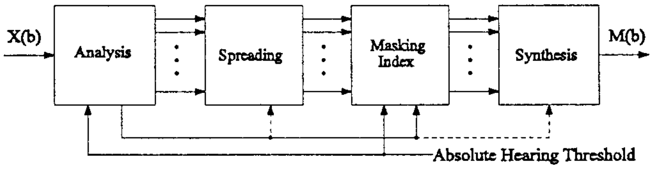

Figure 1: Generalized excitation or masking algorithm. X(b) is a power spectral density, and M(b) is a corresponding masking pattern.

Two analysis functions were considered. The analysis specified in MPEG psychoacoustic model 1 involves separating the narrower, more tone-like spectral components from the wider, more noise-like spectral components, comparing components with the absolute hearing threshold (threshold in quiet), and eliminating weaker tonal components located within one Bark of a stronger tonal component[1]. The second analysis function passes X(b) through unchanged but extracts a tonality parameter α, which is passed on as side information,

$$\alpha = \min[\,-\tfrac{10}{60}\log_{10}(Gm/Am)\,,1\,]\,, \qquad (1)$$

where Gm and Am are the geometric and arithmetic means of X(f) [2]. Note that α increases from 0 to 1 as X(f) becomes less flat.

Spreading functions emulate how a spectral component of an audio signal excites a neighborhood on the basilar membrane. Eight spreading functions were examined. A frequency and level-dependent triangular spreading function appears in [10], and is used in [3]. The i^{th} spectral component, with level L dB SPL, located at b Bark (or f Hz), contributes to the excitation at frequency b+Δ Bark according to:

$$E_i(b+\Delta)=\begin{cases} L+25\Delta \ \ \text{dB}, & \Delta\le 0, \\ L-(24-.2L+230/f)\Delta \ \ \text{dB}, & \Delta\ge 0. \end{cases} \qquad (2)$$

We refer to this spreading function as TRIFL. A frequency-independent version (TRIL) of (2) was created by fixing f at 2.4 kHz (12.5 Bark); this corresponds to the center of the 25-Bark (20 kHz) analysis band. A level-independent version (TRIF) of (2) was constructed using measured long-term average levels as a function of frequency. Figure 2 describes five other spreading functions. The spreading specified by MPEG model 1 (MPEG) is frequency-independent, but its level dependency is displayed in both its downward and upward spreads[1]. The triangular (TRI), flat-topped triangular (FTTRI)[4], and rounded (RND)[5] spreading functions are independent of level and frequency, as is the modified rounded (RNDMOD)[1] spreading function specified in MPEG psychoacoustic model 2. Note that the spreading functions TRI, FTTRI, RND, and RNDMOD share similar asymptotic slopes of approximately +25 and -10 dB/Bark.

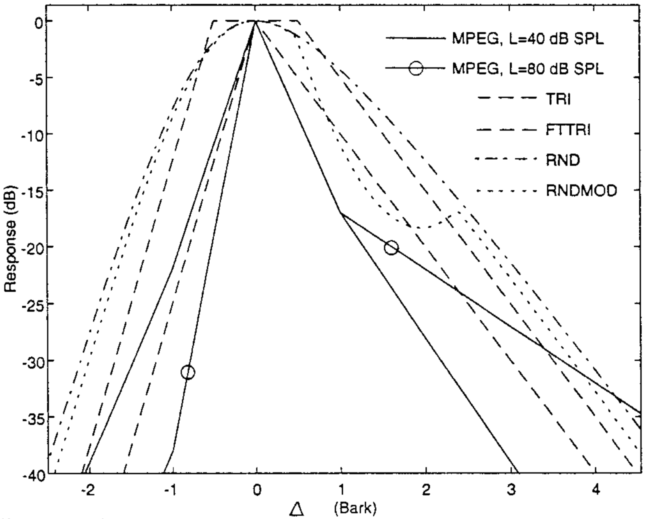

Figure 2: Five spreading functions.

A masking index describes the relationship between an excitation pattern, E(b), and a masking pattern, M(b). Three masking indices were considered; each converts excitation patterns to masking patterns by a frequency-dependent attenuation factor. Expressed in dB, the relationship specified in MPEG model 1 is[1]:

$$M_i(b) = \begin{cases} E_i(b) - (6.025 + .275\,b)\ \text{dB}, & \text{Tonal}, \\ E_i(b) - (2.025 + .175\,b)\ \text{dB}, & \text{Nontonal}. \end{cases} \tag{3}$$

In [2], tonal and nontonal masking indices are weighted by the tonality parameter α,

$$M_i(b) = E_i(b) - [\alpha(14.5 + b) + (1 - \alpha)\,5.5]\ \text{dB}. \tag{4}$$

In (4) α can be replaced with $\bar{\alpha}$ to create a third, "average tonality" masking index. If a masking index is set to 0 dB for all frequencies, then the output of the algorithm in Figure 1 is an excitation pattern.

A synthesis function combines the contributions of the various spectral components into a single excitation or masking pattern. A family of synthesis functions is described by

$$10^{\frac{M(b)}{10}} = \left[\sum_{i=1}^{N} 10^{\frac{M_i(b)}{10} p} \right]^{\frac{1}{p}} , \tag{5}$$

where N is the number of contributions, and p is a positive real number. When p = 1, "masking powers" are added, when p = ½, "masking amplitudes" are added, and as p → ∞, the maximum of the N contributions dominates. The maximum function, "p = ∞" is used in [6], p = 1 in [1,2,4,5,7], p = 0.5 in [10], and p = 0.48 in [8]. In [3], p = 0.4 is found to be most useful when E(b) is used to estimate the perception of coding distortions, and in [11] values of p between .1 and .3 provide the best fit to experimental results. Note that independent of p, the synthesis function (5) can be applied before or after the masking index in (4) with identical results.

The role of the absolute threshold of hearing is discussed in Section 4.3. To avoid taking them out of context, the various components of the MPEG model 1 masking pattern algorithm are used only with each other.

3. PROCEDURE

In order to maximize the generality and applicability of the results, comparisons among the excitation and masking patterns were based on a wide range of audio program material. Twenty, two-second selections were used. Selections included solo instruments (piano, flute, organ, guitar, harpsichord), string and brass groups, full orchestra, big band, pop, country, funk, rock, announcers, and sound effects. The selections were available in digital format (f_s = 44.1 kHz). The monaural sum signal was transformed in blocks of 512 samples (11.6 ms) using a Hamming-windowed FFT. The resulting magnitude squared, X(f), was input directly to the MPEG analysis as well as the tonality calculation in (1), but it was transformed to a Bark scale representation, X(b), for use elsewhere. This transformation resulted in 101 samples spaced at .25-Bark intervals from .25 to 25.25 Bark (25 Hz to 20.2 kHz). At the output of each of the algorithms, the frequency resolution was reduced by taking the minimum of each group of four samples, resulting in 25 samples, uniformly spaced between 1 and 25 Bark. This is specified in [1] and leads to conservative masking patterns.

To compare patterns from "algorithm x" with those from a reference algorithm, the appropriate error, $E_x(b)$-$E_{ref}(b)$ or $M_x(b)$-$M_{ref}(b)$, was formed in the dB domain. For each frequency b, the mean and standard deviation of this error were calculated across all 3440 blocks of audio. Since the errors were often found to be decidedly non-Gaussian, 95% confidence limits were measured and reported as a potentially more useful indication of the deviations one might expect in practice.

4. RESULTS

4.1 Synthesis Functions

For any distribution of addends, it can be shown that the sum in (5) increases as p is decreased[12]. This is the only general relationship among the synthesis functions. However, we observed that for the distributions associated with these audio signals, the "gain" associated with decreasing p values is very well defined. Figure 3 shows the mean error (relative to the case p = 1) for excitation patterns calculated with p = 0.25, 0.5, and "∞." Since the mean error is the difference of two dB values, it can be interpreted as an average gain. The measured limits of the 95% confidence intervals for these errors are also shown. Across most of the band, 95% of the errors have magnitudes less than 6 dB (p = 0.25) or 3 dB (p = 0.5 or "∞"). These results include seven spreading functions (MPEG excluded) and would not be altered by the masking indices under consideration.

The observed response of excitation patterns as p is varied is consistent with (but not necessarily due to) locally flat spectra at L dB SPL. The addends in an excitation version of (5) would then be determined by the spreading function alone:

$$E(b_0)=\frac{10}{p}\log_{10}\left\{\sum_{k=1}^{N}10^{\frac{E_k(b_0)}{10}p}\right\}=$$

$$L+\frac{10}{p}\log_{10}\left\{\sum_{k=0}^{N_1}A^{\Delta pk}+\sum_{k=1}^{N_2}B^{\Delta pk}\right\}\approx L+\frac{10}{p}\log_{10}\left\{\sum_{k=0}^{\infty}A^{\Delta pk}+\sum_{k=1}^{\infty}B^{\Delta pk}\right\} \quad (6)$$

$$=L+\frac{10}{p}\log_{10}\left\{1+\frac{A^{\Delta p}}{1-A^{\Delta p}}+\frac{B^{\Delta p}}{1-B^{\Delta p}}\right\}\text{dB}\,.$$

In (6) we assume the TRI spreading function with slopes of +25 and -10 dB/Bark, resulting in values for A and B that satisfy $10\log_{10}(A)=-25\,\text{dB/Bark}$ and $10\log_{10}(B)=-10\,\text{dB/Bark}$, where Δ is the frequency resolution. (Δ = 0.25 Bark/sample in this case.) While "locally flat spectra everywhere" is certainly not a good assumption, it is probable that the important aspect of (6) is that because spreading functions decay exponentially, the sums tend to be dominated by a few addends contributed by spectral components near b_0. The number of these addends and hence the gain, is a function of the frequency resolution. Figure 3 shows that the magnitude of the gain associated with p is diminished near the band edges, consistent with a reduced number of available addends. Between 4 and 24 Bark, the mean values of the gains in Figure 3 are 37, 11, and -6 dB; these values agree closely with the values of 35, 10, and -4 dB produced by (6). A separate set of measurements was later made using 4 kHz bandlimited speech from 5 different languages. With Δ = 0.5 Bark/sample, mean gains of 26, 7, and -3 dB were measured. These measurements agree well with the values of 26, 6, and -2 dB produced by (6).

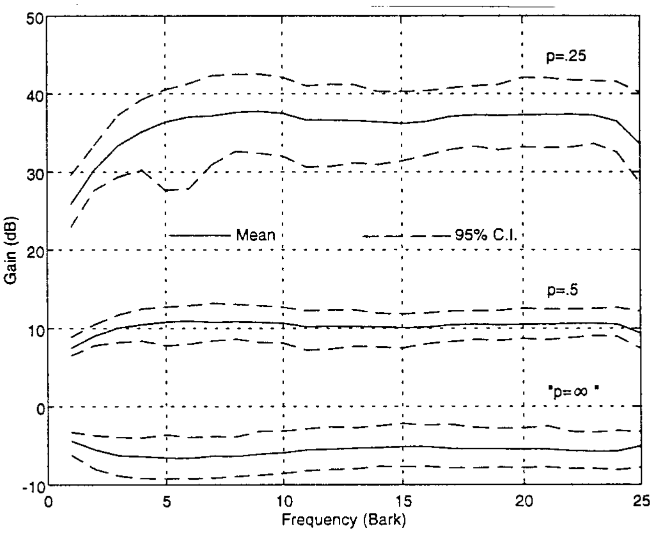

Figure 3: Gain of synthesis function, relative to p = 1. (Δ = .25 Bark/sample.)

4.2 Spreading Functions

Comparisons of the eight spreading functions are reported here for the case p = 1. Results at other p values are comparable, given the uncertainties reported in Section 4.1. The comparisons were based on the excitation patterns produced by the spreading functions. First consider TRIFL as a reference. TRIL shows small deviations from TRIFL, confirming that the frequency dependence described in (2) is minor, and is significant only at lower frequencies. At every frequency, 95% of the errors (TRIL-TRIFL) are between -.5 and +2.5 dB. Above 11 Bark, the errors are all nearly zero. At each frequency, 95% of the errors (TRIF-TRIFL) are between -6 and +3 dB. For (TRI-TRIFL), 95% of the errors are between -5 and +3 dB, at every frequency. Thus, for the audio signals used in this study, coding average level information into frequency dependence does not result in a better approximation to TRIFL. Once level dependence is eliminated, one might as well drop the frequency dependence too, and use the very simple TRI spreading function.

The MPEG spreading function was used only in conjunction with the MPEG analysis function. None of the seven other spreading functions provides a close match to the MPEG excitation patterns across all frequencies. The TRI spreading function comes closest, and at least 95% of the errors, (TRI-MPEG) fall between -4 and +8 dB at each of the sample frequencies between 4 and 20 Bark (430 Hz to 8.4 kHz) inclusive. Below 4 Bark however, the 95% confidence interval widens to [-7, +19] dB and above 20 Bark it is [-7, +14] dB. The error (TRIFL-MPEG) is similar to (TRI-MPEG) at the ends of the band, but generally shows an extra 1 to 2 dB of deviation in the middle of the band. Given the level dependence of the MPEG spreading function, one might expect TRIFL to provide the closer match, but this is not the case. In light of the much smaller sensitivities to spreading function reported in the previous paragraph, it is likely that the MPEG analysis function is a major source of the differences observed here.

Next, taking RND as the reference, a search through the other seven spreading functions showed that the FTTRI function provides the most similar patterns, with at least 95% of the errors (FTTRI-RND) falling between -1 and +4 dB at every frequency. This similarity might be attributed to the identical 1.4 Bark 3-dB bandwidths of these two spreading functions. In contrast, the TRI spreading function has a 3-dB bandwidth of 0.4 Bark. Finally, with RNDMOD as the reference, RND provides the closest match, with at least 95% of the errors (RND-RNDMOD) falling between -2 and +5 dB at every frequency.

4.3 Absolute Hearing Threshold

By definition, no masking pattern can ever fall below the absolute threshold of hearing. Some algorithms enforce this at each frequency by adding (using a chosen value of p) the power that corresponds to the absolute threshold, to the power of the excitation pattern[1,10]. Others calculate a frequency-by-frequency maximum between the excitation pattern and the absolute threshold[2,7]. At a fixed frequency, if the absolute threshold of hearing is T dB SPL, and the level of the excitation pattern is T+Δ dB SPL, then the difference, in dB, between the two rules is

$$\frac{10}{p}\log_{10}\left\{10^{\frac{Tp}{10}}+10^{\frac{(T+\Delta)p}{10}}\right\}-\max(T,T+\Delta)=\frac{10}{p}\log_{10}\left\{1+10^{\frac{-|\Delta|p}{10}}\right\}\text{dB}\,. \quad (7)$$

This difference takes a maximal value of 3/p dB when Δ = 0 dB, and a minimal value of 0 dB as $|\Delta|$ becomes large. Thus, the difference between the two rules is significant only when the excitation pattern is near the absolute threshold of hearing, and will never exceed 3/p dB.

4.4 Masking Indices

The masking indices in (3) were used only with the other elements of the MPEG masking algorithm. The index in (4) and its "average tonality" version were applied to excitation patterns from the TRI and RND spreading functions, resulting in four sets of masking patterns for comparison with the MPEG masking patterns. For the average tonality version, α was replaced with its average value, $\bar{\alpha}=.221$. The standard deviation of α is .090, and the 95% confidence interval for α about $\bar{\alpha}$ is [-.187, +.151]. It follows that for any spreading function, when $\bar{\alpha}$ is substituted for α in (4), 95% of the resulting errors will fall into the interval [-.187·(9+b), +.151·(9+b)] dB. This interval widens with increasing frequency, from roughly ±2 to ±6 dB.

The MPEG masking index can be simplified in an analogous fashion by forming a fixed linear combination of the tonal and nontonal parts of (3). The weight placed on the tonal portion was the average number of tonal components per block (6.01) normalized by the maximum number of tonal components per block (22). Like α, this average tonality parameter is constrained to (0,1]. Its value was .273 which is close to $\bar{\alpha}=.221$. However, the block-by-block correlation between α and the number of tonal components located by the MPEG analysis was very small, indicating that the two parameters do not generally carry the same information. The average tonality masking indices that follow from (3) and (4) are similar:

$$M(b)=\begin{cases} E(b)-(3.1+.2b)\ \mathrm{dB}, & \text{from (3)}, \\ E(b)-(7.5+.2b)\ \mathrm{dB}, & \text{from (4)}. \end{cases} \tag{8}$$

Comparison of the average tonality MPEG masking index with the true MPEG masking index shows that 95% of the errors (MPEGAT-MPEG) are contained in the interval [-2, +4] dB. Of the four other masking patterns calculated, the closest match to the true MPEG masking pattern was provided by the TRI spreading function along with the masking index specified in the second line of (8). The fact that the use of $\bar{\alpha}$ provides a closer match than the use of α is an additional indication that α and the MPEG analysis stage are evidently fairly independent. Figure 4 shows the measured limits for 95% of the errors (TRI-MPEG-mean error) as a function of frequency. The removed mean error was calculated at each frequency across all blocks of audio. As with excitation patterns, errors in the middle of the band tend to be considerably smaller than those at the ends. At least 95% of the errors fall between -6 and +10 dB at each of the sample frequencies between 4 and 20 Bark (430 Hz to 8.4 kHz) inclusive. Given the results of Section 4.2, it is likely that among our candidates, the combination of the TRI spreading function, the masking index specified in the second line of (8), and the synthesis function in (5) with p = 1, would form the best simple algorithm for approximating MPEG masking patterns. As a final test of the generality of this result, five new and different, two-second audio selections were processed. With the mean error of the original 20 selections removed, the interval containing 95% of the remaining error falls largely within the limits described in Figure 4. The only exceptions are at 4 and 5 Bark, where the upper limit for 95% of the errors is 12 dB.

5. SUMMARY

Figure 1 describes a generalized masking or excitation algorithm as a composition of four functions and Section 2 describes options for each of those functions. Measurements with audio program material reveal several interesting results. In Section 4.1 a simple mathematical relationship (6) characterizes the observed impact of p in the synthesis functions defined by (5). A relationship between two popular treatments of the absolute threshold of hearing is noted in (7). The impact of level and frequency dependence on a triangular spreading function is reported in Section 4.2. Several other spreading functions are compared as well. Surprisingly, the best simple approximations to MPEG excitation and masking patterns use level, frequency and tonality-independent rules. The errors associated with these approximations are characterized in Sections 4.2 and 4.4. Knowledge of these errors may be helpful to persons considering reduced complexity computation of auditory excitation and masking patterns.

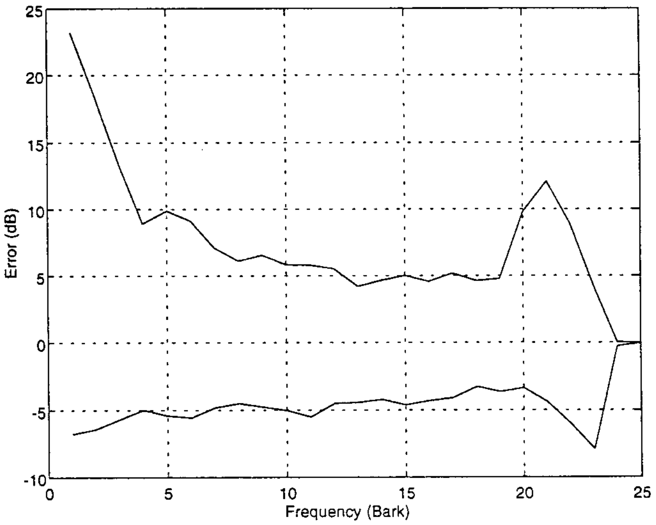

Figure 4: Ninety-five percent confidence intervals for errors of TRI relative to MPEG.

REFERENCES

1. ISO/IEC 11172-3, "Coding of Moving Pictures and Associated Audio..., Part 3: Audio," 1993.
2. Johnston, J.D., "Transform coding of audio signals using perceptual noise criteria," *IEEE J. on Sel. Areas in Com.*, vol. 6, pp. 314-323, Feb. 1988.
3. Beerends, J.G. and Stemerdink, J.A., "A perceptual audio quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 40, pp. 963-978, Dec. 1992.
4. Mahieux, Y. and Petit, J.P., "Transform coding of audio signals at 64 kbit/s," in *Proc. IEEE Global Telecom. Conf.*, 1990, pp. 518-521.
5. Schroeder, M.R., Atal, B.S., and Hall, J.L., "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Am.*, vol. 66, pp.1647-1652, Dec. 1979.
6. Theile, G., Stoll, G. and Link, M., "Low Bit-Rate Coding of High-Quality Audio Signals," in *Proc. 82nd Audio Engineering Society Convention*, 1987.
7. Sinha, D. and Tewfik, A.H., "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. on Signal Process.*, vol. 41, pp. 3463-3479, Dec. 1993.
8. Veldhuis, R.N.J., "Bit rates in audio source coding," *IEEE J. on Sel. Areas in Com.*, vol. 10, pp. 86-96, Jan. 1992.
9. Bullock, T.H., Editor, *Report of the Dahlem Workshop on Recognition of Complex Acoustic Signals*, Life Sciences Report 5, Berlin, Sept. 1976, p. 324.
10. Terhardt, E., "Calculating virtual pitch," *Hearing Research*, vol. 1, pp. 155-182, 1979.
11. Humes, L.E. and Jesteadt, W., "Models of the Additivity of Masking," *J. Acoust. Soc. Am.*, vol. 85, pp. 1285-1294, March 1989.
12. Hardy, G.H., Littlewood, J. E. and Pólya, G., *Inequalities*, Cambridge: University Press, 1952, p. 4.