THE IMPACT OF SIGNAL TEMPORAL STRUCTURE ON AUDITORY SENSITIVITY AND ITS APPLICATION TO AUDIO DYNAMIC RANGE CONTROL

By

QING YANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2011

To my dear parents, Mr. Shuntian Yang and Ms. Suizhen Fu

ACKNOWLEDGMENTS

I would like express my greatest gratitude to my advisor, Dr. John G. Harris, for his his inspiration, guidance and support throughout my entire Ph.D. studies. I would like to extend my thanks to my supervisory committee members Dr. Alice Holmes, Dr. Dapeng Wu, and Dr. Rahul Shrivastav for their insightful comments and suggestions. I greatly appreciate the help of all the volunteers for their participation in the subjective listening tests. I am also very thankful to all the students and staff in Computational Neuro-Engineering Laboratory (CNEL) who ever helped me, encouraged me, inspired me and supported me in my research and life in the past five years. Last but not least, I would like to give my most sincere thanks to my parents Mr. Shuntian Yang and Ms. Suizhen Fu, and my husband Chun-Ming Tang, for their endless love and supports.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2AFC	Two alternative forced choice experiment
ATS	Asymptotic threshold shift
DRC	Dynamic range control
ERB	Equivalent rectangular bandwidth
KFH phase	Key-Fowle-Haggarty phase
ODG	Objective difference grade output from PEAQ algorithm
PDF	Probability density function
PEAQ	Perceptual evaluation of audio quality
PMP	Personal media player
PSD	Power spectral density
PTS	Permanent threshold shift
RMS	Root-mean-squared value
SNR	Signal-to-noise ratio
SPL	Sound pressure level
TK point	Threshold knee-point
TTS	Temporary threshold shift
TWA	Time weighted average

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

THE IMPACT OF SIGNAL TEMPORAL STRUCTURE ON AUDITORY SENSITIVITY AND ITS APPLICATION TO AUDIO DYNAMIC RANGE CONTROL

By

Qing Yang

August 2011

Chair: John G. Harris
Major: Electrical and Computer Engineering

The purpose of this dissertation is to investigate the impact of signal temporal structure on human auditory sensitivity and its application to audio dynamic range control. This dissertation is organized as follows.

Firstly, the relation between the signal temporal structure and human auditory sensitivity is systematically studied with subjective listening tests. The subjective results show that the human auditory system is more sensitive to transient signals than steady signals given the same energy.

Inspired by the impact of signal temporal structure on auditory sensitivity, a high-order spectro-temporal integration model is developed to better predict the audibility thresholds of non-stationary sounds. This higher-order integration model is shown to outperform the existing energy-based and loudness-based audibility prediction models on our experimental data. This model can be extended to provide improved standards for determination of hearing impairment.

We propose the use of dynamic range control (DRC) algorithm for hearing protection. To further improve the conventional DRC algorithm, the level estimation in the traditional DRC framework is extended from second order to higher order. The

objective evaluation results show that higher-order DRC algorithms perform best for moderate-size analysis windows.

CHAPTER 1
INTRODUCTION

The purpose of this study is to investigate how the temporal structure of sound impacts auditory sensitivity and its application to audio dynamic range control. Auditory sensitivity has been studied intensively in the area of psychoacoustics, but most of these studies have focused on human auditory sensitivity to stationary sounds, such as pure tones and octave-band noise. The energy detector is the dominant approach to model auditory sensitivity to these stationary sounds. However, the sounds we deal with in our daily life are mostly non-stationary sounds, e.g., speech and audio signals. One of the most important differences between stationary and non-stationary sounds is that the temporal statistics, including signal mean and variance, vary over time for non-stationary sounds. Given the temporal differences between stationary and non-stationary sounds, the impact of the temporal structure on human auditory sensitivity to non-stationary sounds is a particular research interest in this study.

To find out how the signal temporal structure affects human auditory sensitivity to non-stationary sounds independent of spectral differences, a family of signals with identical power spectra but very different temporal structure is synthesized using phase manipulation. The auditory sensitivity to these synthetic signals is measured based on the audibility thresholds – the lowest signal-to-noise ratios at which the human subjects can just detect the sounds in a given noise environment. The subjective responses in this experiment show that the audibility thresholds are lower for transient signals than the steady signals given the same power spectra, which indicates the human auditory system is more sensitive to the transient signals than the steady signals given the same energy (power spectra). The results from this subjective experiment cannot be

explained with the conventional energy-based model for predicting the auditory sensitivity to stationary sounds.

The new finding in our subjective auditory sensitivity experiment leads to an improved model in predicting the audibility thresholds of non-stationary sounds. Predicting the audibility of audio signals is a crucial task in many acoustical applications. Engineers working on noise monitoring/controlling would be interested in predicting if the aircraft noise in an airport or the rock music from an outdoor stadium would bother the residents in the neighborhood. Designers of cell phone ringtones or audio alarms would like to know which audio signals are more audible and can therefore better alert people. Audibility is also a big concern in the design of hearing aids. The hearing aid fitting is based on an audiogram tested with stationary pure tones typically with a range of frequencies from 0.5-8kHz. Of course, what the hearing impaired people are really interested in is to be able to hear non-stationary speech and music signals using their hearing aids. The mismatch between the audiogram test signals and the real-world signals causes significant difficulty in the hearing aid fitting process. In another related area, audibility prediction has further implication on hearing loss prediction.

In order to better predict the audibility thresholds of non-stationary sounds, there are two challenges that need to be addressed: auditory temporal integration and spectral integration. First of all, since the signal temporal structure plays an important role in determining the human auditory sensitivity to non-stationary sounds, the existing energy-based temporal integration models developed from stationary sounds are not adequate for real world sounds. A computational model that takes the temporal structure into consideration is required for the audibility prediction of non-stationary

sounds. In addition, conventional models are designed for narrowband signals such as pure-tones and octave band noise, and therefore only concerned about the spectral integration within a single auditory frequency channel. A reasonable spectral integration scheme for wideband signals is also required to handle the spectral information from multiple frequency bands in determining the audibility of real-world sounds. Given these two considerations, a high-order spectro-temporal integration model is proposed to predict the audibility of non-stationary sounds and it is shown that this higher-order integration model outperforms all the conventional models based on our experimental data.

With the enhanced understanding of the auditory sensitivity to non-stationary sounds, an application of the auditory sensitivity model is exemplified in an audio dynamic range controller using higher-order level estimation. Dynamic range control (DRC) has been widely used in the design of hearing aids, radio and TV broadcasting, teleconferencing and other acoustical applications. As modern personal media players (PMP) with mass-storage capacities, long battery life, and high output levels, become more and more popular, music-induced hearing loss is becoming more of a social and clinical problem. Listeners often set volume levels based on the intelligibility or detectability of the softest sounds in the audio signals. For audio signals with wide dynamic range, at a given volume level, when the softest sounds are adequately audible, the loudest sounds might be overwhelmingly intense. As shown in previous studies, the loudest transient signals likely cause the most damage to the auditory system (Hamernik and Qiu 2001, Strasser et al. 1999). Therefore, a delicate DRC

algorithm that balances dynamic range reduction and perceptual concerns would be beneficial to protect the hearing of music listeners.

As the audio signals are mostly used for entertainment purposes, a primary concern for reducing the dynamic range of audio signals is the consequent quality degradation. A desirable dynamic range controller should provide the optimal balance of both dynamic range reduction and audio quality. In other words, for a given dynamic range reduction, the designed DRC should offer the best audio quality for fidelity requirements; for a given audio quality requirement, the designed DRC should offer the most reduction on signal dynamic range for hearing protection purposes. The best compromise between the dynamic range reduction and the audio quality is realized in our proposed DRC using higher-order level estimation.

The remainder of this dissertation is organized into five chapters. Chapter 2 reviews the existing studies on auditory temporal and spectral integration. Chapter 3 explores how the temporal structure impacts the auditory sensitivity in human subjective experiments. Chapter 4 proposes a higher-order specto-temporal integration model for audibility prediction and compares it with the existing models using our experimental data. Possible extensions of the proposed higher-order integration model are also summarized. Chapter 5 discusses a practical application of the auditory sensitivity model to audio dynamic range control. Chapter 6 concludes the dissertation.

CHAPTER 2
TEMPORAL AND SPECTRAL INTEGRATION FOR AUDITION

In this chapter, existing studies of auditory temporal integration and spectral integration will be reviewed, including the most important psychoacoustic experiments, classical theories and computational models for both stationary and non-stationary sounds. Relevant model simulations and their implications will be discussed as well.

2.1 Temporal Integration

Auditory temporal integration often refers to the ability of the auditory system to integrate information across the time to improve the detectability of a signal. In this section, conventional experiments and models that explain the auditory temporal integration process are introduced in details.

2.1.1 Masking Experiments

In a typical auditory masking experiment, the subjects are asked if they hear the target signal in the stimulus or not in a simple detection task, or in which stimulus they hear the target signal given multiple choices. The signal levels are then adjusted according to the subjective responses. The lowest signal levels at which the subjects can just detect the target signals in quiet or in noise is defined to be the absolute audibility thresholds or the masked audibility thresholds. As Figure 2-1 illustrates, both absolute and masked thresholds depend on the duration of the test sinusoidal tone. The absolute threshold, also known as the threshold in quiet, is shown by dotted lines and thresholds for tone bursts masked by uniform masking noise of 60 and 40 dB sound pressure level (SPL) are indicated by solid and dashed lines, respectively.

In spite of the different masking noise levels, both absolute and masked thresholds as a function of tone durations are parallel curves in Figure 2-1. The dependence on

duration shows a constant test-tone threshold for durations longer than 200 ms corresponding to that of long-lasting sounds. For durations shorter than 200 ms, both the absolute and the masked thresholds increase with decreasing duration at a rate of 10 dB per decade. This behavior can be described by assuming that the human auditory system integrates the sound intensity over a period of 200 ms (Fastl and Zwicker 2006).

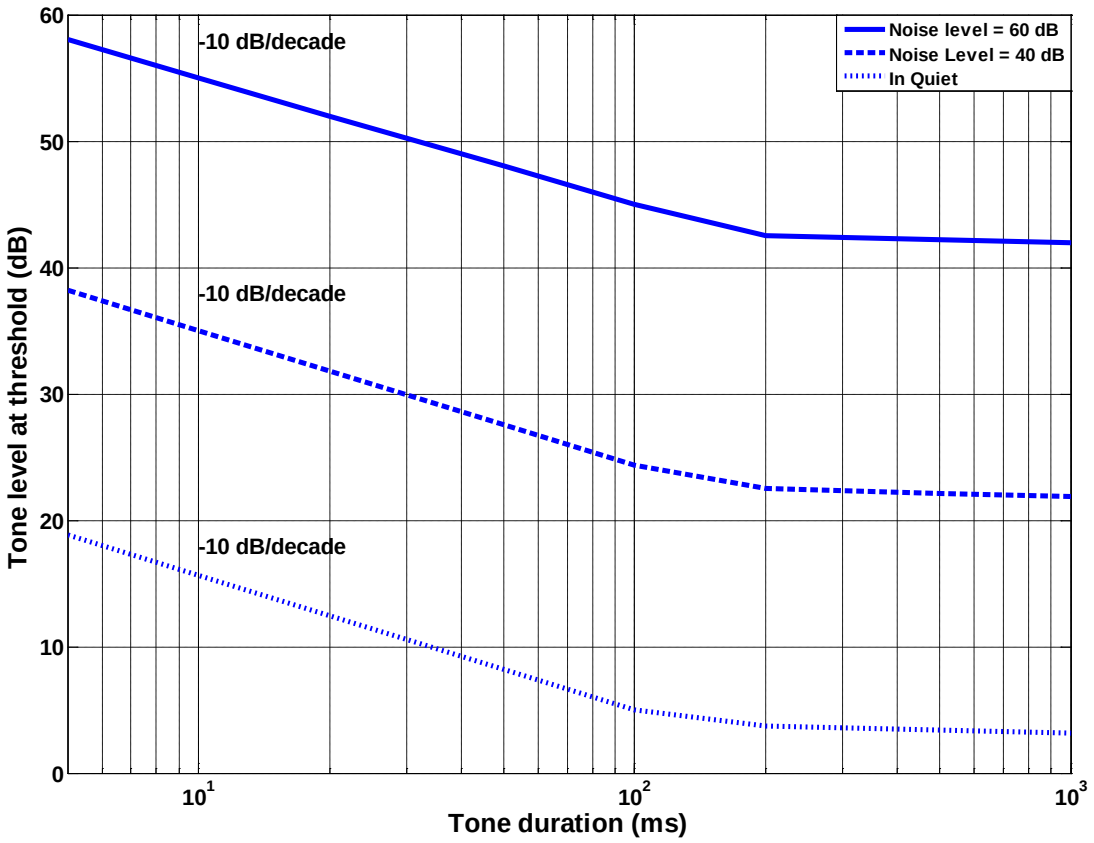


Figure 2-1. Just audible levels of 1 kHz tone bursts as a function of tone duration in quiet condition (dotted line) and masked by uniform masking noise of given levels (solid for noise of 60 dB and dashed line for noise of 40 dB). This figure is adapted from Fastl and Zwicker 's book Psychoacoustics published in 2006.

2.1.2 Classical Theory – Energy Integrator

The simplest explanation of the intensity-duration trade-off is an integration process.

An energy integrator is typically composed of three parts: a critical band filter, a rectifier and an integrator, as shown in Figure 2-2. The critical band filter ensures only the signal energy from the same critical band is integrated. The rectifier is usually a square-law device that provides a quantity whose average is a value monotonically related to the signal intensity. The integrator provides the basis for the final decision.

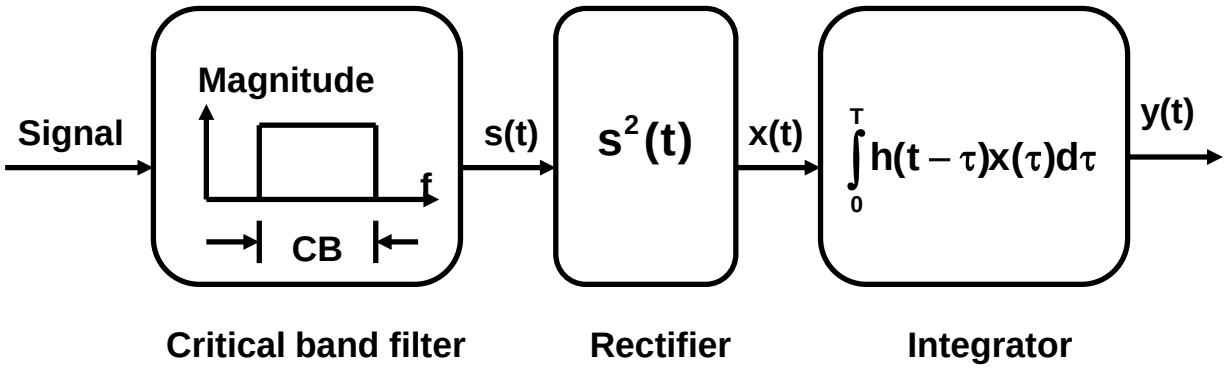


Figure 2-2. Block diagram of energy integrator

Suppose $x(t)$ is the input to the integration process. If we weight the input by a function $h(t)$ and integrate the weighted values, we obtain the output $y(t)$ as the following convolution:

$$y(t) = \int_{-\infty}^{t} h(t-\tau)x(\tau)d\tau \qquad (2\text{-}1)$$

The simplest form for $h(t)$ is the rectangular window,

$$h(t) = \begin{cases} 1 & 0 < t < \tau \\ 0 & \text{otherwise} \end{cases} \qquad (2\text{-}2)$$

This rectangular weighting function simply integrates all of the $x(t)$ occurring within time τ.

To better model the intensity-duration tradeoff, Muson (1947), Plomp and Bouman (1959), and Zwislocki (1960) introduced an exponential function for $h(t)$ in their seminal papers.

$$h(t) = \begin{cases} e^{-t/\tau} & t > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2-3}$$

This integrator weights past time by an amount of $e^{-t/\tau}$, so that recent events carry more weights, and events more than 3τ in the past are greatly attenuated and have little impact on the present output. This choice of $h(t)$ results in a simple first-order lowpass filter (Eddins and Green 1995).

The energy integrator predicts the signal audibility threshold solely dependent on the signal energy. It works well for narrowband stationary signals such as pure tones. However, for non-stationary signals, such as speech and audio, there are strong suggestions that the auditory sensitivity of impulsive signals is in part related to the peak levels (Price and Wansack 1985, Erdreich 1985). As a result, such energy-based audibility prediction underestimates the audibility of many transient or impulsive sounds.

2.1.3 Detectability Index (d')

The field of signal detection theory can be used to quantify the detectability of a signal. The detectability of a signal depends both on the separation and the spread of the noise-alone and signal-plus-noise curves as shown in Figure 2-3. Detection is made easier either by increasing the separation (stronger signal) or by decreasing the spread

(less noise). In either case, there is less overlap between the probabilities of occurrence curves.

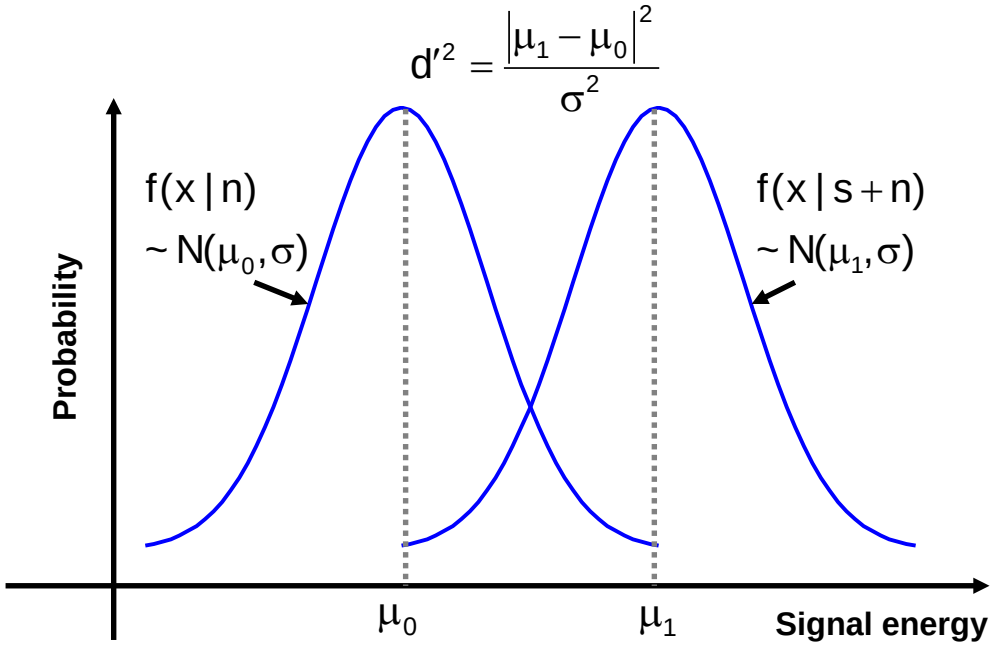


Figure 2-3. Illustration of detectability index d'

The most widely used measure for signal detectability is the detectability index d'

$$d' = \text{separation / spread}$$

$$= \frac{|\mu_1 - \mu_0|}{\sigma} \tag{2-4}$$

Assuming we have two hypotheses,

$$\begin{cases} H_0 : x(i) = n(i) & \text{noise} \\ H_1 : x(i) = s(i) + n(i) & \text{signal} + \text{noise} \end{cases} \tag{2-5}$$

and the conditional density functions for these two hypotheses are Gaussian distributions $N(\mu_0, \sigma)$ and $N(\mu_1, \sigma)$, respectively.

Let $T(x) = \sum_{i=1}^{N} x^2(i)$, we have

$$T'(x) = \frac{T(x)}{\sigma^2} \sim \begin{cases} \chi_N^2 & \text{under} \quad H_0 \\ \chi_N^2(\lambda) & \text{under} \quad H_1 \end{cases} \tag{2-6}$$

where $\lambda = \frac{\sum_{i=1}^{N} s^2(i)}{\sigma^2} = \frac{E_s}{\sigma^2}$

Using the properties of chi-squared distribution, we have

$$\mu_0 = N \tag{2-7}$$

$$\mu_1 = \lambda + N \tag{2-8}$$

$$\sigma_0^2 = 2N \tag{2-9}$$

$$\sigma_1^2 = 4\lambda + 2N \tag{2-10}$$

Assuming the signal power is very weak and there are enough samples, we have

$$\sigma_1^2 = 4\lambda + 2N \approx 2N = \sigma_0^2 \tag{2-11}$$

$$d'^2 = \frac{|\mu_1 - \mu_0|^2}{\sigma_0^2} = \frac{\lambda^2}{2N} = \frac{E_s / \sigma^2}{2N} \tag{2-12}$$

From equation 2-12, it can be seen that, d' is actually an estimate of the signal-to-noise ratio. Since d' is monotonically related to the expected percentage correct, it is frequently employed in signal detection (Green and Swets 1974, Kay 1998).

2.1.4 Recent Model – Multiple Looks

To account for temporal integration, both the rectangular integrator and the exponential integrator use a long window or time constant, typically hundreds of milliseconds. On the other hand, models proposed to explain temporal resolution, such as modulation detection, gap detection and certain temporal aspects of non-simultaneous masking, usually assume a lot shorter integration windows.

To solve this “resolution-integration” discrepancy, Viemeister and Wakefield propose a multiple-look model for temporal integration, which assumes that long-term integration does not occur, instead, the listeners can take multiple “looks” during a long duration signal and combine the information from all “looks” optimally to detect a signal (Viemeister and Wakefield 1991).

Assume that the temporal window for each look is a 3ms rectangular window, and that successive windows are contiguous and nonoverlapping. Also assume that the window samples are mutually independent and optimally combined. According to Green’s multiple observation theory (Green and Swet 1974), the overall detectability for n looks is the square root of the sum of the squares of the individual looks d_i':

$$d_n' = \sqrt{\sum_{i=1}^{n} (d_i')^2} \tag{2-13}$$

Furthermore, assume that

$$d_i' = k_i I \tag{2-14}$$

where I is the signal intensity and k_i is the weight for the i-th look.

In order to fit the integration data for pure tones in quiet shown in Figure 2-1

$$I = (1 - e^{-t/Tc}) I_\infty \tag{2-15}$$

where t is the signal duration, I_∞ is the threshold for very long duration signals, and T_c is the time constant used to fit the data, Viemeister and Wakefield derived a formula for the weights

$$k_i = \sqrt{\beta e^{-i(\tau/Tc)} (1 - e^{-i(\tau/Tc)})} \tag{2-16}$$

where τ is the duration of the temporal window and at threshold β is a constant given by

$$\beta = 2(\tau / T_c)(d' / I_\infty)^2 \qquad (2\text{-}17)$$

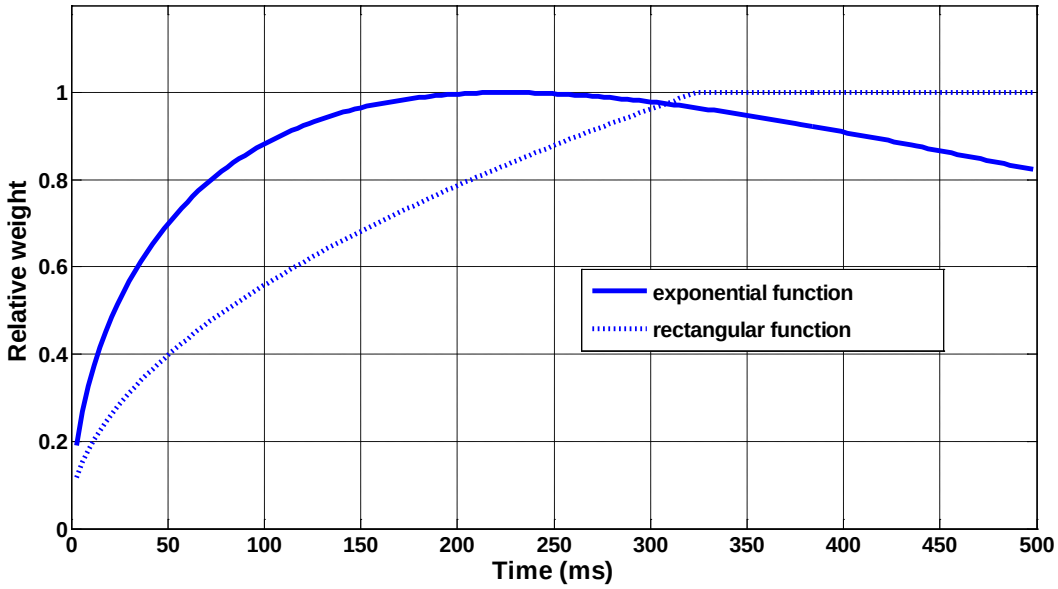


Figure 2-4. Weighting function for multiple-look model with a time constant of 325ms.

The weighting functions for the multiple-look model are plotted in Figure 2-4. The solid line represents the weighting function derived from the exponential integrator in equation 2-15 and the dashed-dotted line represents the weighting function derived from the rectangular window integrator $I = C / t \ (t < T_c)$ and $I = C \ (t \geq T_c)$.

From Figure 2-4, we can see that the weighting functions derived from different integration functions are significantly different, even though their integration functions look similar.

Since this weighting function is derived from the pure tone temporal integration data, it might be applicable only for signals with flat envelopes. The weighting function for non-stationary signals, especially transients, could be considerately different. Moreover, the temporal integration rule for non-stationary signals is not yet clear.

Without this prior knowledge, a general weighting function that works for an arbitrary non-stationary signal is hard to derive.

2.1.5 Simulation Results and Discussions

To compare the audibility prediction models discussed in the previous sections, Figure 2-5 illustrates the intensity-duration tradeoffs for a 1 kHz pure tone for three integration models: rectangular-window energy integrator, exponential-window energy integrator and the multiple-look model using an exponential window (plotted by solid, dashed and dotted lines, respectively). The window size/time constants for both energy integrators are 200ms.

In all three cases, they model the intensity-duration tradeoff -10 dB/decade very well. The thresholds of exponential-window energy integrator and the multiple-look model are around 3 dB different from that of the rectangular-window energy integrator, but they converge at around 1 second.

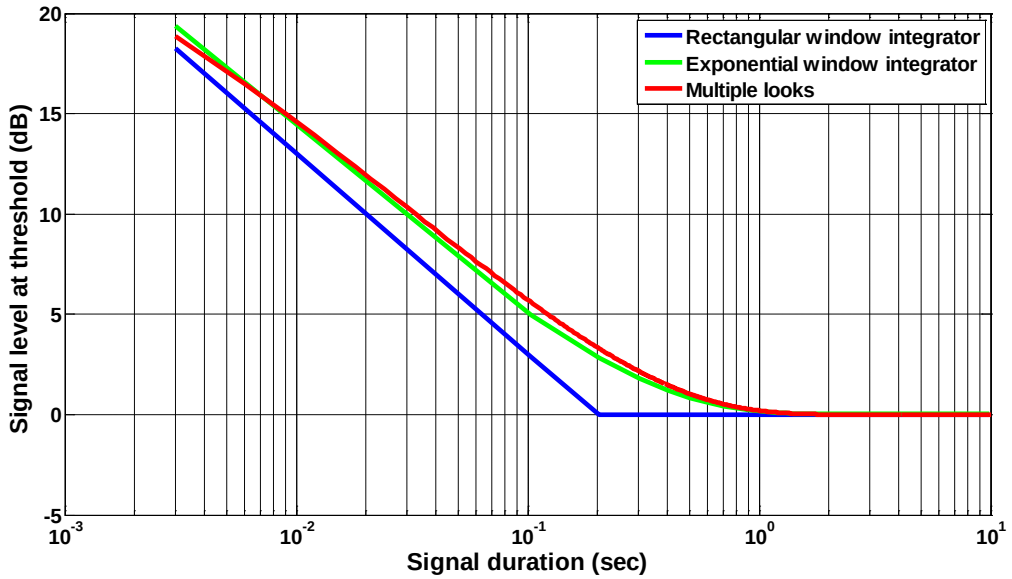


Figure 2-5. Predicted signal thresholds from three temporal integration models

Since all three models are developed from narrowband stationary signals, the goal for all these models is to predict the -10dB/decade intensity-duration tradeoff for narrowband signals with flat envelopes. Signals such as speech and audio are both non-stationary and wideband. First of all, the -10dB/decade rule is not necessarily fit for non-stationary signals. Solely integrating the signal powers is not sufficient to predict the audibility threshold for non-stationary sounds, especially for music signals which have a wide dynamic range. What is the appropriate quantity to integrate in the time domain is the key topic that will be explored in the later parts of this dissertation. On the other hand, real world signals are rarely confined within a single critical band. How to generalize a narrowband temporal integration model to a two-dimensional spectro-temporal surface for wideband signals is another key problem to discuss in this dissertation.

2.2 Spectral Integration

Auditory spectral integration often refers to the ability of the auditory system to sum information across a range of frequencies to improve the detectability of a signal in a masking experiment. There are two levels of spectral integration discussed in this section, within critical band and across critical bands.

2.2.1 Critical Band Theory

Fletcher first measured the threshold for detecting a sinusoidal signal as a function of the bandwidth of a band-pass noise masker. The threshold of the signal increased at first as the noise bandwidth increases. As soon as the noise bandwidth reaches the critical bandwidth, the detection threshold starts to flatten off (Fletcher 1940).

These observations can be explained with the signal processing in the peripheral auditory system. The basilar membrane in the cochlea can be represented as a series

of band-pass filters, called auditory filters. The detection of the signal is thought to be governed by the auditory filter that is centered around the signal. When the power ratio between the tone and the masker at the output of this filter exceeds a certain criterion value, the tone is assumed to be detectable. As long as the noise bandwidth is less than the auditory filter bandwidth, an increase in the noise bandwidth results in more noise passing through the auditory filter, which will lead to an increase in the detection threshold. Once the noise bandwidth exceeds the auditory filter bandwidth, the added noise power will be rejected by the band-pass auditory filter and thus have no impact on the detection threshold. Fletcher's classical critical band theory suggests that, signals in each critical band are processed independently. Signal powers are integrated only when they fall in the same critical band.

Later on, Moore and Glasberg proposed the equivalent rectangular bandwidth (ERB) based on their notch-noise masking experiments. ERB is said to provide a more accurate approximation of the auditory frequency response than critical bandwidth. The values of ERB are close to critical band scale but tend to be smaller at low frequencies than the values of critical bandwidth (Moore 2003).

The empirically fit equation describing the value of the ERB as a function of center frequency f (in kHz) is (Glasberg and Moore 1990)

$$\mathrm{ERB}(f) = 24.7(4.37f + 1) \tag{2-18}$$

A formula relating the number of ERBs to frequency f (in kHz) is given as (Glasberg and Moore 1990)

$$\text{Number of ERBs}(f) = 21.4\log_{10}(4.37f + 1) \tag{2-19}$$

This ERB scale is used to design the front-end auditory filterbank in our proposed high-order audibility prediction model that will be explained in Chapter 4.

2.2.2 Experiments on Complex Signals

Gassler measured the detection threshold for complex signals consisting of evenly spaced sinusoids. As the number of frequency components in a sound was increased, the threshold, specified in terms of total energy, remained constant until the overall spacing of the tones reached the critical bandwidth. Thereafter the threshold increased by about 3 dB per doubling of bandwidth. These results indicated the energies of the individual components in a complex sound will sum in the detection of that sound, provided that all components lie within the same critical band. When the components are distributed over more than one critical band, the detection is based on the single critical band that gives the highest detectability (Malmierca and Irvine 2005).

However, other results indicate that simultaneous presence of signal energy in different critical bands aids auditory detection. Green showed that psychometric functions for complex sounds composed of 12 or 16 equally detectable components were largely parallel to those for a single tone; the thresholds for the complex tones, in terms of the level per component, were about 6 dB below the thresholds for the single components in isolation (Green 1958). Buus et al. compared the masked threshold of a 450ms 18-tone complex relative to a single pure tone. Their results showed that the thresholds for an 18-tone complex, in terms of level per tone, is consistently lower than for pure tones and the level decrease followed the $10\log10(\sqrt{n})$ dB rule, where n is the number of frequency components (Buus 1986). Van den Brink and Houtgast further pointed out the masked threshold of a broadband brief signal (typically 10 ms or less)

decreases even faster as the signal bandwidth increases. Instead of $10\log10(\sqrt{n})$ dB, the detection threshold for brief compound signals in terms of level per component is lowered by $16\log10(\sqrt{n})$ dB, with n being the number of 1/3 octave bands that equally excited by the input signal (Van den Brink and Houtgast 1990, Malmierca and Irvine 2005).

How the auditory system improves the signal detection given the wideband information is the key to generalize to more sophisticated real-world signals such as speech and audio. Clearly, it is not because of the integration of the signal powers across frequencies. Classical critical band theory has clarified that only the signal powers that fall in to the same critical band are integrated and signals in different critical bands are processed independently. To explain the improved detectability of wideband signals, there are two models researchers generally support, multi-band energy detector and independent threshold model, which will be introduced in the next two sections.

2.2.3 Multi-Band Energy Detector

Green proposed a multi-band energy detector, to explain the improved masked threshold of a multi-component signal with respect to the threshold of each of the single-component stimuli (Green 1958).

This model postulates that the energy within each auditory channel is transformed into a Gaussian distributed decision variable $x_i \sim N(\mu_i, \sigma_i)$. The detectablity in each channel d_i' is proportional to its signal intensity. It also assumes that the detection for wideband signal depends on a weighted sum of decision variable in each channel

$$y = \sum_{i=1}^{n} w_i x_i \qquad (2\text{-}20)$$

It further assumes the human auditory system can optimally combine the information across channels and make a final decision. Green proved that, to make the detection optimal, the weights must be inversely proportional to the variance and proportional to the detectability in each channel, i.e.,

$$w_i = \frac{d_i'}{\sigma_i} \tag{2-21}$$

The resulting detectability of a wideband signal

$$d_n' = \sqrt{\sum_{i=1}^{n} (d_i')^2} \tag{2-22}$$

With Green's multi-band energy detection model, it is easy to explain the improved detectability of wideband signals. Suppose a wideband signal is equally detectable in all the critical bands, the overall detectability is $d_n' = \sqrt{n} d_i'$. Since d_i' is proportional to the signal intensity, the threshold corresponding to some fixed level of performance (e.g., $d' = 1.16$ to track a percentage correct of 79%), for the n component signal will be

$$L_n = L_1 - 10\log 10\left(\sqrt{n}\right) \quad \text{dB} \tag{2-23}$$

where L_1 is the threshold for a single component and L_n is the level per component in a complex signal at threshold.

Despite the convenience of explaining the detectability improvement of wideband signals using this multi-band energy detector, it does not seem physiologically plausible to assume that there is interaction among different critical bands below threshold, especially among those distant critical bands.

2.2.4 Independent Threshold Model

The independent threshold model was proposed by Schafer and Gales (Schafer and Gales 1949) and used by Scholl (1961) to explain the different psychometric functions for detection of narrow-band and wideband signals (Buus et al. 1986).

According to this model, an increase in detectability results from multiple channels is not because their information is integrated to form a single basis for decision, but rather because each additional channel presents another independent detection opportunity. In other words, it is a result of probability summation that causes the detectability improvement of wideband signals. This model also assumes statistically independent observation in each auditory channel.

The rule used to combine decisions is to make a positive response when any one of the several observations is positive. Under this rule, the probability of detection based on n observations is

$$p_n = 1 - \prod_{i=1}^{n} (1 - p_i) \tag{2-24}$$

If the individual probabilities are all equal, this formula becomes

$$p_n = 1 - (1 - p_i)^n \tag{2-25}$$

To treat the forced choice task with this model, it may be noted that, the probability of correct response on a single trial is

$$P(C) = p + (\frac{1}{m})(1 - p) \tag{2-26}$$

where m is the number of intervals in a trial. Hence, the expression for $P_n(C)$ can be written by substituting p_n in equation 2-22 for p in equation 2-23, with the result that

$$P_n(C) = 1 - \prod_{i=1}^{n}(1 - p_i) + (\frac{1}{m})\prod_{i=1}^{n}(1 - p_i) \quad (2\text{-}27)$$

Scholl showed that the summation of probabilities could give rise to steeper psychometric functions for wideband than narrowband signals. However, this finding is critically dependent on the assumed form of the psychometric function. Buus et al. derived a possible psychometric function for this model which could predict the $10\log10(\sqrt{n})$ dB improvement in detectability for n component signal.

$$P(C) = 0.5(2 - 0.5^{10^{0.2SL}}) \quad (2\text{-}28)$$

where P(C) is the probability of correct response in a two alternative forced choice (2AFC) task and SL is the sensation level in dB (Buus 1986).

In terms of physiology, the place coding theory says, signal information is encoded in the response of the auditory nerve fibers, which discharge at the rate in response to the movement of the basilar membrane at that location. Auditory nerve fibers at different locations fire independently and the ones that discharge at the greatest rate determine the signal audibility threshold. Also in terms of psychoacoustics, the critical band theory specifies only the signal powers in the same critical band are integrated together and different critical bands function independently. Hence, the independent threshold model is consistent with our existing knowledge both in physiology and psychoacoustics, which we prefer to embrace.

2.3 Non-Stationary Sounds

The existing temporal integration models, including the energy integrator and the multiple-look model, have focused on stationary signals. Some studies suggest that non-stationary sounds can be characterized by their crest factors and kurtosis (Erdreich

1986, Hamernik and Qiu 2001). Crest factor, which is the ratio of the peak level and the root-mean-square (rms) value, is sensitive only to signals with single largest peak and thus is not a good predictor (Erdreich 1986). Kurtosis is defined as the fourth cumulant divided by the square of the second cumulant and is often used as a measure of peakedness of the signal probability distribution. Kurtosis is correlated with auditory sensitivity of impulsive noise (Hamernik and Qiu 2001), however, it is not clear how to quantify its influence explicitly in a prediction model. More recently, Moore and Glasberg proposed an audibility prediction model for time-varying sounds based on loudness measurements. Psychoacoustic basics related to loudness and loudness based audibility prediction model will be introduced in the next subsections.

2.3.1 Loudness

Loudness is a psychological description of magnitude of auditory sensation (Flecter and Munson 1933), which varies directly with the sound intensity, frequency and bandwidth.

Frequency dependence: Equal loudness contours illustrate the frequency dependence of the loudness of pure tones. Based on equal loudness contours, for pure tones of different frequencies, the same intensity does not necessarily produce the same loudness. Also the contours are flatter at higher loudness levels than at lower loudness, which is the result of steeper function relating loudness to intensity at low frequencies than at medium frequencies.

Level dependence: Loudness as a function of signal intensity follows Steven's power law,

$$N' = k \cdot I^{0.3} \tag{2-29}$$

where N' is loudness, I is intensity and k is a constant depending on the listener and unit employed (Stevens 1955).

Bandwidth dependence: Zwicker et al. found out in their experiment, when total energy was held constant, as the bandwidth of a signal increased, its loudness stayed roughly the same until the bandwidth was greater than the critical bandwidth. Zwicker et al. explained, if loudness in each critical band followed Steven's law, then distributing the intensity of a sound across n critical bands, would result in an increase in total loudness, because $k \cdot I^{0.3} < n \cdot k \cdot (I/n)^{0.3}$ when n>1. This is Zwicker's theory of critical bandwidth for loudness summation (Zwicker et al. 1957).

Loudness scale: Stevens was the first to introduce the "sone" as the unit of loudness, where 1 sone was defined as the loudness of a 1 kHz pure tone with a sound pressure level of 40 dB. Loudness of other sounds would be determined by comparing with this reference, 1kHz 40 dB tone. For example, a sound judged by listeners to be twice as loud as the reference would have a loudness of 2 sones. (Moore 1995)

2.3.2 Audibility Prediction Based on Time Varying Partial Loudness

Glasberg and Moore proposed an audibility prediction model for time varying sounds based on the average short-term partial loudness (STPL). Partial loudness is defined as the loudness of a target sound presented in the context of other sounds, which is different from loudness in absolute quiet. The stages of this model are summarized as follows and illustrated in Figure 2-6. (Moore and Glasberg 2005)

1. A given signal and noise go through a finite impulse response filter representing the signal transmission of outer and middle ear.
2. Calculation of multi-resolution spectrum using the Fast Fourier Transform (FFT) for both signal and noise. To give adequate spectral resolution at low frequencies combined with adequate temporal resolution at high frequencies, six FFTs are

calculated in parallel with longer signal segments for low frequencies and shorter signal segments for high frequencies as illustrated in Figure 2-7. The windowed segments are zero padded and use a 2048-point FFT. All FFTs are updated at 1 ms intervals.

3. Calculation of excitation patterns from physical spectrum of signal and noise.

4. Transformation of excitation patterns to specific partial loudness.

5. Calculation of instantaneous loudness by integrating specific loudness across frequencies.

6. Calculation of short-term partial loudness by low-pass filtering instantaneous loudness along the time axis.

7. Average the short-term partial loudness over the entire duration of the given signal to determine the signal audibility.

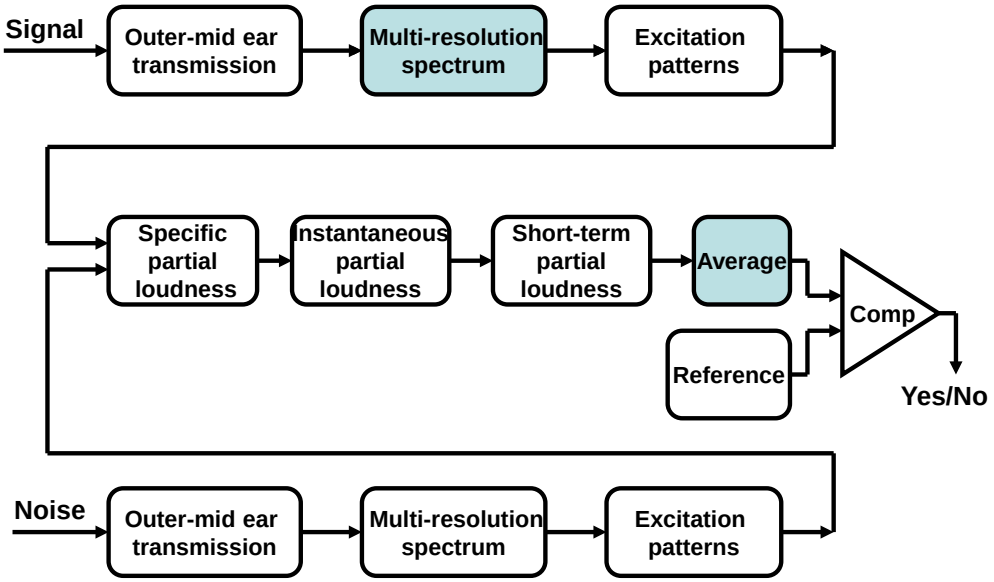


Figure 2-6. Block diagram of Moore and Glasberg's audibility prediction model

There are two drawbacks involved in the loudness-based audibility prediction model – the spectrum calculation and the loudness average process, which are shaded in Figure 2-6. Even though this model calculates loudness based on the multi-resolution spectrum, in each auditory frequency band, signals are still processed by a window of fixed size. This fixed-size window is not necessarily the ideal segmentation for some transient signals. As an example shown in Figure 2-8, if the transient signal is

segmented into two fixed-size windows as illustrated by dotted lines, neither of the windows is audible and they will be both given a loudness value close to zero sones. After lowpass filtering and averaging, the average short-term loudness will remain close to zero and the signal will be judged as not audible. However, if a different segmentation is used as illustrated in solid line, the audibility decision might turn around because the signal in one of the window is assessed as audible and the overall signal has better chance to exceed the audibility threshold.

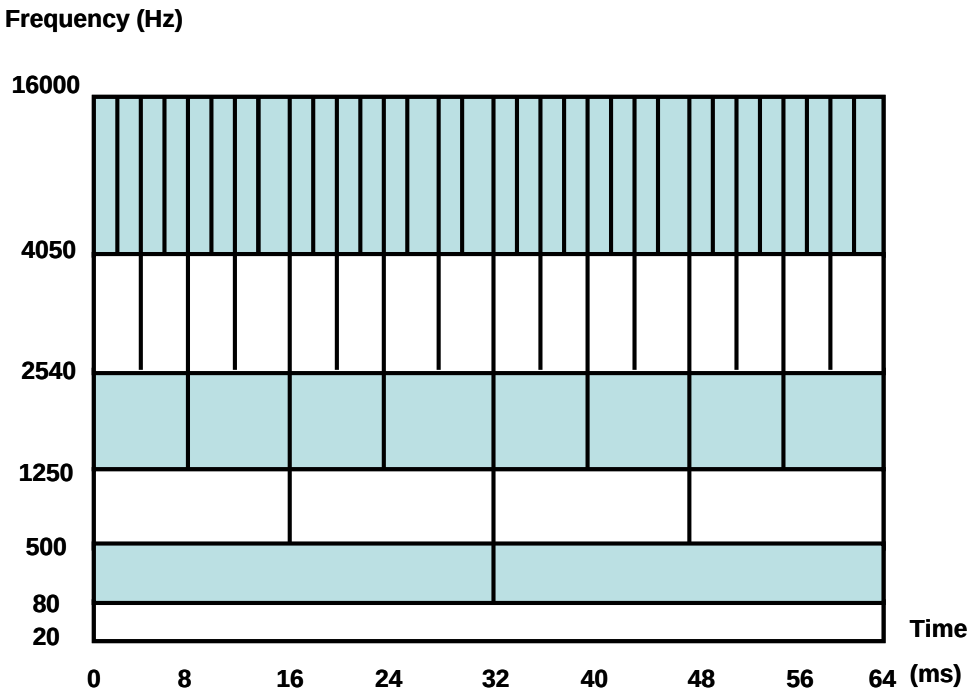


Figure 2-7. Multi-resolution spectrum in Moore and Glasberg's loudness model

For some impulsive signals, such as percussion music, the signal may last a long time but its energy is highly concentrated for only a very brief time interval. The audibility of the signal is actually caused by only a few frames that have high enough loudness. Even the signal is actually audible to the subjects, when counting its STPL over time, most of the frames may have loudness close to zero and the resulting average short-term loudness may become smaller than the pre-determined reference

(Figure 2-9). Hence, using average STPL over the duration of the signal tends to underscore the impact of transients on the signal audibility.

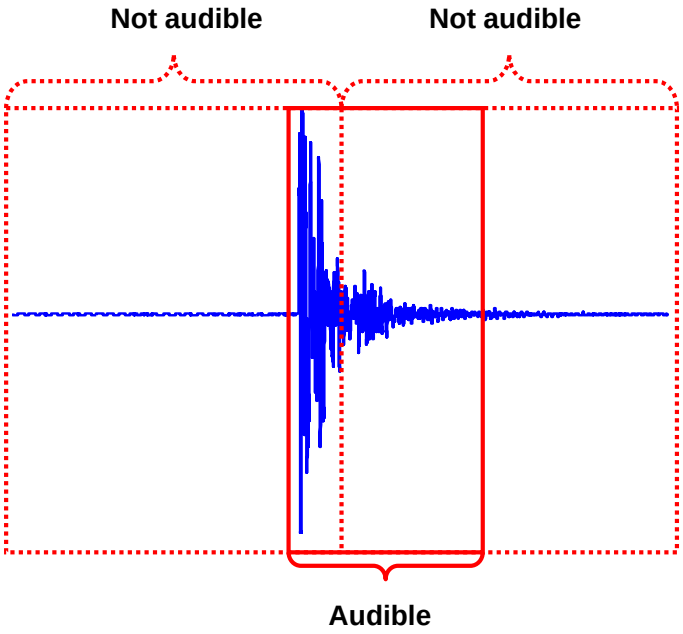


Figure 2-8. Illustration of different segmentation schemes

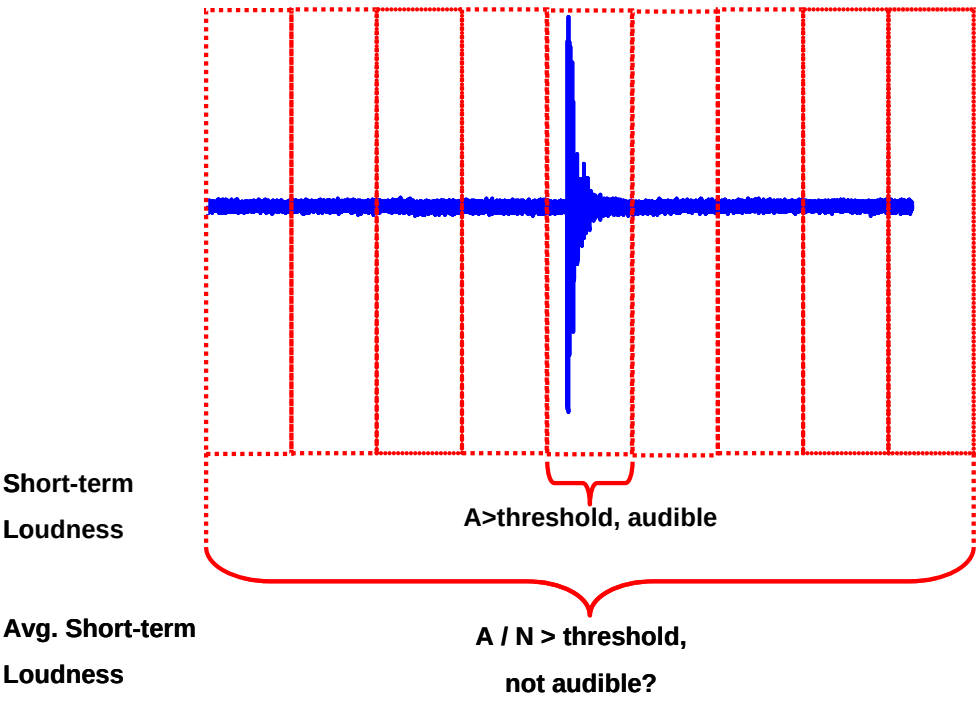


Figure 2-9. Illustration of averaging loudness across windows

Moreover, studies by Willera et al. suggest that the most efficient spectro-temporal integration for the auditory system is confined to a narrow time window and critical bandwidth (Willera et al. 1990). Summing up the specific loudness over all ERB bands and averaging the low-pass filtered instantaneous loudness over the duration of the signal, are actually giving equal weights to the loudness of all ERB bands and all time frames. This scheme is not consistent with the spectro-temporal integration efficiency of the auditory system.

2.4 Summary

In this chapter, we have reviewed temporal integration and spectral integration for audition. For temporal integration, both the energy detector and the multiple-look model are derived from the narrowband stationary pure tone psychoacoustic data and aim to model the -10 dB/decade intensity-duration tradeoff. However, the integration rule for wideband non-stationary signals such as speech and audio are not necessarily the same, so the energy-based metric may not be sufficient to predict the audibility thresholds of real world wideband non-stationary signals. Future chapters in this dissertation will explore the most appropriate quantity to integrate in the time domain.

As for auditory spectral integration, it is commonly agreed that signal powers in the same critical band are integrated. Both the multi-band energy detector and the independent threshold model can explain the improved detectability of a wideband signals. However, the multi-band energy detector is built on the assumption that auditory system can combine information optimally to make a final audibility decision for a wideband signal. It is still under debate whether across-frequency spectral integration is physiological plausible. On the other hand, the independent threshold model postulates our auditory system processes signals in different critical band independently

and the improved detectability of wideband signals results from the probability summation of multiple bands, which is consistent with our existing knowledge on critical bands and place coding. As a result, we use the independent threshold model as the basis for our more sophisticated spectro-temporal audibility prediction model for wideband non-stationary signals.

CHAPTER 3
SUBJECTIVE EXPERIMENTS ON NON-STATIONARY SOUNDS

This chapter aims to explore how the temporal structure of sounds impacts human auditory sensitivity. The design of the subjective experiment, including the test signal synthesis and the experiment paradigm, will be discussed in details. Observations from the subjective listening test show that, in spite of identical power spectra, the more impulsive the signal is, the lower the subjective audibility threshold will be, and the more detectable to the human auditory system.

3.1 Signal Synthesis

As mentioned in Chapter 2, the signal energy is not sufficient to characterize non-stationary signals. There is strong evidence that shows the human auditory system is more sensitive to transient signals than stationary signals given the same signal energy (Hamernik et al. 2001, Price and Wansack 1985, Erdreich 1986). In order to verify this concept, we generated a family of signals that have the same windowed Fourier spectra but different temporal envelopes. We then used these generated signals in subjective listening tests. By factoring out the spectral differences, we can clearly see how the temporal differences solely affect human audibility thresholds.

The key step of the experiment is to construct temporal waveforms from the Fourier magnitude and phase spectra. The solution is derived from the Fourier transform equation. In the Fourier transform pair in time and frequency domain, the temporal envelope is controlled by the spectral magnitude and phase. Since we do not want to change the magnitude spectrum, it leaves only the phase to manipulate the temporal envelope. By maintaining the original signal magnitude spectrum and replacing the phase spectrum with some designed phases, and applying the inverse

discrete Fourier transform (IDFT), a given signal temporal envelope can be changed over each selected time window without introducing any difference in the signal magnitude spectrum. The phase solution for our signal synthesis will be discussed in the next two sections.

3.1.1 PDF (Probability Density Function)-Based Phase Manipulation

Hsueh and Hamernik proposed a pdf-based phase algorithm for random noise synthesis as shown in Figure 3-1. By keeping the overall magnitude spectrum and changing the phase probability distribution (e.g., standard deviation) within certain selected frequency bands (on-band), peaks in the random waveform can be constructed from the phase manipulated frequency bands. In this way, an entire family of signals can be produced having the same magnitude spectra but temporal statistical characteristics (e.g., skewness and kurtosis) that vary along the continuum from Gaussian through non-Gaussian to purely impulsive (Hsueh and Hamernik 1990).

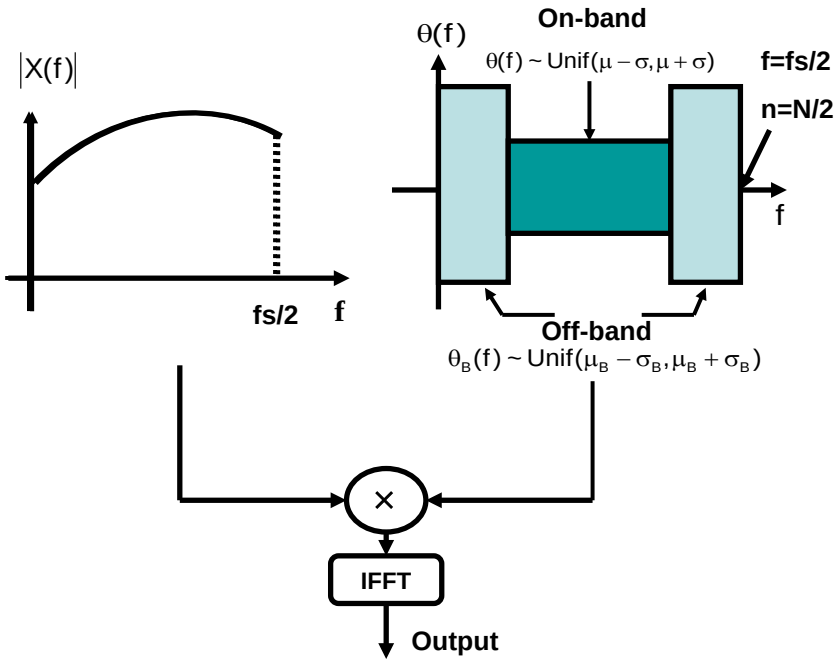


Figure 3-1. Block diagram of pdf-based phase algorithm

This pdf-based phase algorithm is able to synthesize random noise with different temporal structure and a given power spectrum for auditory sensitivity experiments (Hamernik and Qiu 2001, Yang and Harris 2009), however, when it is applied on audio synthesis frame by frame, it creates significant artifacts in the processed audio signals.

3.1.2 Key-Fowle-Haggarty Phase

Quatieri and McAulay proposed a peak-to-rms (root mean square) reduction technique for speech signals based on a sinusoidal model (Quatieri and McAulay 1991). We modified Quatieri and McAulay's method to manipulate the audio signal temporal waveforms without introducing any spectral differences. By maintaining the original signal magnitude spectrum and replacing the phase spectrum with the Key-Fowle-Haggarty (KFH) phase, and applying the inverse discrete Fourier transform (IDFT), a given signal envelope can be optimally flattened over each selected time window. This technique does not introduce any difference into the signal magnitude spectrum and assumes that the duration-bandwidth product is large enough (Fowle 1964). The diagram for our signal synthesis is shown in Figure 3-2. The filtering procedure before phase manipulation is to select single/multiple critical band(s) to be included in the synthesized signals. The KFH phase can be computed as

$$\theta_{KFH}(\omega) = -L\int_0^{\omega}\int_0^{\beta} \hat{M}^2(\alpha)\,d\alpha\,d\beta \qquad (3\text{-}1)$$

where $\hat{M}^2(\alpha)$ is the signal power normalized by its energy and L is the frame size in samples.

The KFH phase solution was originally designed to reduce the dynamic range of radar signals without changing the signal spectrum. Since the goal is to minimize the

dynamic range of the temporal envelope, the ideal case is to make the temporal envelope become a constant level, in other words, the dynamic range is zero. Given the magnitude spectrum, this constant temporal envelope can be derived using Parseval's theorem. The phase problem can be simplified as to find a spectral phase as a function of the signal magnitude spectrum through the Fourier transform equation under the constraint of a given constant temporal envelope. With a series of mathematical manipulation, including stationary phase theory and Taylor series expansion, an approximate solution was found by Key et al. at MIT Lincoln labs (Fowle 1964). The detailed mathematical derivation of the KFH phase for manipulation of signal temporal structure can be found in the Appendix.

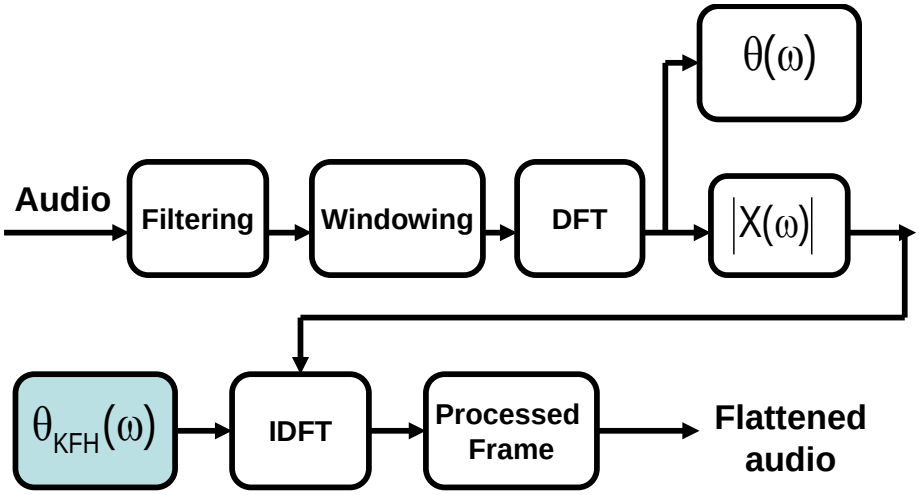


Figure 3-2. Block diagram of the KFH phase manipulation system

It can be seen from equation 3-1, the KFH phase is a function of the signal power spectrum and the selected window size. Since the magnitude spectrum is determined by the original signal, the peak-to-rms reduction for the entire signal depends on the window length used in the phase manipulation. Figure 3-3 shows a set of signals synthesized using the KFH phase algorithm. The original signal shown at the upper left

panel is an exponentially decaying 1 kHz sinusoid repeated at a period of 100 ms and the rest of the signals are phase manipulated signals with different window size. As the window size increases from 25 ms to 100 ms, the signal temporal envelope becomes flatter. When the window size exceeds 100ms, the period of the original signal, the signal envelopes are not further flattened. As a result, for periodic or quasi-periodic signals (e.g., voiced speech), choosing the pitch period as the window size L in equation 3-1 will optimally flatten the overall signal. For non-periodic signals, we can treat the overall signal as one period, the longer the window size used for KFH phase manipulation, the flatter the signal envelope will be.

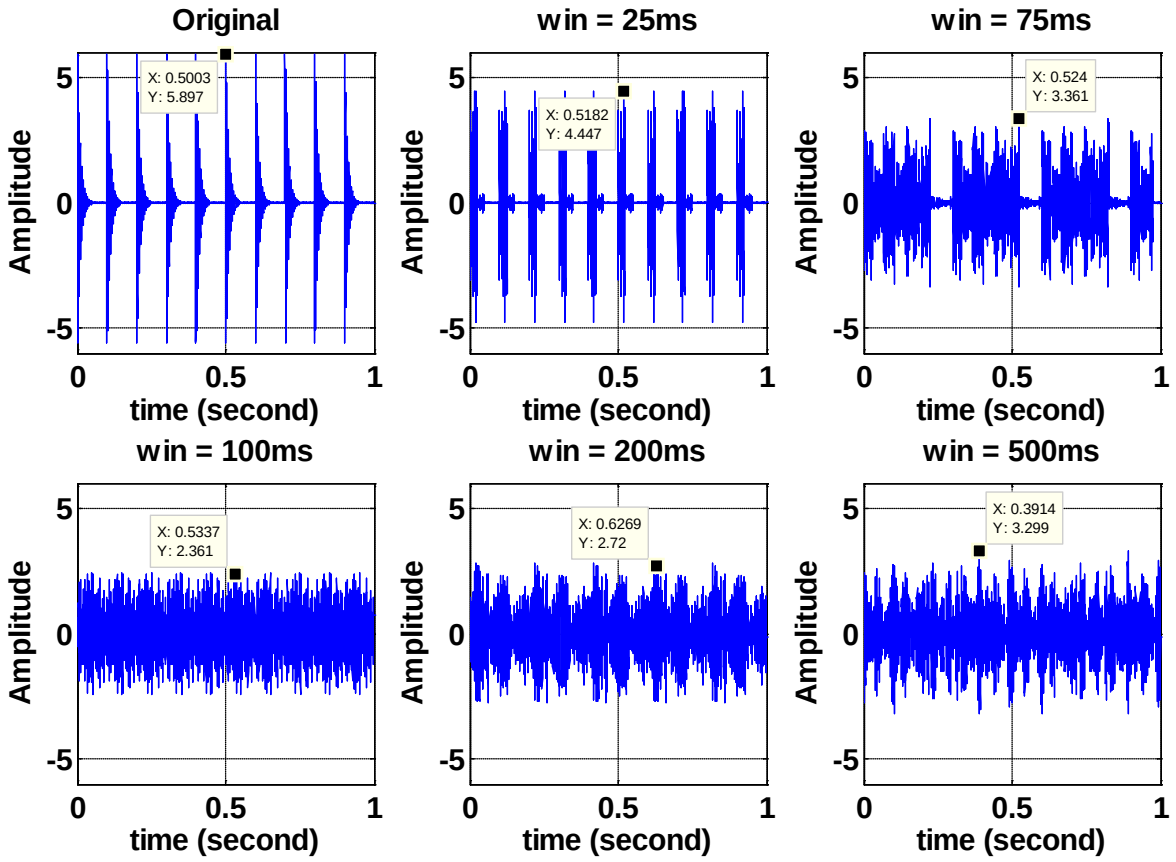


Figure 3-3. Synthetic signals from KFH phase algorithm

3.2 Subjective Listening Experiments

3.2.1 Test Signal Analysis

We study two special cases – single critical-band (1 kHz, 2 kHz and 4 kHz) and three critical-band (1kHz + 2kHz + 4kHz) audio signals. Two wide-band music clips are selected of typical Cuban percussion instruments: the conga and the bongo drum. Each clip is about 2-seconds long and sampled at 32 kHz.

For the single band case, these two signals are first ERB (Equivalent Rectangular Bandwidth)-filtered with center frequencies of 1kHz, 2kHz and 4kHz and fed into the phase manipulation system. The single-band signals are segmented into 50% overlapped frames with duration of 64 ms and 512 ms. For the three-band case, three ERB filtered single-band phase manipulated signals are summed into one three-band signal.

Figure 3-4 illustrates the power spectral densities (PSD) of single/three-band signals estimated using the Welch method (8 ms rectangular window with 50% overlap). The blue line shows the ERB filtered audio without phase manipulation, and the red line shows the phase-manipulated signal with a window size of 64 ms, and green line shows the phase-manipulated signal with a window size of 512 ms. The root-mean-squared values (RMS) of all three signals are normalized to unity. Theoretically, the processed audio and the input signal should have identical spectra window by window, but overall, there are some slight differences after connecting all the frames into one signal due to different window size used in the power spectral density estimation. However, the spectral differences are too small to cause a difference in audibility thresholds. On the other hand, Figure 3-5 shows the large temporal difference among the processed signals. Even though these signals have unity root-mean-squared power but they have

very different temporal structure. The peak-to-rms ratio differences are as large as 8.59 dB for narrowband signals and 10.8 dB for three-band signals.

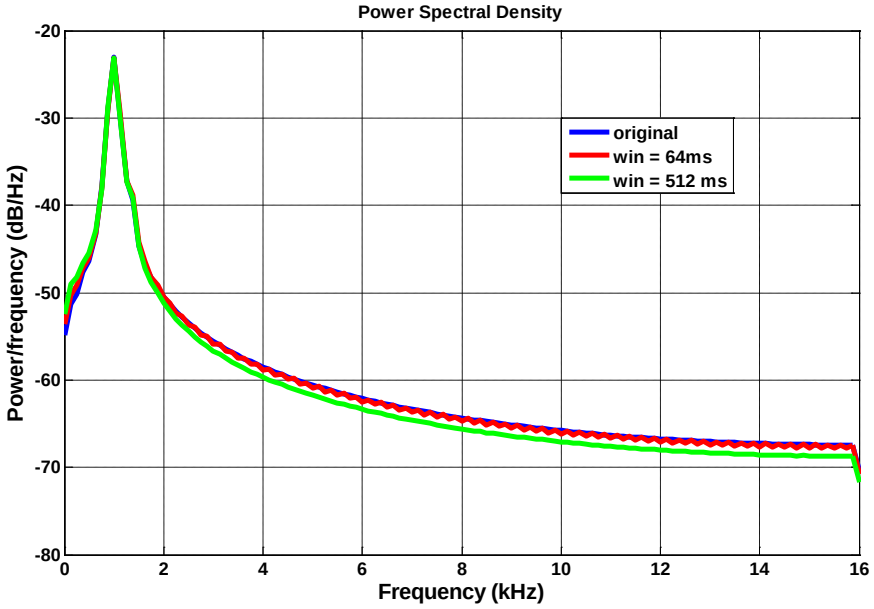
3.2.2 Experimental Paradigm

The tested audio stimuli are all played binaurally with added white noise through headphones (SENNHEISER HD 25-1) to subjects with normal hearing. The number of subjects is extended from one (Yang and Harris 2010) to four. Each signal is tested five times and the results are averaged. Levitt's adaptive two alternative forced choice (2AFC) paradigm was used to determine the masked signal-to-noise ratios (SNR) at threshold for all stimuli (Levitt 1971). There are two intervals in each trial: white noise alone and signal plus the same noise. Three correct responses count as a successful trial and one incorrect response counts as a failure trial, Hence, thresholds are determined at the 79% correct point. There are twelve reversals used for each test. The step size is initially set to 5 dB and reduced to 2 dB after the first four reversals and 1 dB after the first six reversals. The first 4 reversals are discarded and the following 8 reversals are used to calculate the masked SNR at the audibility threshold.

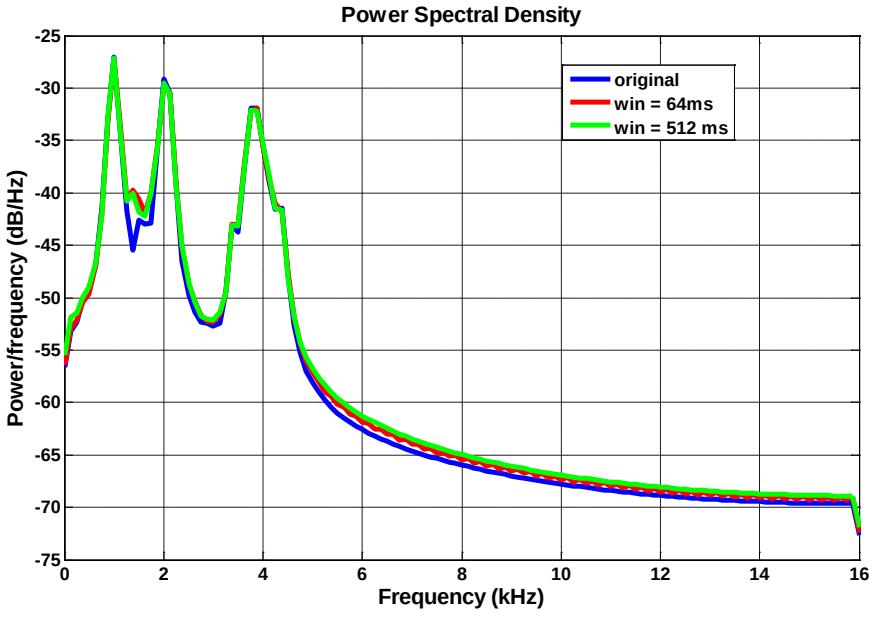
3.2.3 Discussion

The normalized masked SNRs in decibels for the single- and three-band signals are shown in Table 3-1 and Figure 3-6. From this table and figure, we can see that eight groups of signals have almost the same spectra within groups, but have 4 (Conga 1 kHz) ~ 11 dB (conga 4kHz) difference in audibility thresholds. Since the spectra are nearly identical, the factors that cause these differences in audibility thresholds must be temporal differences. In each of these eight groups, the original signal is the most impulsive signal (and have the lowest thresholds) while the phase manipulated signal with window size of 512 ms is the flattest signals (and the highest thresholds), indicating

that the auditory system is more sensitive to impulsive signals than to flat signals of the same power spectra.

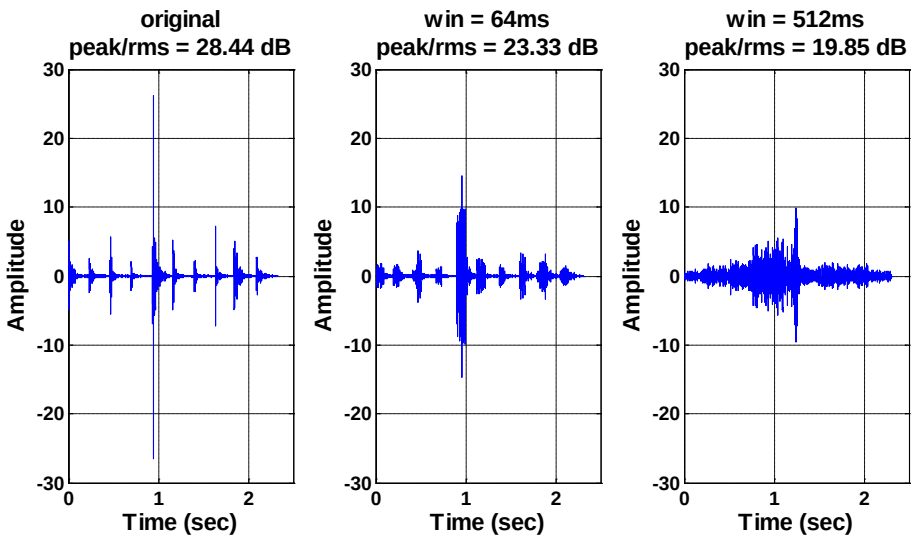


A

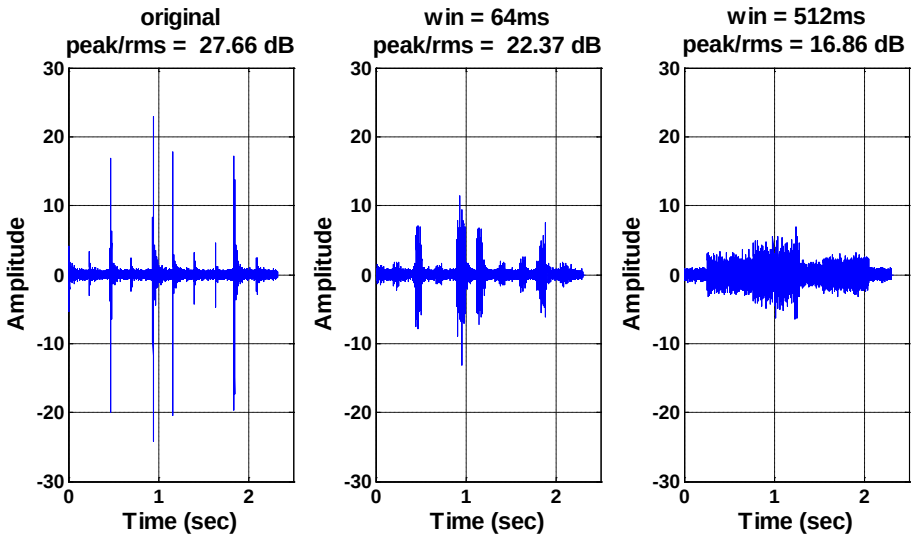


B

Figure 3-4. Power spectral densities (PSD) of A) single-band signals B) three-band signals



A

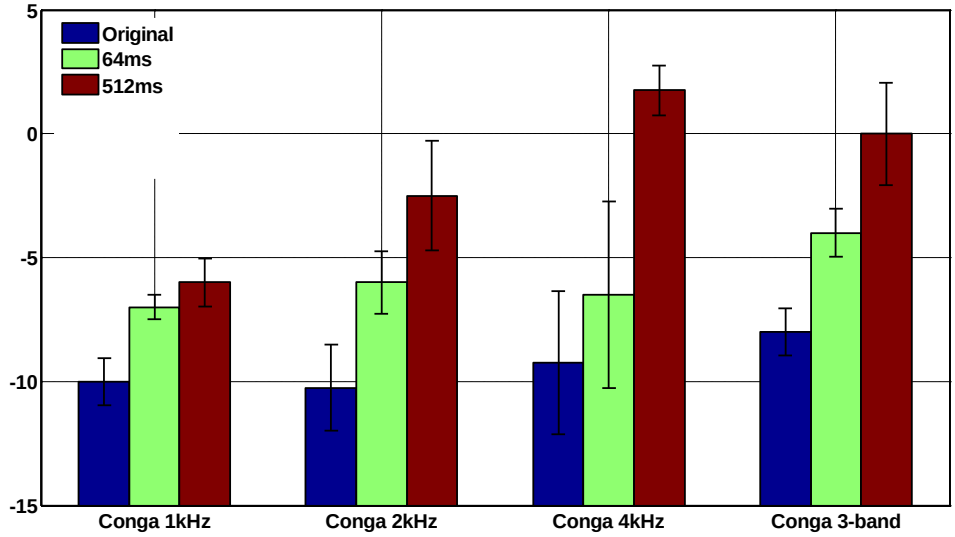


B

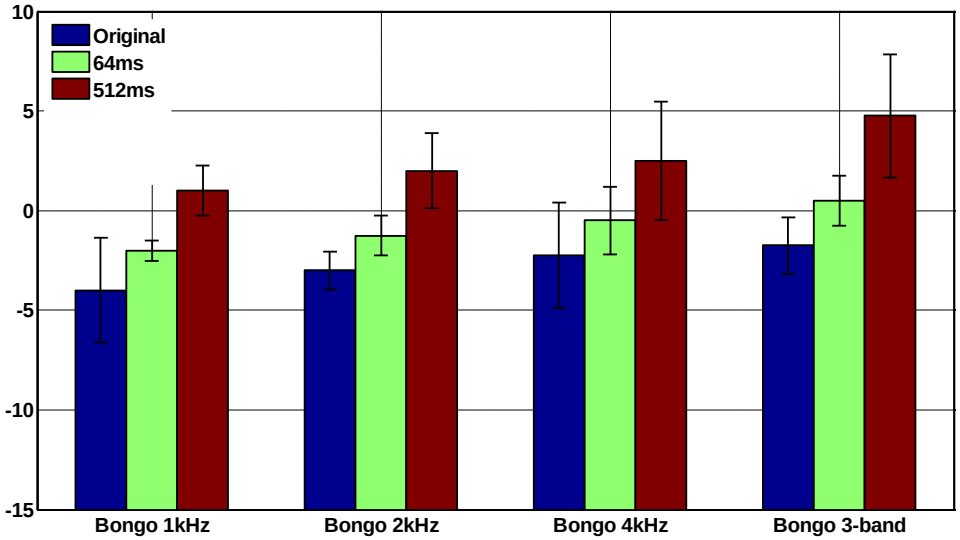
Figure 3-5. Time domain plots of single/three-band signals with corresponding PSDs shown in Figure 3-4. A) single-band signals. B) three-band signals.

Table 3-1 Average Subjective thresholds for single/ three-band signals normalized by the threshold of 1 KHz tone

Instrument	Window (ms)	1 KHz (dB)	2 KHz (dB)	4KHz (dB)	3-band (dB)
Conga	Original	-10.00	-10.25	-9.25	-8.00
Conga	64	-7.00	-6.00	-6.5	-4.00
Conga	512	-6.00	-2.50	1.75	0.00
Bongo	Original	-4.00	-3.00	-2.25	-1.75
Bongo	64	-2.00	-1.25	-0.5	0.50
Bongo	512	1.00	2.00	2.50	4.75



A



B

Figure 3-6. Bar plot of subjective thresholds for single/ three-band signals normalized by the threshold of 1 kHz tone that are listed in Table 3-1. A) Subjective thresholds for conga signals. B) Subjective thresholds for bongo signals.

3.3 Summary

The goal of this chapter is to find out how signal temporal structure impacts human auditory sensitivity. To exclude the spectral differences and observe how the temporal structure alone affects the signal audibility, two phase manipulation approaches are considered for test signal synthesis. By changing the standard deviation (STD) of the phase distribution, Huesh and Hamernik's pdf-based approach is effective in generating a family of noise signals with different impulsiveness given a pre-selected power spectrum. However, when applied to audio signal synthesis, this approach causes a lot of artifacts and severely degrades the perceptual quality of the audio. On the other hand, by changing the manipulation window size, the KFH phase approach can modify the signal temporal structure without modifying the signal power spectra, while retaining the perceptual quality of synthetic audio at an acceptable level.

Signals synthesized with the KFH phase algorithm were tested in a standard two alternative forced choice subjective (2AFC) experiment. The subjective responses show that, given the nearly identical power spectra, the more impulsive the signal is, the lower the audibility threshold is. In other words, the human auditory system is more sensitive to impulsive signals than the steady signals of the same power spectra, which cannot be explained by the conventional second-order energy detector. To better predict the audibility of non-stationary signals, a higher-order integration model will be proposed in the next chapter.

CHAPTER 4
A HIGHER-ORDER SPECTRO-TEMPORAL INTEGRATION MODEL FOR AUDIBILITY PREDICTION

In this chapter, a higher-order model to determine the audibility of audio signals is presented. Previous models have been energy based (second-order) and adequate only for stationary, narrow-band signals. Music, speech and other audio signals are non-stationary and wideband so traditional energy-based models poorly predict the audibility of these sounds. The predictions from the higher-order model are compared to actual subjective listening tests to show that the higher-order, wide-band technique outperforms previous models. Possible extensions of our higher-order integration model are proposed as well.

4.1 Basic Model

4.1.1 Model Description

The goal of our model is to predict the lowest signal-to-noise ratio (SNR) at which human subjects can detect a stimulus in the presence of a white noise masker. The signal can be wideband or narrowband, stationary or transient. The block diagram of the model is shown in Figure 4-1.

The model consists of a front-end auditory filter bank, followed by a nonlinear operator which is then low-pass filtered and thresholded to determine the signal audibility. The most important part in this model is the non-linear block shaded in blue. In conventional energy detectors, this non-linear block is typically a second-order operator, i.e., $n = 2$, while in our proposed model, it will be replaced by a higher-order operator, $n > 2$.

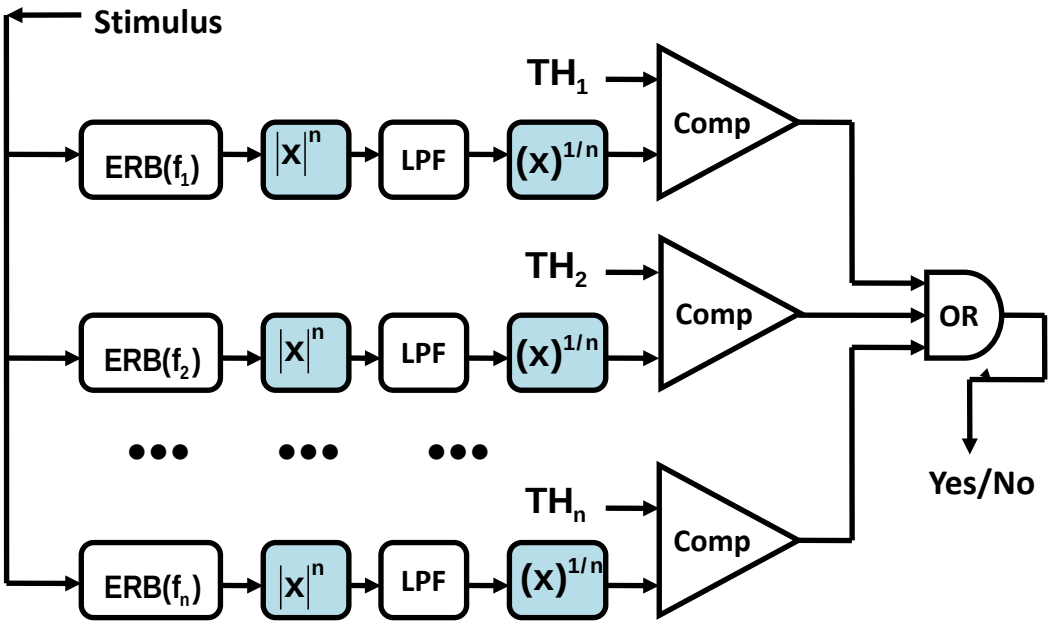


Figure 4-1. Block diagram of higher-order spectro-temporal integration model

4.1.1.1 Front-end auditory filter bank

In this model, an unknown stimulus first goes through an ERB (Equivalent Rectangular Bandwidth) filter bank, which aims to mimic human auditory frequency selectivity. Based on the critical band theory, critical bands are independent auditory processing channels, i.e., only spectral components that fall into the same critical band are processed together. The ERB scale recommended by Moore and Glasberg is close to the critical band scale and better fits the data in notch-noise masking experiments.

The empirically fit equation for ERB is

$$\mathrm{ERB}(f) = 24.7(4.37f/1000 + 1) \tag{4-1}$$

where f is the center frequency in hertz.

The ERB scale is defined as the number of ERBs for each frequency

$$\text{number of ERBs} = 21.4\log_{10}(4.37f/1000 + 1) \tag{4-2}$$

The ERB filterbank in the front-end of the model provides a relatively accurate approximation to the auditory frequency response.

4.1.1.2 Non-linear processing

The non-linear block $|x|^n$ in Figure 4-1 is typically a rectifier or square operator, i.e., $n = 2$. However, by analyzing the subjective thresholds of audio signals in section 3.2, we have shown that these energy-based audibility models do not accurately predict the audibility of transient stimuli. We replace the second-order processing with higher-order processing: $|x(n)|^2 \rightarrow |x(n)|^n$. Intuitively, this gives more emphasis to the transient (larger) parts of the signal. In section 4.1.2, by considering orders from second-order to tenth-order, we find the optimal values of n to predict audibility.

4.1.1.3 Low-pass filtering

As is well known, the human auditory system cannot keep track of the very fine structure of the stimuli. Auditory temporal integration is traditionally modeled as a combination of a square law rectifier and a low-pass filter controlled by a time constant, which is also known as leaky integration. The low-pass filter can be formulated as

$$S'(n+1) = \alpha S(n+1) + (1-\alpha)S'(n) \tag{4-3}$$

where $S'(n)$ is the output of the low-pass filter at time n, and $S(n)$ is the n-th output from the non-linear block. α is a constant related to the low-pass filter time constant T_c, which can be calculated as

$$\alpha = 1 - \exp(-T_i / T_c) \tag{4-4}$$

where T_i is the time interval between successive values.

The time constant T_c is chosen to be the temporal integration limit of the auditory system 200ms for an attack ($S(n) > S'(n-1)$) and 500ms for a release ($S(n) \leq S'(n-1)$).

4.1.1.4 Decision making

Finally, we need to decide whether the observation is signal plus noise or noise alone.

$$H_0 : x(k) = n(k)$$

$$H_1 : x(k) = s(k) + n(k)$$

where $s(k)$ denotes the signal samples, $n(k)$ denotes the noise masker samples and $x(k)$ denotes the observation samples. The output of each band is compared with a pre-determined frequency dependent threshold TH_i as shown in Figure 4-1, which is found based on the subjective thresholds of the narrowband audio signals.

To determine the threshold TH_i in Figure 4-1, we first calibrate the narrowband audio signals at their just audible levels as shown in Table 3-1, we then feed each calibrated narrowband audio to our model and obtain the individual n-th root output. The n-th root outputs of the narrowband audio signals with the same center frequency are averaged to set the pre-determined threshold TH_i for the corresponding ERB band.

Finally the model will output a binary decision for each band. If a stimulus in any ERB band is audible, then the overall stimulus is audible. The lowest SNR for a stimulus to be audible in the white noise masker is the masked audibility threshold for this stimulus.

4.1.2 Model Evaluation and Discussion

Six three-band audio signals were fed into the higher-order integration model shown in Figure 4-1 and compared with the frequency-dependent thresholds TH_i. The prediction errors of both the loudness-based models and the higher-order models are listed in Table 4-1. The prediction error is calculated from the predicted thresholds subtracted by the actual subjective thresholds listed in the Table 3-1.

In Table 4-1, the audio signals C1-C3 and B1-B3 represent two groups of three-band signals, conga and bongo, in three cases, the signal without phase manipulation, and the phase manipulated signal with window size of 64 ms and 512 ms, respectively. The final column of the table shows the average absolute prediction errors of all six signals for each model. The average prediction errors are also plotted in Figure 4-2.

Models that are compared in Table 4-1 are as follows.

Loud-1 denotes the original loudness-based model proposed by Glasberg and Moore (Glasberg and Moore 2005).

Loud-2 denotes a modified loudness-based model, in which instead of averaging short-term loudness over the duration of the signal, instantaneous loudness is low-pass filtered with the same time constant as our higher-order model (200ms for an attack and 500ms for a release) and the output of the low-pass filter is compared with the pre-selected threshold set by a 1 kHz tone at threshold.

HO-2—HO-10 denotes our higher-order spectro-temporal integration model shown in Figure 4-1 using orders from two to ten.

HO-Inf represents the higher-order integration model using infinite order. Let x(i) be the signal samples, N be the total number of samples, and M be the maximum absolute sample value, we have

$$\lim_{n\to+\infty}\left(\sum_{i=1}^{N}|x(i)|^n\right)^{1/n}=\lim_{n\to+\infty}\left(M^n\sum_{i=1}^{N}\frac{|x(i)|^n}{M^n}\right)^{1/n}=M\lim_{n\to+\infty}\left(\sum_{i=1}^{N}\left|\frac{x(i)}{M}\right|^n\right)^{1/n}=M \qquad (4\text{-}5)$$

From the simple derivation in equation 4-5, we can see that the integration model with infinite order is essentially a signal peak detector.

Table 4-1 Prediction errors for three-band audio signals in decibels

Model	C1	C2	C3	B1	B2	B3	Average
Loud-1	3.00	2.00	1.00	1.75	1.50	-0.75	1.67
Loud-2	-2.00	-1.00	2.00	0.25	0.50	-1.75	1.25
HO-2	3.50	-0.5	-2.50	1.75	0.00	-3.75	2.00
HO-3	1.00	-0.5	-1.50	0.75	0.50	-2.75	1.17
HO-4	0.00	0.00	-1.00	-0.25	1.00	-2.25	0.75
HO-5	-0.50	0.00	-1.00	-0.75	1.50	-1.75	0.92
HO-6	-0.50	0.00	-1.00	-0.75	1.50	-1.75	0.92
HO-7	-1.00	0.00	-0.50	-1.25	1.50	-1.75	1.00
HO-8	-1.00	0.00	-0.50	-1.25	2.00	-1.75	1.08
HO-9	-1.00	0.00	-0.50	-1.25	2.00	-1.75	1.08
HO-10	-1.00	0.00	-0.50	-1.75	2.00	-1.75	1.17
HO-Inf	-1.00	0.00	-0.50	-2.25	2.50	-1.75	1.33

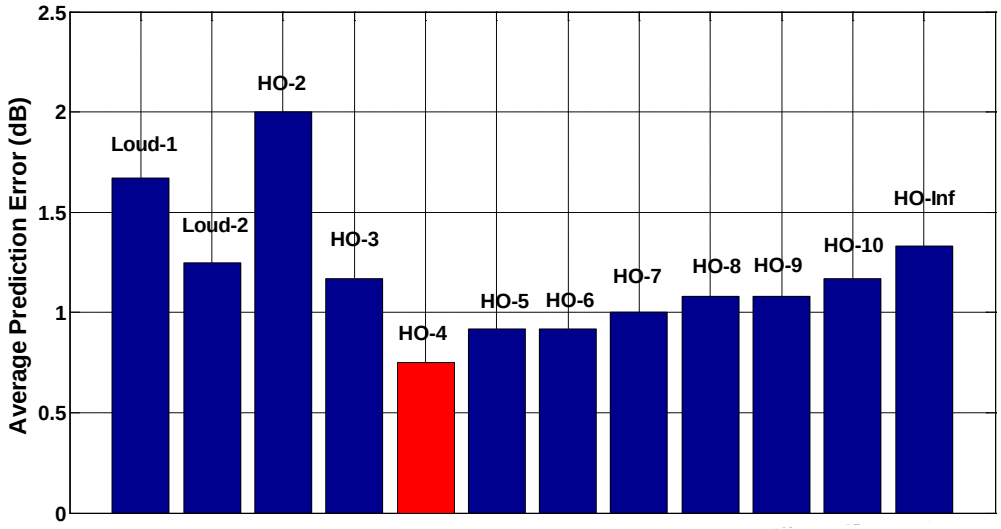


Figure 4-2. Average prediction error for three-band signals in decibels.

The observations from Table 4-1 and Figure 4-2 are as follows.

Loudness-based model versus higher-order models: from Figure 4-2, we can see that, in terms of average prediction error, the higher-order models HO-3—HO-10 outperform the two loudness-based models Loud-1 and Loud-2. Overall, the fourth-order model is the best audibility predictor and the second-order model is the worst one on our experimental data.

Loudness-based models: for two loudness-based models Loud-1 and Loud-2, simply by replacing the average of short-term loudness with the maximum of the low-passed instantaneous loudness as an alternative predictor, the average prediction error is decreased by 0.42 dB and the individual prediction error for the most impulsive signal C1 is decreased by 5 dB, which indicates that the average of loudness over the whole duration of a signal seriously underestimates the audibility of an impulsive signal that has energy highly concentrated in a brief time interval.

Models of all orders: Among all the higher-order models HO-2—HO-Inf, the fourth-order model outperforms the others. When the order equals two, the predictor acts as a single-band energy detector with limited memory. As the order increases, the predictor emphasizes more of the signal peaks. It turns out that some moderate order between the second order and the infinite order (peak detection) best predicts all the tested audio signals, which indicates both the signal energy and signal peak contribute to the signal audibility.

4.2 Model Extension I

As mentioned in Chapter 2, the complex tone experiments (Green 1958, Buus et al. 1986, Van den Brink and Houtgast 1990) show that, equally detectable signal components in isolated critical bands improve the detectability of the overall signal.

However, our basic model in the previous section predicts the audibility based on the most detectable critical band and hence cannot explain the decreased audibility thresholds of complex tones. To overcome this problem, an extended model that incorporates auditory spectral integration is proposed in this section.

4.2.1 Extended Model with Spectral Integration Based on Probability Summation

As stated in Chapter 2, the independent threshold model assumes statistically independent observations in each auditory channel and makes decisions based on probability summation across auditory channels, which is shown to be a better balance of explaining the detectability improvement of wideband signals and being consistent with the critical band theory than the multi-band integration model. As a result, the independent threshold spectral integration model is incorporated in our extended audibility prediction model.

The rule used to combine decisions in the independent threshold model is to make a positive response when any one of the several observations is positive. Under this rule, the probability of detection based on observations in M critical bands is

$$P = 1 - \prod_{i=1}^{M} (1 - p_i) \tag{4-6}$$

where p_i denotes the detection probability of the i-th critical band.

The psychometric function proposed by Green and Luce (Green and Luce 1975) is most commonly used to calculate the auditory detection probability for each critical band

$$p_i = 1 - \exp(-\alpha I^{\beta}) \tag{4-7}$$

I represents the signal intensity, and α and β are both free parameters.

Let I_{TH} and p_{TH} be the signal intensity and detection probability at threshold for a given critical band, and I_i and p_i be a random intensity and the corresponding detection probability for the same signal within the same critical band, we have

$$p_{TH} = 1 - \exp(-\alpha I_{TH}^{\beta}) \tag{4-8}$$

$$p_i = 1 - \exp(-\alpha I_i^{\beta}) \tag{4-9}$$

From equation (4-8), we have

$$-\alpha = \frac{\ln(1 - p_{TH})}{I_{TH}^{\beta}} \tag{4-10}$$

Substituting equation (4-10) into equation (4-9), we have the individual probability p_i as a function of the signal intensity I_{TH}, the detection probability at threshold p_{TH} for the same critical band and the pre-selected parameter β as follows.

$$p_i = 1 - \exp\left(\ln(1 - p_{TH}) \cdot \left(\frac{I_i}{I_{TH}} \right)^{\beta} \right) \tag{4-11}$$

where the frequency dependent I_{TH} and p_{TH} can be determined by the single-band signals at thresholds.

The free parameter β can be set based on the value that best fits the experimental data. The value of β controls the amount of emphasis that puts on the most detectable auditory frequency channels. The overall extended model is shown in Figure 4-3.

4.2.2 Model Evaluation and Discussion

$\beta = 2$, 4, and 6 are used to test the extended model shown in Figure 4-3. The prediction errors on our experimental data are listed in Table 4-2. By putting more emphasis on the most detectable auditory frequency channels and attenuating the

contribution of the least detectable auditory frequency channels, $\beta = 6$ performs the best in predicting the audibility thresholds of all the tested audio signals.

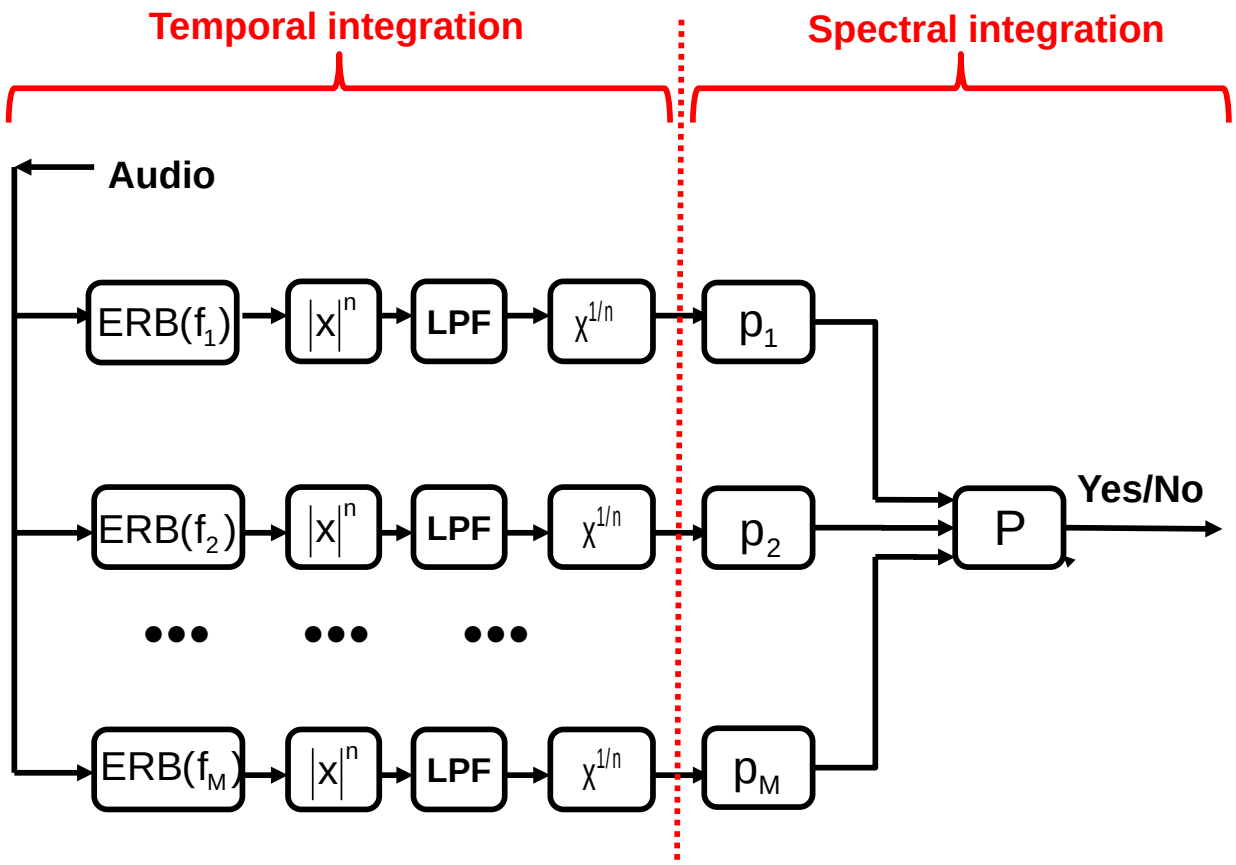


Figure 4-3. Block diagram of extended higher-order integration model for audibility prediction

By comparing Table 4-1 and Table 4-2, the extended model with $\beta = 6$ performs very close to the basic model. In this evaluation, since the tested audio signals have power distributed very differently across auditory frequency channels, the extended model with $\beta = 6$ is essentially reduced to the basic model, making audibility decision based the most detectable auditory channel solely for the tested audio signals.

The advantage of this extended model is in predicting the decreased audibility thresholds of signals that have equivalent or comparable detectability in multiple auditory channels. The value of β can be optimized with more subjective data.

Table 4-2 Average prediction error in decibels (dB) for different beta parameters

Beta	2	4	6
HO-2	2.75	2.08	2.00
HO-3	2.58	1.08	1.17
HO-4	2.58	0.92	0.75
HO-5	2.67	1.08	1.08
HO-6	2.67	1.17	0.92
HO-7	2.67	1.17	1.08
HO-8	2.58	1.17	1.08
HO-9	2.58	1.25	1.08
HO-10	2.58	1.25	1.25

4.3 Model Extension II

4.3.1 Analogy between Audibility Perdition and Hearing Loss Prediction

The human auditory system has evolved so that it can detect sounds with displacements in the sub-angstrom range while, at its upper limits, it can faithfully encode sounds over a dynamic range of 120 dB SPL (Henderson and Hamernik 1995).

Incoming sound travels down the external ear (pinna and meatus) and causes the tympanic membrane to vibrate. These vibrations are transmitted through the middle ear by three ossicles to the oval window, which is the opening of the spiral-shaped cochlear. When the oval window is set in motion by the incoming sound, a pressure difference is applied across the basilar membrane (BM) and causes the BM to move. When the BM moves up and down, a shearing motion is created between the BM and the tectorial membrane. As a result, the hairs at the top of the outer hair cells are displaced, which leads to the excitation of the inner hair cells and leads in turn to the generation of action potentials in the neurons of the auditory nerve (Moore 2003).

However, the process cannot continue indefinitely. Over-stimulation might cause cellular fatigue in our auditory system due to a change of the metabolic process in the inner ear and result in a temporary threshold shift (TTS). With greater exposure to sounds, the auditory periphery, or cochlea, progressively deteriorates, which results in mechanical damage to the auditory system. The damage caused by this kind of intense noise is pervasive and affects virtually all of the cellular subsystems of the inner ear (sensory cells, nerve endings and vascular supply). Sounds such as gunfire and certain industrial impacts—peak levels greater than 125 dB—are especially hazardous to the cochlea and very likely to cause permanent threshold shift (PTS). (Henderson and Hamernik 1995).

Although the mechanisms for audibility of weak sounds and temporary/permanent threshold shifts caused by moderate to intense sounds are very different in terms of biology, they do share some common characteristics. The characteristics are level dependence, frequency dependence, and leaky integration and temporal effects discussed in the following sections.

4.3.1.1 Experiments on stationary sounds

Frequency place theory and half-octave shift : The frequency place theory proposed by Von Bekesey states that the traveling wave forms a displacement envelope on the basilar membrane that peaks at different places along the cochlear according to the excitation frequency (Bekesey 1960). As a result of the place theory, if the stimulus, pure tone or narrowband noise, is low in level, audibility is usually restricted to a narrow range near the stimulus frequency. At high exposure levels, the signal energy at the center frequency and temporal threshold shift (TTS) are also correlated, except that the TTS spreads predominantly toward the high frequencies and

the maximum hearing loss typically shifts to a point one-half octave above the center frequency of the stimulus (Quaranta et al. 1998).

Level dependent asymptotic threshold shift (ATS): For long exposures, e.g., 16-24 hours, Temporary threshold shifts (TTS) increase for about 8 h and then reach a plateau called asymptotic threshold shift (ATS), which sometimes is interpreted as the upper value of PTS that can occur from the same exposure that lasts for years(Quaranta et al. 1998). Same as audibility thresholds, there is a clear relation between the sound level and the ATS. Systematic experiments on TTS by Mills et al.(1979) showed, in the frequency region of greatest loss, ATS increased about 1.7 dB for every 1 dB increase in the level of the noise above a critical level, which can be described as

$$\mathrm{ATS} = 1.7(\mathrm{OBL} - \mathrm{C}) \tag{5-1}$$

where OBL is the octave band noise level in decibels and C is the frequency dependent critical level (Mills et al. 1979, Melnick 1991).

Exponentially growth and recovery of TTS: One of the most intensive studies on TTS was conducted by Mills and his colleagues. In their study, 60 human subjects were divided into 8 groups and exposed to a diffuse sound field for 16-24 h to an octave-band noise centered at 4, 2, 1, or 0.5 kHz. Sound-pressure levels were varied from 75 dB to 88 dB on different exposure occasions. Temporary threshold shifts (TTS) increased for about 8 h and then reached a plateau or asymptote. After termination of the exposure, recovery to within 5 dB of pre-exposure thresholds was achieved within 24 h or less. Recovery can be described by a simple exponential function with a time constant of 7.1 h. Mills et al. also summarized the relation between TTS and exposure

duration as simple exponential functions using their own data as well as the previously reported TTS data (Mills et al. 1979). Similar to the case of soft sounds, the growth and recovery of TTS caused by moderate to high level octave band noises is also a temporal integration process. TTS can also be characterized by leaky integrators with different time constants for growth and recovery, as in the case for audibility.

Given the analogy between audibility and TTS in level dependence, frequency dependence and temporal integration for stationary sounds, we believe that our higher-order audibility prediction model could be extended to TTS prediction for non-stationary sounds using a leaky integrator with different time constant for growth and recovery.

4.3.1.2 Inadequacy of existing safety standards for non-stationary sounds

The data sets used to determine the degree of hearing loss caused by noise were collected in the late 1960s and early 1970s in predominantly white, adult male populations that were exposed to industrial noise. These data were instrumental in developing standards (ISO 1999, ANSI S3.44-1996) to describe the relationship between noise exposure and noise-induced permanent threshold shift (NIPTS) and regulations by the Occupational Safety and Health Administration (OSHA 1983) and safety recommendations by the National Institute for Occupational Safety and Health (NIOSH 1998). Current understanding from these data is that a maximum exposure of 85 dB, A-weighted (dBA), for an 8-hour daily exposure over a working lifetime of 40 years results in roughly 8% of exposed persons having a hearing handicap, owing to the wide variability seen in susceptibility to noise-induced hearing loss across individuals (Fligor 2006).

Exchange rate of exposure level and duration: As mentioned in Chapter 2, for stationary signals less than 200ms, there is a clear duration-intensity tradeoff for

audibility threshold and sound duration. Risk assessment for hearing loss suggested by current standards is also equalized using a similar “exchange rate” that is an increment, or decrement, of decibels that requires halving, or doubling, the allowable exposure time, respectively. Use of the exchange rate allows for determining the Time-Weighted Average (TWA) exposure level to compare risk for noise-induced permanent threshold shift (NIPTS) despite different durations and levels of exposure. Currently, the OSHA limits the maximum permissible exposure limit to 90 dBA for an 8-hour TWA, using a 5-dB exchange rate. While not exceeding this permissible exposure limit, workers exposed to 85 dBA, 8-hour TWA, must be enrolled in a Hearing Conservation Program, including annual audiometry, use of hearing protection devices, and education on risks and prevention of NIHL. This 85 dBA, 8-hour TWA (with 5-dB exchange rate) is considered the Action Level by OSHA (OSHA 1983). In contrast, NIOSH recommends an exposure limit of 85 dBA for an 8-hour TWA, using a 3-dB exchange rate (NIOSH 1998). Most developed western countries limit worker exposure to 85 dBA, 8-hour TWA using the 3-dB exchange rate. And the 3-dB exchange rate is exactly based on equivalent energy calculation (Fligor 2006).

Energy-based PTS prediction in current standard (ISO 1990:1999) : Similar to energy detectors for audibility threshold prediction, the measure of exposure to noise for a population at risk is the averaged A-weighted sound exposure (time-integrated squared sound pressure) $E_{A,T}$, and the related the equivalent continuous A-weighted sound pressure level $L_{Aeq,T}$ over an average working day (assumed to be of 8 h duration), for a given number of years of exposure, which are both second-order energy-based measures restricted in each octave band.

The prediction method presented in the ISO standard is based primarily on data collected with essentially broad-band, steady non-tonal noise. For tonal noise and/or impulsive/impact noise, the standard only suggests they are about as harmful as a steady non-tonal noise that is approximately 5 dB higher in level (ISO 1990:1999). This guidance doesn't take the specific signal time structure into consideration and it offers very limited help for practical use.

4.3.1.3 Hearing risk associated with signals of different time structures

As verified in Chapter 3, for non-stationary sounds of equivalent energy, the more impulsive the signal is, the lower its audibility threshold is, and thus the easier for human auditory system to detect it. There is growing evidence that shows hearing loss induced by intense sounds is also affected by the signal temporal structure (Strasser et al. 1999, Hamernik et al. 2001 and 2003) – for signals of same energy above some critical level, the more impulsive the signal is, the more damaging to human auditory systems.

TTS Experiments on human subjects: In order to disclose the actual physiological responses to exposures which varied with respect to the temporal structure and the semantic quality of sounds, Strasser et al. conducted a series of tests where physiological costs associated with varying exposures were measured audiometrically. In this study, 10 subjects with normal hearing participated in a test series with four exposures (white noise, industrial noise, heavy metal music, and classical music) , which were characterized by the same level 94 dB(A), for an hour. The physiological responses to the four exposures were assessed by Integrated Restitution TTS (IRTTS), which is computed as the integral of the regression function TTS (t) from 2 min after the exposure to the point t. The IRTTS is a numeric value for

the total threshold shift (in dB x min) which has to be "paid" by the hearing in physiological costs for the exposure. The results showed that, the industrial noise had an IRTTS value of 631 dBmin in relation to 424 dBmin qualified as responses to white noise brought about an increase of approximately 50% in total physiological cost. Heavy metal music was also associated with tremendous physiological cost (637 dBmin). Classical music was accompanied with the lowest physiological cost (160 dBmin). The results indicate time structure does affect the noise induced hearing risk and energy-equivalent approach could lead to dangerously wrong assessment of hearing risk (Strasser et al. 1999).

PTS Experiments on animal subjects : The most systematic studies on hearing loss induced by noise signals of equivalent energy were conducted by Hamernik and his colleagues. In their studies, groups of chinchillas were exposed to either a Gaussian noise or one of the non-Gaussian noises at 100 dB(A) SPL. All exposures had the same total energy and approximately the same flat spectrum but their statistical properties were varied to yield a series of exposure conditions that varied across a continuum from Gaussian through various non-Gaussion conditions to pure impact noise exposures. Trauma, as measured by asymptotic and permanent threshold shifts (ATS, PTS) and by sensory cell loss, was greater for all of the non-Gaussian exposure conditions. In their chinchilla model, PTS and outer hair cell loss were monotonically related to the signal kurtosis $\beta(t)$. Besides, the frequency-specific OHC loss produced by non-G noise exposures is well correlated the frequency specific kurtosis $\beta(f)$. In conclusion, in spite of the same signal energy and spectra, the more transient the signal was, the more serious trauma developed on the animal subjects (Hamernik et al. 2001 and 2003).

To sum up, for stationary sounds, audibility and hearing loss are both a temporal integration process in individual octave bands with exponential growth and recovery. Existing prediction models for audibility and hearing loss are both based on equivalent energy concept – same energy causes same detectability/damage. For non-stationary sounds, both audibility and hearing loss are affected not only by the signal energy but also the signal temporal structure.

4.3.2 Extended Model Suggested for Hearing Loss Prediction

Given the similarity between audibility and hearing loss prediction for stationary sounds, extending our higher-order integration model to hearing loss prediction can potentially lead to a more accurate prediction for hearing loss induced by non-stationary signals. A possible framework to predict the lowest signal level that causes hearing loss for a given audio signal is illustrated in Figure 4-4.

The main differences between this hearing loss prediction model in Figure 4-4 and the audibility prediction model in Figure 4-1 are in the design of the low-pass filter and the pre-determined thresholds for individual auditory channels, which are shaded in Figure 4-4. As mentioned in the previous section, compared with audibility prediction, temporary/permanent hearing loss takes a lot longer time to build up, hence, a temporal integration processing with a longer time constant needs to be embedded in the design of the low-pass filter. In addition, the pre-determined thresholds for each auditory channel need to be modified based on the experimental data for hearing loss. Given more subjective data, the higher order for auditory temporal integration can also be optimized for hearing loss prediction.

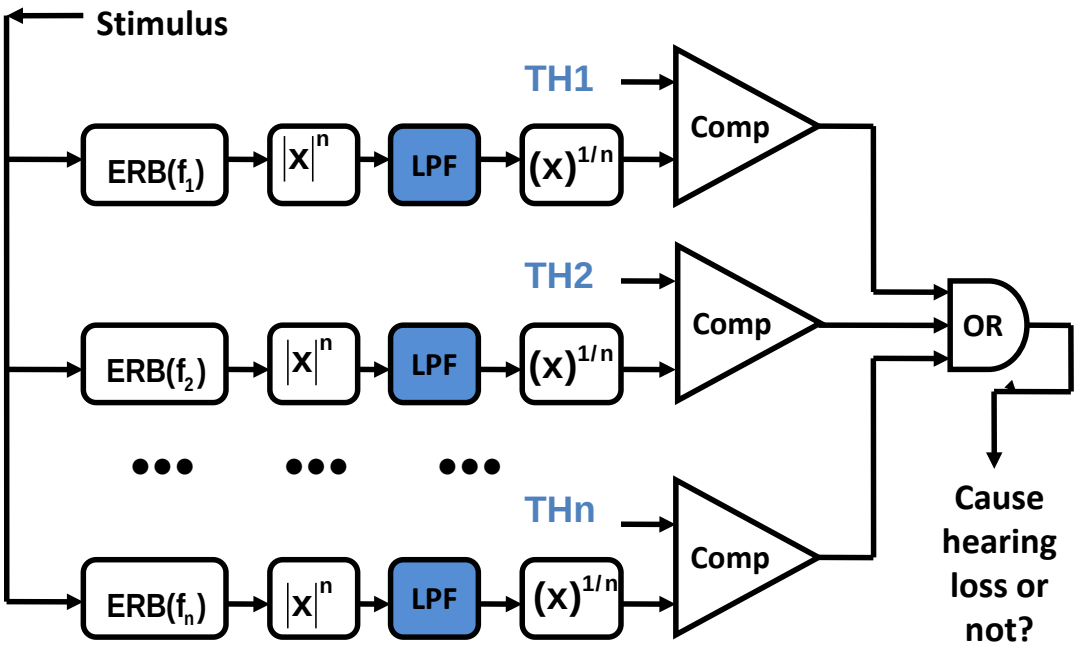


Figure 4-4. Block diagram of the higher-order integration model for predicting the threshold of a given audio signal at which possible hearing loss will be induced

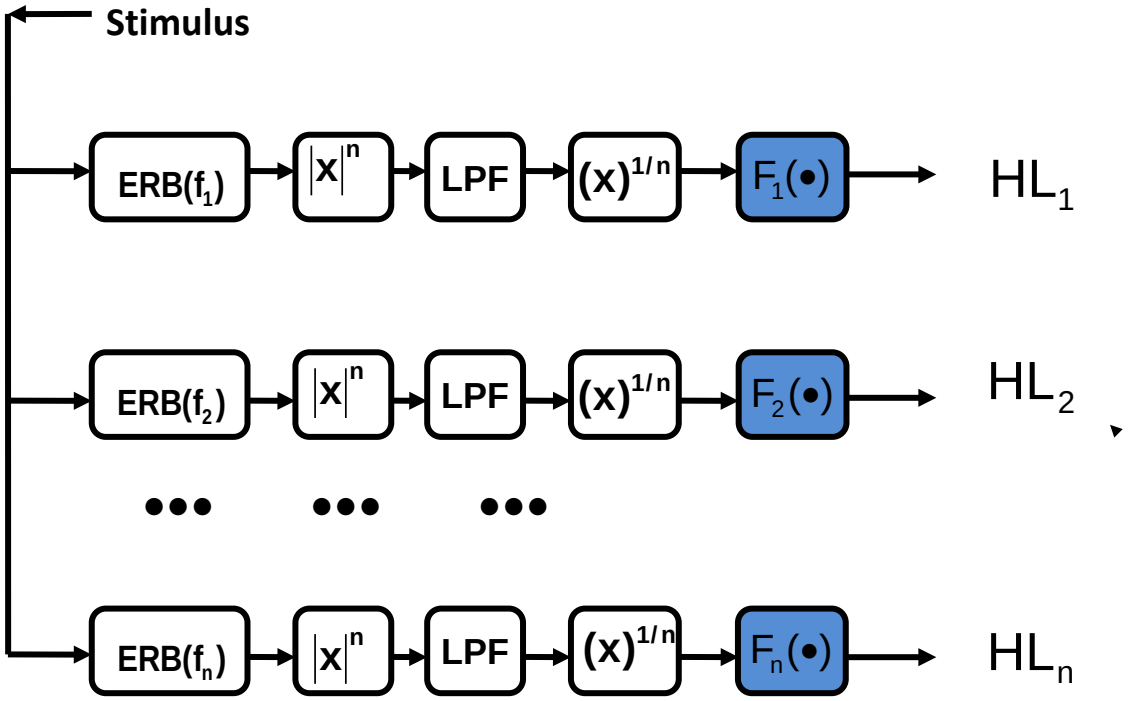


Figure 4-5. Block diagram of higher-order integration model for predicting the possible hearing risk that is induced for a given audio signal at a fixed signal level

Alternatively, a similar framework can also predict the potential hearing risk induced by a given audio signal at a pre-selected level. A possible modified model is illustrated in Figure 4-5. In this model, the output of the higher-order integrator is fed into a pre-determined function $F_i(\bullet)$ and the potential hearing loss for each critical band induced by the given audio signal is calculated accordingly. With sufficient subjective data, the relation between the output of the higher-order integration and the consequent hearing loss can be formulated for each critical band.

4.4 Summary

We propose a higher-order spectro-temporal integration model for predicting the audibility thresholds of audio signals. By using higher-order temporal integration in individual ERB bands and making decision on the most audible ERB band, our basic prediction model outperforms existing loudness-based and energy-based models. With more data, we will determine the precise order of the model.

For better generalization of our basic model, a spectral integration process based on probability summation across all auditory channels is incorporated into the basic model. The prediction performance of this extended model is shown to be very close to our basic model with our experimental data.

The analogy between audibility and hearing loss prediction has been well summarized. Given the similarity shared by both audibility and hearing loss prediction, two hearing loss prediction models extended form our basic higher-order integration model have been proposed. The refinement and validation of these two models require more subjective data.

CHAPTER 5
HIGHER-ORDER LEVEL ESTIMATION FOR AUDIO DYNAMIC RANGE CONTROL

In Chapter 2, the human auditory system has been shown to be more sensitive to transient signals than stationary signals given the same energy. The conventional dynamic range control (DRC) algorithm is based on second-order level estimates (i.e., energy or root-mean-squared value). Since the second-order estimate cannot adequately characterize the auditory perception of non-stationary audio signals, the conventional second-order DRC algorithm is extended using a higher-order level estimate in this study. The perceptual quality and the dynamic range reduction effectiveness are evaluated for both second-order and higher-order DRC algorithms. Evaluation results show that higher-order DRC algorithms with moderate-size analysis windows offers the best balance of perceptual quality and dynamic range reduction.

5.1 Dynamic Range Control for Hearing Protection

Dynamic range control (DRC) has been widely used in the design of hearing aids, radio and TV broadcasting, teleconferencing and other acoustical applications. As modern personal media players (PMP) with mass-storage capacities, long battery life, and high output levels, become more and more popular, music-induced hearing loss is becoming more of a social and clinical problem. Listeners often set volume levels based on the intelligibility or detectability of the softest sounds in the audio signals. For audio signals with wide dynamic range at a given volume level, when the softest sounds are adequately audible, the loudest sounds might be overwhelmingly intense. Studies have shown that the loudest and most transient sounds cause the most hearing risk. In order to determine the actual physiological responses to exposures which varied with respect to the time structures, Strasser and his colleagues tested ten normal hearing human

subjects with a series of four exposures (white noise, industrial noise, heavy metal music, and classical music), which were characterized by the same level 94 dB(A). The experimental data showed that the heavy metal music induced the most integrated temporary threshold shifts on human subjects, followed by the industrial noise, white noise and classical music. These results indicate that acoustical signals that are rich in intense transients induced the most hearing risk at a given sound level (Stresser et al 1999). Hamernik and his colleagues conducted the most systematic studies on hearing loss induced by noises of equivalent energy. In their studies, groups of chinchillas were exposed to a family of noise exposures that had the same sound level 100 dB(A) and approximately the same flat spectra but very different time structures. Their results showed that the more transient the signal was, the more serious was the trauma developed in the animal subjects (Hamernik and Qiu 2001). Since the intense transient signals are more damaging to the auditory system at a given sound pressure level, a delicate DRC algorithm that balances the dynamic range reduction and perceptual concerns would be beneficial to protect the hearing of music listeners.

As recorded audio signals are mostly used for entertainment purposes, a primary concern for reducing the dynamic range of audio signals is the consequent quality degradation. Conventional DRC algorithms based on second-order measurements (root-mean-squared value or energy) normally provide a trade-off of perceptual quality and dynamic range reduction. This study will extend the conventional second-order DRC using a higher-order level estimate and further investigate the impact of temporal structure on the optimal order selection for a given audio signal.

In the rest of this chapter, our higher-order DRC system will be first introduced briefly. The objective audio quality and the dynamic range reduction for both our proposed approach and the existing approach will then be evaluated. The relation between the optimal DRC order and the temporal structure for a given audio signal will be investigated as well. Finally, section 5.3 summarizes the results and offers suggestions for future work.

5.2 High-Order Dynamic Range Control

5.2.1 Model Description

There are generally two types of topologies for DRC systems, feedback and feed-forward. Feedback systems use the output signal to control the gain. Since the signal levels must reach the system output before the compensation gains can be generated, the feedback systems allow overshoots to occur at the output and thus do not handle transients well. The feed-forward systems overcome this limitation by using the input signal levels to control the gains and better suppress the transients. Feed-forward designs also have an advantage in system stability (Schneider and Brennan 1997). Hence, the feed-forward topology is chosen for our DRC system. The block diagram of the complete feed-forward DRC system used in this study is shown in Figure 5-1.

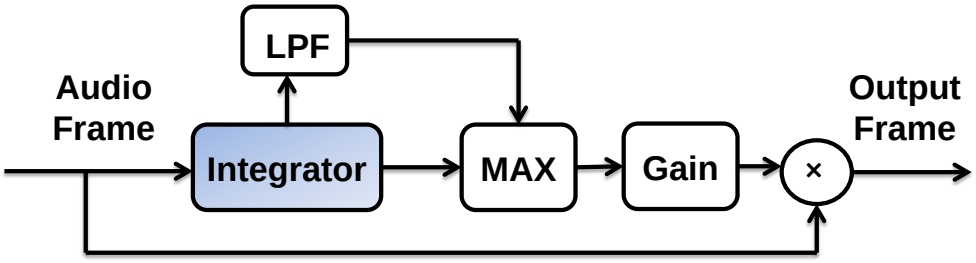


Figure 5-1. Block diagram of the dynamic range control system

Pitch and timbre are the most important aspects in music perception. Variations in pitch are central to our experience of melody, harmony, and key. Timbre allows us to

distinguish one instrument from another. The physical variables that contribute to our experience of timbre include the spectral energy distribution (i.e., bandwidth and concentration), temporal envelope, and transient components of a tone. A multiple frequency channel DRC is widely used for speech signals. However, music requires that the balance between the lower-frequency fundamental energy and the higher-frequency harmonic energy remains intact to achieve optimal sound quality. An imbalance in the amplification of low- and high-frequency channels will always affect timbre, may also lead to problems for musical pitch perception and ultimately distort the intent of the musician(Chasin 2003). As a result, a single frequency channel processing is more common for music to preserve timbres, because the short-term spectral distribution is unaltered. Some studies attempted to design multi-channel DRC systems for music signals to achieve more flexibility in frequency equalization, but they still need to ensure that the adjacent channel controllers were set similarly (Schmidt and Rutledge 1996). For simplicity, we chose to use a single frequency band for our DRC system.

The input/output mapping function is usually determined by three aspects: the compression ratio, the threshold knee-point (TK) where the DRC algorithm starts, and the smoothness of the transition between different linear segments. To simplify the system, we use only the TK point to control the output dynamic range, in other words, the audio frames with levels higher than the TK point will be reduced to a fixed level otherwise they will be linearly amplified. The simplified input-output mapping function is shown in Figure 5-2. To calculate the TK point, we first construct a histogram of the audio levels across all the frames, obtain the cumulative density function (CDF), and pick a level at a certain cumulative probability as the TK point. The higher the selected

cumulative probability, the less the signal will be compressed, which normally leads to better perceptual quality and less dynamic range reduction.

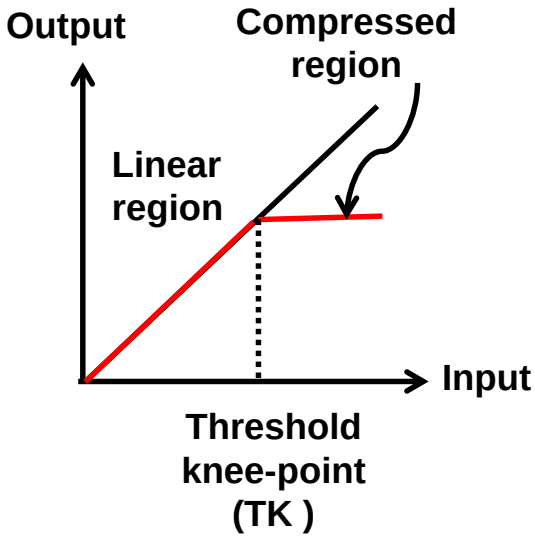


Figure 5-2. Illustration of the input-output mapping function for the DRC algorithm in this study

The key part in our DRC design is the level estimation in the integrator block for each frame as shaded in Figure 5-1. Conventional dynamic range controllers estimate the compensation gains based on the second-order measurement root-mean-squared value (RMS) or energy. Our study in Chapter 3 has shown that the human auditory system is more sensitive to impulsive signals than stationary signals given the same energy. There is also growing evidence that shows hearing loss induced by intense sounds is affected by the signal temporal structure (Hamernik and Qiu 2001). These results cannot be adequately explained by energy-based (second-order) models. On the other hand, our higher-order spectro-temporal integration model outperforms the existing second-order models in predicting the audibility of non-stationary audio signals (Chapter 4) and higher-order statistics such as kurtosis have been shown to be correlated to the hearing loss induced by non-stationary noise (Hamernik and Qiu

2001). Inspired by the application of higher-order processing in auditory modeling of non-stationary signals, we replaced the second-order level estimate in the conventional DRC with a higher-order level estimate,

$$S(k) = \left(\sum_{i=1}^{M} |x(i)|^2 \right)^{1/2} \rightarrow S(k) = \left(\sum_{i=1}^{M} |x(i)|^N \right)^{1/N} \quad (5\text{-}1)$$

where $s(k)$ is the instantaneous level for the k-th frame, $x(i)$ is the i-th sample in a given frame, M is the number of samples in each frame, and N (N>2) is the higher order for the non-linear operation.

The instantaneous levels across frames are then smoothed by a first-order low-pass filter as follows.

$$S'(n) = \alpha S'(n-1) + (1-\alpha) S(n) \quad (5\text{-}2)$$

where $S'(n)$ is the smoothed output of the low-pass filter for the nth frame, and α is a constant related to the low-pass filter time constant $T_C = 5\,\text{ms}$

$$\alpha = 1 - \exp(-T_i / T_C) \quad (5\text{-}3)$$

T_i is the time interval between the start points of successive frames. The final estimated level for the n-th frame is determined by the maximum of the smoothed level $S'(n)$ and the instantaneous level $S(n)$ for each value of n,

$$L(n) = \max(S'(n), S(n)) \quad (5\text{-}4)$$

which ensures that the estimated levels across frames are smoothed for relatively stationary frames and the impulsive frames can be quickly suppressed as well.
The hope is that the intense transient parts of the audio signals can be further reduced by the higher-order non-linear processing and the degradation in perceptual quality can also be minimized. The objective quality evaluation will be addressed in section 5.2.2.

5.2.2 Evaluation and Results

Forty-five audio recordings are selected from the European Broadcasting Union (EBU) sound quality assessment materials (SQAM) (European Broadcasting Union 2008) which covers a variety of audio contents such as single instruments, vocals, solo instruments, vocal and orchestra, orchestra and pop music. All the audio recordings are originally sampled at 44100 Hz and converted to a 48000 Hz sample rate.

The audio signals are first segmented into 50% overlapped frames with a pre-determined frame size ranging from 128 samples/frame (2.7 ms/frame) to 4096 samples/frame(85.3 ms/frame). The signal levels for each frame are calculated using order N =2, 4, 6, and infinity as shown in equation 5-1, which are then low-pass filtered. To calculate the threshold knee point (TK) point, the corresponding cumulative probability functions of the signal levels across all frames are obtained through their histograms. The TK point is varied from the signal level at 10% cumulative probability to the signal level at 95% cumulative probability with a spacing of 5%. The processed audio signals are normalized to the same overall root-mean-squared value as the corresponding unprocessed audio signal. The peak-to-RMS ratio reduction relative to the unprocessed signal are calculated every 2 seconds and averaged out.

The perceptual quality of the processed audio signals is evaluated by the basic version of ITU standard objective measure of perceived audio quality, a.k.a., PEAQ (International Telecommunication Union 2001, Kabal 2002). The basic version of this standard used in this study (Figure 5-3), calculates a number of psycho-acoustical measures, including the signal bandwidth, noise loudness, noise-to-mask ratio, modulation difference, detection probability and error harmony structure based on the excitation patterns from auditory pre-processing. All of these psycho-acoustical

measures are input to a well-trained neural network with a single hidden layer with three nodes to give a measure of the quality difference between the unprocessed signal and the DRC processed signal. The objective difference grades (ODG) that output from the neural network range from 0 to -4, where 0 corresponds to an imperceptible impairment and -4 to an impairment judged as very annoying.

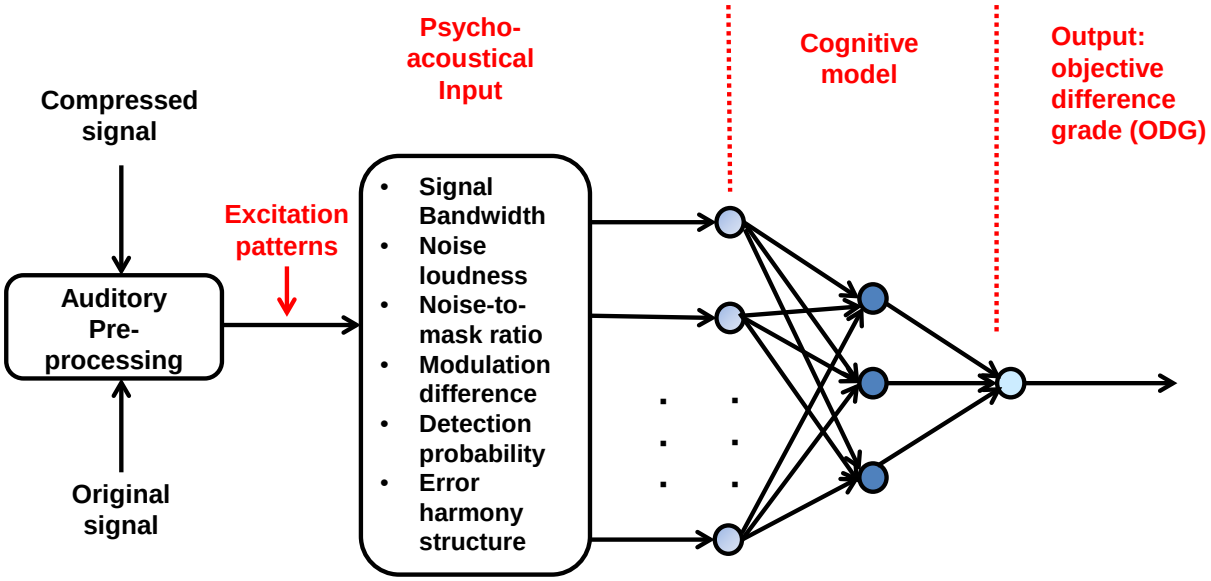
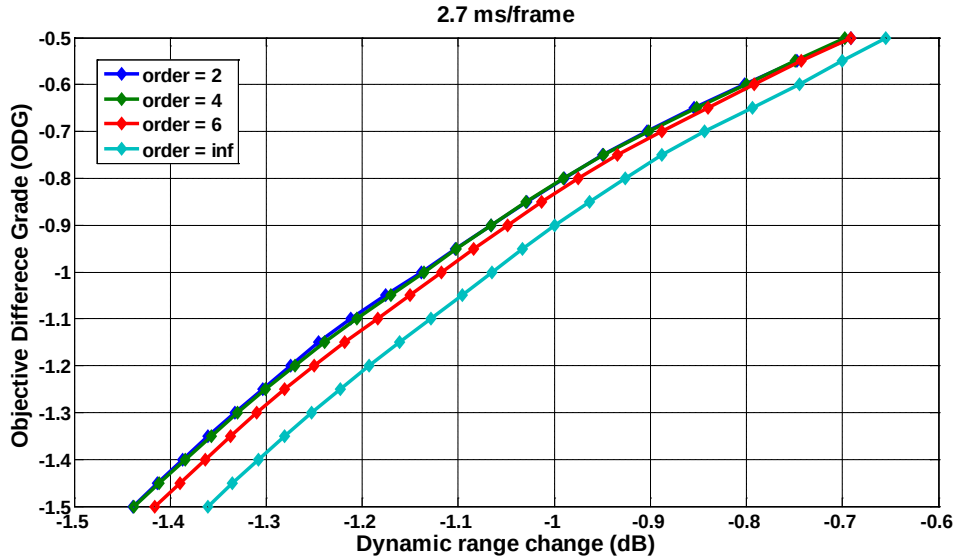


Figure 5-3. Block diagram of ITU-R BS.1387-1perceptual evaluation of audio quality (PEAQ)

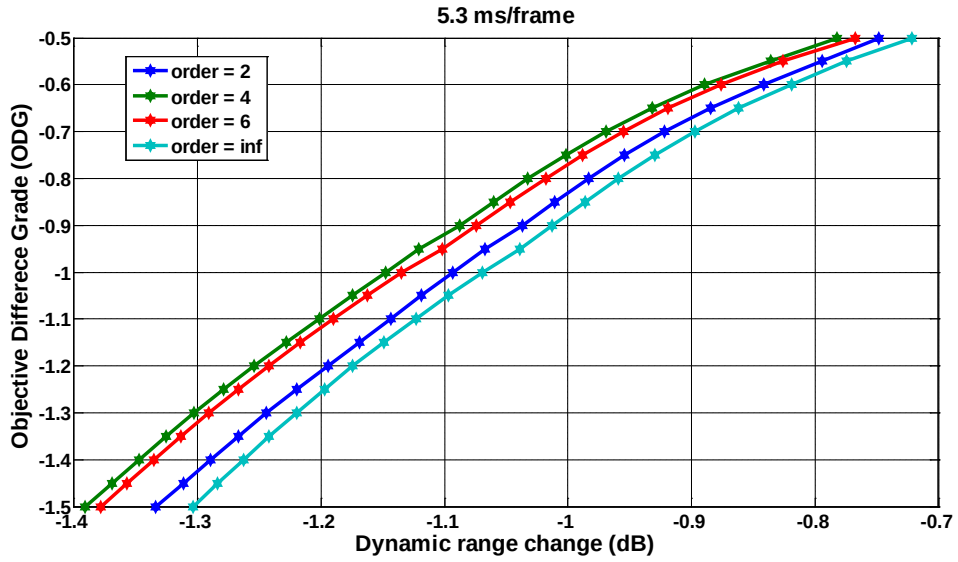
The dynamic range reduction of each processed audio signal is interpolated for objective difference grades from -0.1 to -1.5 with a spacing of -0.05. The interpolated dynamic range reduction values of all the audio signals for a given analysis window size are averaged and an operational curves of average dynamic range reduction and objective difference grades is generated for each analysis window size. The reason why objective quality scores of -0.1 to -1.5 is of interest in this study is that the DRCs using long windows tend to operate in the region of high quality and lower dynamic range reduction. For DRCs using really long windows, the low quality region of the operational curves are heavily dependent on extrapolation and thus not very reliable. To compare

the operational curves of dynamic range reduction and audio quality across windows, objective difference grade of -0.1 to -1.5 is a shared operational region of DRCs using different window sizes. In addition, processed audio signals with objective difference grades within the range of -0.1 to -1.5 are generally considered as imperceptible to slightly perceptible degradation relative to the unprocessed signals, which is necessary for audio entertainment.

The operational curves of average dynamic range reduction and objective audio quality are shown in Figure 5-4. In each subplot, for all four types of DRC algorithms from the second order to the infinite order, the objective perceptual quality generally decreases as the signal dynamic range decreases. When the dynamic range is reduced severely, the audio quality is seriously degraded; when the dynamic range is changed very little, the audio quality is almost the same as the original. Between these two extremes, the leftmost DRC offers the best balance of dynamic range reduction and objective perceptual quality compared to the other DRC algorithms. The interpretation for Figure 5-4 is that, given the same objective perceptual quality, the signals processed by the leftmost DRC have larger dynamic range reduction than the others, which can better protect the hearing of audio listeners. On the other hand, given the same dynamic range reduction between the input and output, the signals processed by the leftmost DRC offer better perceptual quality than the other signals, which can better satisfy the fidelity requirement.

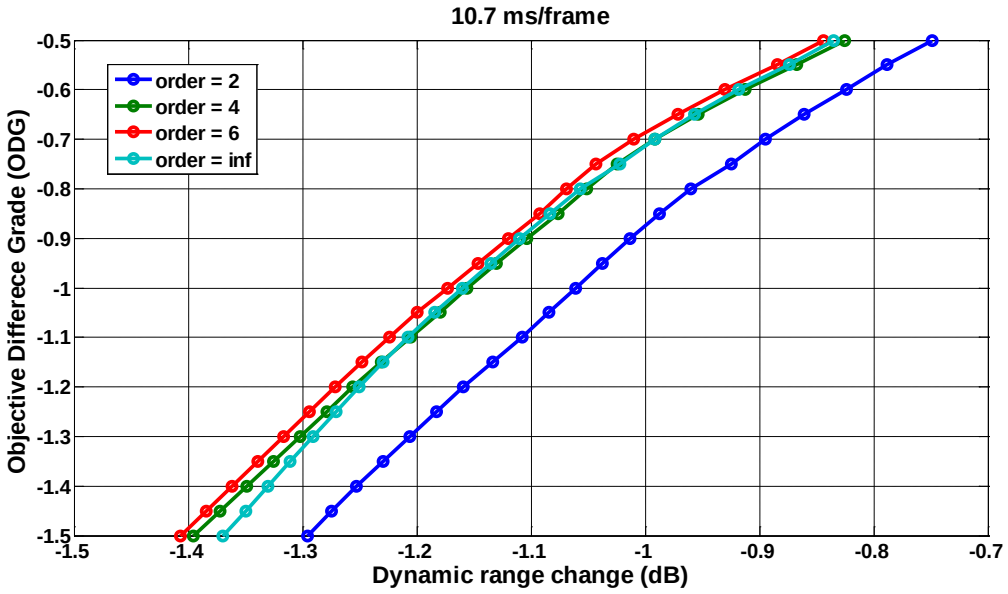


A

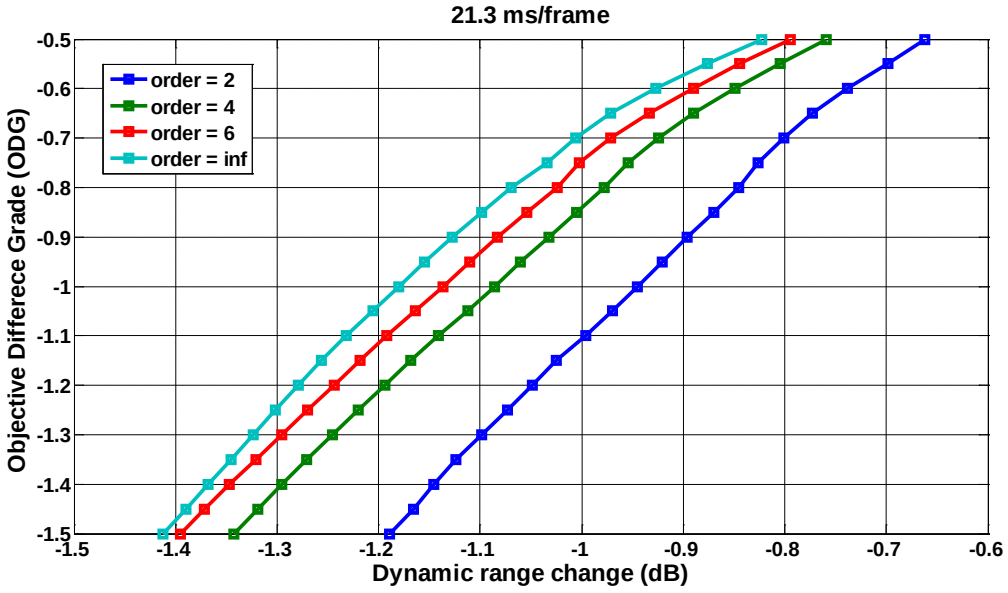


B

Figure 5-4. Operational curves of average dynamic range reduction and objective audio quality for given analysis window sizes A) 2.7 ms/frame B) 5.3 ms/frame C) 10.7 ms/frame D) 21.3 ms/frame E) 42.7 ms/frame F) 85.3 ms/frame

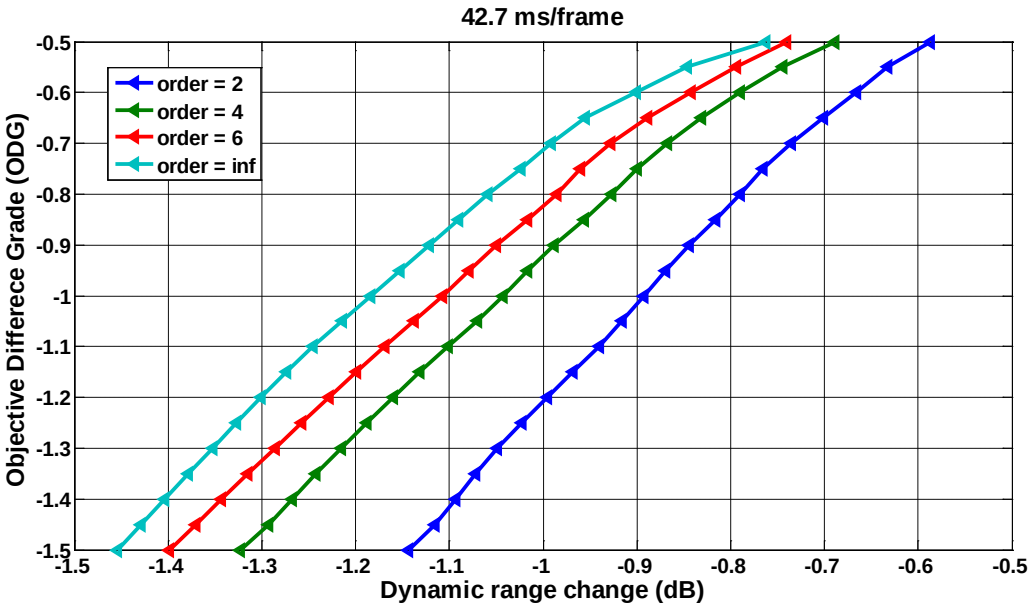


C

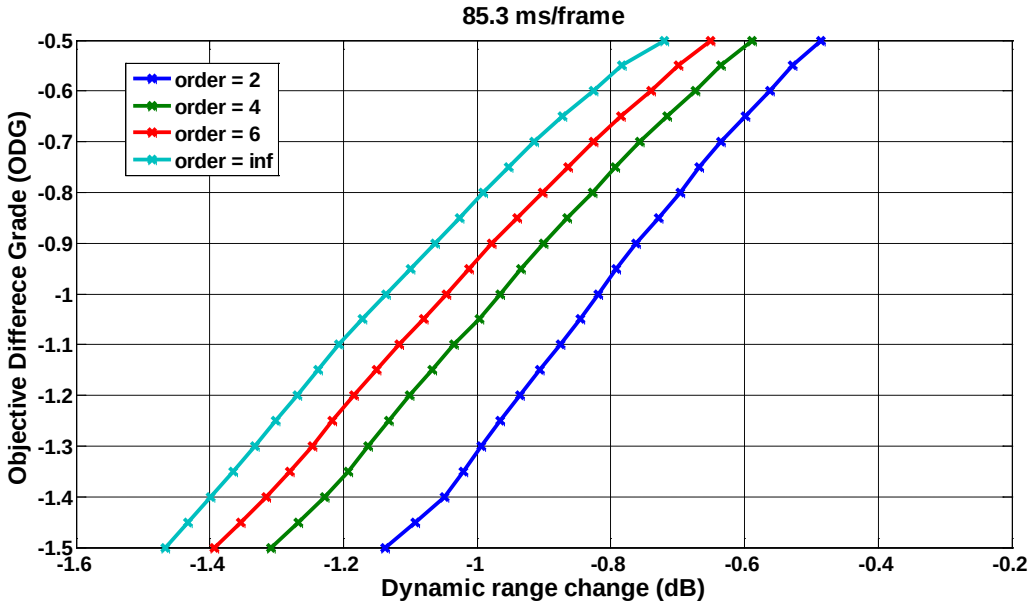


D

Figure 5-4. Continued.



E



F

Figure 5-4. Continued.

Depending on the window size used for DRC segmentation, the optimal order to achieve the best balance of dynamic range reduction and objective audio quality could be different. As shown in the first subplot of Figure 5-4 (A), given a window size of 2.7 ms/frame, the leftmost signal that achieves the best balance between dynamic range

reduction and audio quality is the one processed by the second-order DRC algorithm (represented in blue line). As the window size increases, the optimal order for DRC level estimation also shifts to higher order. With window sizes of 5.3 ms/frame and 10.7 ms/frame, audio signals processed by the fourth order (represented in green line) and sixth order (represented in red line) outperform the signals processed by second-order DRC in compromising dynamic range reduction and audio quality (Figure 5-4 (B) and (C)). When the window size is raised to 21.3 ms/frame and above, the signal processed by infinite-order DRC (signal peak detector) surpasses the signal processed by the fourth and sixth orders and becomes the best in the joint evaluation of dynamic range reduction and audio quality, which are illustrated in the in Figure 5-4 (D)-(E).

Table 5-2 list the impact of window size and DRC order on individual psychoacoustic measures used in PEAQ at a given dynamic range change of -1 dB. These psychoacoustic measures include noise-to-mask ratio, windowed modulation difference, average block distortion, harmonic structure of the error, average modulation difference, distortion loudness, maximum filtered probability of detection, and relatively disturbed frames. All these measure values are normalized to [0,1] for comparison convenience. The corresponding indexes are explained in Table 5-1.

Table 5-1 Psychoacoustic measures used in PEAQ basic version and the corresponding index used in Table 5-2

Index	Psychoacoustic measures
M 1	Noise-to-mask ratio
M 2	Windowed modulation difference
M 3	Average block distortion
M 4	Harmonic structure of the error
M 5	Average modulation difference 1
M 6	Average modulation difference 2
M 7	Distortion loudness
M 8	Maximum filtered probability of detection
M 9	Relatively disturbed frames

Table 5-2 Normalized psychoacoustic measures for a given dynamic range change of -1 dB with different window sizes

Window size = 2.7 ms/frame									
Order	M1	M2	M3	M4	M5	M6	M7	M8	M9
2	0.3687	0.1045	0.6394	0.0194	0.1753	0.0288	0.0404	0.9959	0.3187
4	0.3699	0.1015	0.6327	0.0230	0.1730	0.0287	0.0448	0.9971	0.3073
6	0.3781	0.1040	0.6372	0.0247	0.1784	0.0297	0.0482	0.9960	0.3124
Inf	0.4001	0.1140	0.6559	0.02956	0.1991	0.0335	0.0569	0.9960	0.3327
Best Order	2	4	4	2	4	4	2	2	4
Window size = 5.3 ms/frame									
Order	M1	M2	M3	M4	M5	M6	M7	M8	M9
2	0.3478	0.1355	0.6990	0.0154	0.2231	0.0373	0.0304	0.9913	0.3187
4	0.3267	0.1320	0.6872	0.0175	0.2142	0.0375	0.0314	0.9960	0.2812
6	0.3300	0.1373	0.6854	0.0191	0.2226	0.0398	0.0335	0.9956	0.2853
Inf	0.3556	0.1550	0.7008	0.0213	0.2568	0.0486	0.0402	0.9963	0.3181
Best Order	4	4	6	2	4	2	2	2	4
Window size = 10.7 ms/frame									
Order	M1	M2	M3	M4	M5	M6	M7	M8	M9
2	0.3731	0.1475	0.7329	0.0117	0.2452	0.0402	0.0238	0.9893	0.3330
4	0.3372	0.1383	0.7159	0.0131	0.2306	0.0396	0.0223	0.9682	0.2923
6	0.3315	0.1404	0.7098	0.0133	0.2350	0.0413	0.0228	0.9801	0.2877
Inf	0.3281	0.1453	0.7069	0.0142	0.2514	0.0471	0.0243	0.9731	0.2858
Best Order	Inf	4	Inf	2	4	4	4	4	Inf
Window size = 21.3 ms/frame									
Order	M1	M2	M3	M4	M5	M6	M7	M8	M9
2	0.4368	0.1814	0.7786	0.0086	0.2614	0.0453	0.0232	0.9960	0.3615
4	0.3547	0.1362	0.7228	0.0129	0.2167	0.0360	0.0196	0.9849	0.2835
6	0.3388	0.1317	0.7075	0.0137	0.2132	0.0358	0.0194	0.9797	0.2600
Inf	0.3279	0.1374	0.6993	0.0160	0.2276	0.0407	0.0209	0.9671	0.2416
Best Order	Inf	6	Inf	2	6	6	6	Inf	Inf
Window size = 42.7 ms/frame									
Order	M1	M2	M3	M4	M5	M6	M7	M8	M9
2	0.4713	0.1794	0.7975	0.0073	0.2558	0.0440	0.0217	0.9995	0.3783
4	0.3702	0.1278	0.7399	0.0118	0.2124	0.0329	0.0184	0.9976	0.2885
6	0.3423	0.1229	0.7182	0.0133	0.2046	0.0322	0.0180	0.9917	0.2547
Inf	0.3156	0.1308	0.6912	0.0153	0.2210	0.0377	0.0201	0.9916	0.2365
Best Order	Inf	6	Inf	2	6	6	6	Inf	Inf

Table 5-2 Continued.

Window size = 85.3 ms/frame									
Order	M1	M2	M3	M4	M5	M6	M7	M8	M9
2	0.5125	0.1649	0.8301	0.0046	0.2385	0.0579	0.0403	1.0000	0.4848
4	0.4314	0.1255	0.7988	0.0071	0.1927	0.0302	0.0159	1.0000	0.3766
6	0.3854	0.1089	0.7741	0.0085	0.1789	0.0271	0.0147	1.0000	0.3189
Inf	0.3499	0.1070	0.7385	0.0111	0.1750	0.0275	0.0148	0.9918	0.2657
Best Order	Inf	Inf	Inf	2	Inf	6	6	Inf	Inf

Since the dynamic range change is fixed at -1 dB, the optimal DRC order for an individual psychoacoustic measure with a given window size is determined by the lowest measure value, which are listed in the last row of Table 5-2. As it can be seen in Table 5-2, as the window size increases, the optimal DRC orders for individual psychoacoustic measures generally increases except harmonic structure of the error (M4), which is consistent with the relation between the window size and optimal DRC order for audio quality and dynamic range reduction shown in Figure 5-4. In general, the lower the measure value is, the less impact on the audio quality degradation. The consistency between the individual psychoacoustic measures and the overall audio quality scores is not surprising.

This relation between the window size and the optimal order could be explained by the signal stationarity within the window duration. For short windows of a few milliseconds, the signal stays relatively stationary and signal dynamic range is limited within the each window, hence, the second-order measure is sufficient to characterize the signal in each window. When the window size increases, more transient components are likely to be included within each window and the signal becomes more and more dynamic, as a result, a higher order is needed to measure the non-stationary signal within each window.

To compare the operational curves of dynamic range reduction and objective audio quality across different window sizes, the best operational curves for each given window size in the Figure 5-4 are selected and reproduced in Figure 5-5. It can be seen from Figure 5-5, neither shortest signal segmentation (e.g., 2.7 ms/frame and 5.3 ms/frame) nor the longest signal segmentation (e.g.,85.3 ms/frame) gives the best joint performance of dynamic range reduction and audio quality. The DRC algorithms with really short window sizes are effective in reducing the dynamic range, but tend to degrade the audio quality severely; on the other hand, the DRC algorithms using extremely long window sizes are good at preserving the audio quality, but are limited in compressing the dynamic range. The most desirable window sizes for DRC signal segmentation are from 10.7 ms/frame to 42.7 ms/frame. The optimal DRC orders that correspond to these window sizes are from 6-th order and infinite order. In general, for window sizes from 10 ms/frame to 50ms/frame, audio signals could be very non-stationary. For this window size, higher-order level estimation is necessary to optimize the tradeoff between dynamic range reduction and audio fidelity.

In practice, given the specifications in affordable delay time, buffer sizes, computation capacity, there could be occasions that window sizes are fixed for DRC processing and a suboptimal window size must be used. In this case, the DRC level estimation order should be chosen based on the average signal temporal structure within the given analysis window.

The ITU objective measure of perceived audio quality, PEAQ, is specifically designed for reliably evaluating the perceptual quality of audio codecs and may not be optimal for this DRC application. The psycho-acoustical aspects considered in this

algorithm are targeted at the artifacts introduced by bit-rate reduction in audio encoding and decoding system. When the PEAQ is applied to audio dynamic range controllers, it gives a relevant difference grades for processed signals and original signals, which is valuable for preliminary study but could be biased relative to the subjective evaluation. More subjective experiments are required to show that these objective results are consistent with subjective responses.

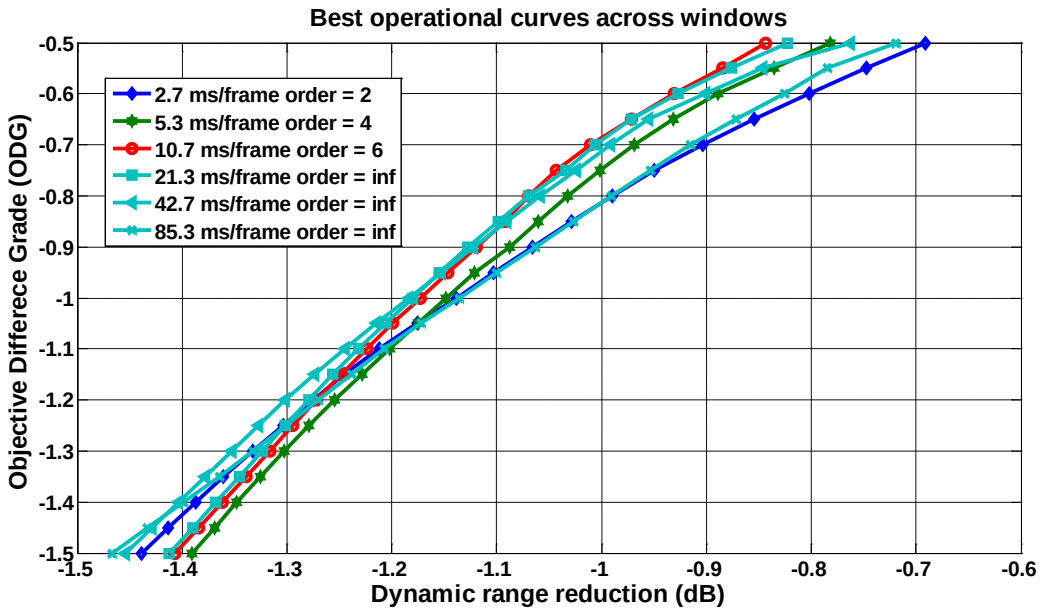


Figure 5-5. Operational curves with the best compromise of dynamic range reduction and audio quality for window sizes from 2.7 ms/frame to 85.3 ms/frame.

5.3 Summary

The conventional second-order DRC is extended to a higher-order DRC algorithm. By examining a variety of audio signals using level estimates with different orders, we find that the optimal level estimate for the DRC algorithm that achieves the best balance of perceptual quality and dynamic range reduction relies on the signal temporal structure within the pre-determined analysis windows, the longer the analysis window is, the more transient the signal is in each window, the higher the order of DRC that is

preferred. Higher-order DRC with moderate window size gives the best performance in general. Further experiments are required to show the objective results presented in this chapter are consistent with the human subjective tests.

CHAPTER 6
CONCLUSION

This aim of this dissertation is to study the impact of temporal structure on human auditory sensitivity of non-stationary sounds and its application to audio dynamic range control. There are three major contributions given as follows.

First, a systematic subjective listening test has been designed and conducted to investigate how temporal structure solely affects human auditory detection of non-stationary sounds independent of spectral differences. The subjective experimental data confirm that the human auditory system is more sensitive to transient signals than steady signals given the same energy, which gives a strong counter-example against the equivalent energy theory for auditory temporal integration.

Since these temporal effects impact the auditory sensitivity to non-stationary sounds, a higher-order spectro-temporal integration model that emphasizes the transient parts relative to the steady parts of an audio signal, is proposed to predict the audibility thresholds of non-stationary sounds. By doing higher-order temporal integration in each critical band and making the decision based on the most audible band, the proposed higher-order integration model outperforms the existing energy-based and loudness-based audibility prediction models on our human subjective data. We expect that this higher-order methodology can be used to develop improved standards for determining hearing risks of transient sounds.

We propose to use audio dynamic range reduction (DRC) to reduce the risk of music induced hearing loss while listening to music. Inspired by the success of the higher-order model in predicting audibility of non-stationary sounds, an audio dynamic range control (DRC) algorithm using higher-order level estimation is proposed for both

hearing protection and audio fidelity. Based on the objective evaluation, the optimal order for the DRC algorithm to achieve the best balance of dynamic range reduction and audio quality is dependent on the signal temporal structure within a given analysis window. In general, for moderate sized temporal windows, the higher-order DRC algorithms provide the optimal tradeoff between dynamic range reduction and signal fidelity. Further subjective experiment is needed to confirm the objective results are consistent with the subjective evaluations.

APPENDIX
DERIVATION FOR KEY-FOWLE-HAGGARTY PHASE SOLUTION FOR WAVEFORM MANIPULATION

(Key et al 1961, Fowle 1964, Quatieri and McAulay1991)

Let x(t) be the given signal and $\hat{x}(t)$ be its Hilbert transform . Then the analytic signal representation r(t) is given by

$$r(t) = x(t) + j\hat{x}(t) \tag{A-1}$$

Let $|r(t)|$ be the envelope and $\varphi(t)$ is the phase of r(t), the analytic signal be written as

$$r(t) = |r(t)| \exp(j\varphi(t)) \tag{A 2}$$

which has a Fourier transform

$$M(\omega) = |M(\omega)| \exp(j\psi(\omega)) \tag{A 3}$$

The signal design problem can be stated as follows: given a time-domain envelope $|r(t)|$ and a frequency-domain spectral magnitude $|M(\omega)|$, find the phase $\varphi(t)$ in time and $\psi(\omega)$ in frequency such that the following Fourier transform relation is satisfied:

$$|r(t)| \exp(j\varphi(t)) = F^{-1}\{|M(\omega)| \exp(j\psi(\omega))\} \tag{A 4}$$

where "F^{-1}"denotes inverse Fourier transform.

$$\begin{aligned} r(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |M(\omega)| \exp(j\psi(\omega)) \cdot \exp(j\omega t) d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |M(\omega)| \exp(j(\omega t + \psi(\omega))) d\omega \end{aligned} \tag{A 5}$$

According to the principle of stationary phase, the integral of a rapidly oscillating function has little value except in regions where the phase is "stationary" or where the derivative of the phase is zero. In (A 5) we have a stationary point where

$$\frac{d}{d\omega}(\omega t + \psi(\omega)) = 0 \qquad \text{(A 6)}$$

Let us represent the value of ω which satisfies (A 6) by λ. Then we have

$$\psi'(\lambda) = -t \qquad \text{(A 7)}$$

Under the assumption that for each value of time t, there is only one stationary point, (A 5) then becomes

$$r(t) \cong \frac{1}{2\pi}\int_{\lambda-\varepsilon}^{\lambda+\varepsilon} |M(\omega)| \exp(j(\omega t + \psi(\omega)))d\omega \qquad \text{(A 8)}$$

Following the method of stationary phase, the phase function in (A 5) is expanded in a 3-term Taylor's series about the stationary point $\omega = \lambda$ and

$$\omega t + \psi(\omega) \cong [\lambda t + \psi(\lambda)] + [t + \psi'(\lambda)](\omega - \lambda) + \frac{\psi''(\lambda)}{2}(\omega - \lambda)^2 \qquad \text{(A 9)}$$

The second term is zero based on (A 7). Then r(t) is given by

$$r(t) \cong \frac{1}{2\pi}|M(\lambda)| \exp\{j[\lambda t + \psi(\lambda)]\} \cdot \int_{\lambda-\varepsilon}^{\lambda+\varepsilon} \exp[j\frac{\psi''(\lambda)}{2}(\omega - \lambda)^2]d\omega \qquad \text{(A 10)}$$

Setting $x = \omega - \lambda$ and $\sigma^2 = \frac{-1}{j\psi''(\lambda)}$, (A 10) becomes

$$r(t) \cong \frac{1}{2\pi}|M(\lambda)| \exp\{j[\lambda t + \psi(\lambda)]\} \cdot \int_{-\varepsilon}^{\varepsilon} \exp[\frac{-x^2}{2\sigma^2}]dx \qquad \text{(A 11)}$$

The statement that the phase is dispersive at $\omega = \lambda$ infers that $\psi''(\lambda)$ is large and σ^2 is therefore small. We assume σ^2 is small enough to cause the entire area under the Gaussian function in (A 11) to be obtained within the limits $\pm\varepsilon$ even though ε is small. Hence, we have

$$r(t) \cong \frac{1}{2\pi}|M(\lambda)| \exp\{j[\lambda t + \psi(\lambda)]\} \cdot \int_{-\infty}^{+\infty} \exp[j\frac{\psi''(\lambda)}{2}(\omega - \lambda)^2]d\omega \qquad \text{(A 12)}$$

The integral may be evaluated as follows,

$$\int_{-\infty}^{+\infty} \exp(\frac{-x^2}{\sigma^2})dx = \sqrt{2\pi}\sigma \qquad \text{(A 13)}$$

where

$$\sigma = \frac{\exp(\pm j\frac{\pi}{4})}{\sqrt{|\psi''(\lambda)|}} \qquad \text{(A 14)}$$

We take + associated with $\psi''(\lambda) > 0$, - with $\psi''(\lambda) > 0$.

Thus we obtain

$$r(t) = \frac{|M(\lambda)|}{\sqrt{\psi''(\omega)}} \exp\left[j(\lambda t + \psi(\lambda) \pm \frac{\pi}{4}) \right] \qquad \text{(A 15)}$$

The modulus |r(t)| is

$$|r(t)| = \frac{|M(\lambda)|}{\sqrt{\psi''(\lambda)}} \qquad \text{(A 16)}$$

To make |r(t)| constant, inspection of (A 16) shows that we must set

$$|\psi''(\lambda)| = c|M(\lambda)|^2 \qquad \text{(A 17)}$$

We differentiate (A 7) with respect to λ to obtain

$$\psi''(\lambda) = -\frac{dt}{d\lambda} \qquad \text{(A 18)}$$

Then we square both sides of (A 16) and use (A 18) to get

$$|r(t)|^2 dt = |M(\lambda)|^2 d\lambda \qquad \text{(A 19)}$$

We integrate (A 19) and have

$$\int_{-\infty}^{t} |r(t)|^2 dt = \int_{-\infty}^{\lambda} |M(\omega)|^2 d\omega \qquad \text{(A 20)}$$

Let $|r(t)| = A$ and A is constant, we have

$$\int_{-\infty}^{t} A^2 dt = A^2 t = \int_{-\infty}^{\lambda} |M(\omega)|^2 d\omega \qquad \text{(A 21)}$$

It follows

$$t = (\int_{-\infty}^{\lambda} |M(\omega)|^2 d\omega) / A^2 \qquad \text{(A 22)}$$

By substituting (A 22) in to (A 7), we have

$$\psi'(\lambda) = (-\int_{-\infty}^{\lambda} |M(\omega)|^2 d\omega) / A^2 \qquad \text{(A 23)}$$

Integrate both side of (A 17), we have

$$\psi'(\lambda) = c \int_{-\infty}^{\lambda} |M(\omega)|^2 d\omega \qquad \text{(A 24)}$$

By comparing (A 23) and (A 24), we have

$$c = \frac{-1}{A^2} \qquad \text{(A 25)}$$

Parseval's theory requires that

$$\int_{0}^{T} |r(t)|^2 dt = \int_{-\pi}^{\pi} |M(\omega)|^2 d\omega \qquad \text{(A 26)}$$

Because for analytic signal r(t), we have $M(\omega) = 0(\omega < 0)$, (A 26) can be re-written as

$$\int_{0}^{T} |r(t)|^2 dt = \int_{0}^{\pi} |M(\omega)|^2 d\omega \qquad \text{(A 27)}$$

By substituting $|r(t)| = A$ in to (A 27), we have

$$\int_0^T A^2 dt = \int_0^{\pi} |M(\omega)|^2 d\omega \qquad \text{(A 28)}$$

$$\Rightarrow A^2 T = \int_0^{\pi} |M(\omega)|^2 d\omega \qquad \text{(A 29)}$$

$$\Rightarrow A^2 = (\int_0^{\pi} |M(\omega)|^2 d\omega)/T \qquad \text{(A 30)}$$

Combining (A 30) and (A 25), we have

$$c = -\frac{T}{\int_0^{\pi} |M(\omega)|^2 d\omega} \qquad \text{(A 31)}$$

By substituting (A 31) into (A 17), we have

$$|\psi''(\lambda)| = \frac{-T|M(\lambda)|^2}{\int_0^{\pi} |M(\omega)|^2 d\omega} \qquad \text{(A 32)}$$

$$\Rightarrow \psi(\lambda) = \frac{-T\int_0^{\lambda}\int_0^{\beta} |M(\alpha)|^2 d\alpha d\beta}{\int_0^{\pi} |M(\omega)|^2 d\omega} = -T\int_0^{\lambda}\int_0^{\beta} \hat{M}(\alpha)^2 d\alpha d\beta \qquad \text{(A 33)}$$

where $\hat{M}^2(\alpha) = \frac{|M(\alpha)|^2}{\int_0^{\pi} |M(\omega)|^2 d\omega}$

LIST OF REFERENCES

American National Standards Institute (1996). ANSI S3.44-1996 American National Standard Determination of Occupational Noise Exposure and Estimation of Noise-induced Hearing Impairment.

Buus, S., E. Schorer, et al. (1986). "Decision rules in detection of simple and complex tones." The Journal of the Acoustical Society of America 80(6): 1646-1657.

Chasin, M. (2003). "Music and hearing aids." Hearing Journal 56(7): pp 36,38,40-41.

Chasin, M. (2006). "Hearing Aids for Musicians." Hearing Review.

Eddins, D. A. and D. M. Green (1995). Temporal integration and temporal resolution. Hearing., San Diego, CA, US: Academic Press: 207-242.

Erdreich, J. (1985). "Distribution based definition of impulse noise." The Journal of the Acoustical Society of America 77(S1): S19-S19.

European Broadcasting Union (2001). Recommendation ITU-R BS.1387-1: Method for objective measurements of perceived audio quality.

European Broadcasting Union (2008). Sound Quality Assessment Material recordings for subjective tests.

Fastl, H. and E. Zwicker (2006). Psychoacoustics: Facts and Models, Springer-Verlag New York, Inc.

Fletcher, H. (1940). "Auditory Patterns." Reviews of Modern Physics 12(Copyright (C) 2010 The American Physical Society): 47.

Fletcher, H. and W. A. Munson (1933). "Loudness of a Complex Tone, Its Definition, Measurement and Calculation." The Journal of the Acoustical Society of America 5(1): 65-65.

Fligor, B. J. (2009). "Risk for Noise-Induced Hearing Loss From Use of Portable Media Players: A Summary of Evidence Through 2008." Perspectives on Audiology 5(1): 10-20.

Fowle, E. (1964). "The design of FM pulse compression signals." Information Theory, IEEE Transactions on 10(1): 61-67.

Glasberg, B. R. and B. C. J. Moore (1990), “Derivation of auditory filter shapes from notched-noise data.” Hearing Research, 47: 103-138.

Glasberg, B. R. and B. C. J. Moore (2005). "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds." Journal of Audio Engineering Society, 53: 906-918.

Green, D. M. (1958). "Detection of Multiple Component Signals in Noise." The Journal of the Acoustical Society of America 30(10): 904-911.

Green, D. M. and R. D. Luce (1975). "Parallel psychometric functions from a set of independent detectors." Psychological Review 82(6): 483-486.

Green, D. M. and J. A. Swets (1974). Signal detection theory and psychophysics, Oxford, England: Robert E. Krieger.

Hamernik, R. P. and W. Qiu (2001). "Energy-independent factors influencing noise-induced hearing loss in the chinchilla model." The Journal of the Acoustical Society of America 110(6): 3163-3168.

Hamernik, R. P., W. Qiu, et al. (2003). "The effects of the amplitude distribution of equal energy exposures on noise-induced hearing loss: the kurtosis metric." J Acoust Soc Am 114(1): 386-95.

Henderson, D. and R. P. Hamernik (1995). "Biologic bases of noise-induced hearing loss." Occup Med 10(3): 513-34.

Hsueh, K. D. and R. P. Hamernik (1990). "A generalized approach to random noise synthesis: Theory and computer simulation." The Journal of the Acoustical Society of America 87(3): 1207-1217.

International Organization for Standardization (1999). ISO 1999:1990 Acoustics - Determination of occupational noise exposure and estimation of noise-induced hearing impairment.

Kabal, P. (2002). An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality.

Kay, S. M. (1998). Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory Prentice Hall PTR.

Levitt, H. (1971). "Transformed Up-Down Methods in Psychoacoustics." The Journal of the Acoustical Society of America 49(2B): 467-477.

Malmierca, M. and D. Irvine, Eds. (2005). Auditory spectral processing, Academic Press.

Melnick, W. (1991). "Human temporary threshold shift (TTS) and damage risk." The Journal of the Acoustical Society of America 90(1): 147-154.

Mills, J. H., R. M. Gilbert, et al. (1979). "Temporary threshold shifts in humans exposed to octave bands of noise for 16 to 24 hours." The Journal of the Acoustical Society of America 65(5): 1238-1248.

Moore, B. C. J., Ed. (1995). Hearing (Handbook of Perception and Cognition), Academic Press.

Moore, B. C. J. (2003). An introduction to the psychology of hearing (5th ed.), San Diego, CA, US: Academic Press.

National Institute for Occupational Safety and Health (1998). NIOSH 98-126 Criteria for a Recommended Standard: Occupational Noise Exposure.

Occupational Safety & Health Administration (1986). OSHA 1910.95 Occupational noise exposure.

Price, G. R. and S. Wansack (1985). "A test of predicted maximum susceptibility to impulse noise." The Journal of the Acoustical Society of America 77(S1): S82-S82.

Quaranta, A., P. Portalatini, et al. (1998). "Temporary and permanent threshold shift: an overview." Scand Audiol Suppl 48: 75-86.

Quatieri, T. F. and R. J. McAulay (1991). "Peak-to-RMS reduction of speech based on a sinusoidal model." Signal Processing, IEEE Transactions on 39(2): 273-288.

Schafer, T. H. and R. S. Gales (1949). "Auditory Masking of Multiple Tones by Random Noise." The Journal of the Acoustical Society of America 21(4): 392-397.

Schmidt, J. C. and J. C. Rutledge (1996). Multichannel dynamic range compression for music signals. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on.

Schneider, T. and R. Brennan (1997). A multichannel compression strategy for a digital hearing aid. Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on.

Stevens, S. S. (1955). "The Measurement of Loudness." The Journal of the Acoustical Society of America 27(5): 815-829.

Strasser, H., H. Irle, et al. (1999). "Physiological cost of energy-equivalent exposures to white noise, industrial noise, heavy metal music, and classical music." Noise Control Engineering Journal 47(5): 187-197.

van den Brink, W. A. and T. Houtgast (1990). "Efficient across-frequency integration in short-signal detection." Journal of the Acoustical Society of America 87(1): 284-291.

van den Brink, W. A. C. and T. Houtgast (1990). "Spectro-temporal integration in signal detection." The Journal of the Acoustical Society of America 88(4): 1703-1711.

Viemeister, N. F. and G. H. Wakefield (1991). "Temporal integration and multiple looks." The Journal of the Acoustical Society of America 90(2): 858-865.

Von Békésy, G. (1960). Experiments in hearing, Oxford, England: Mcgraw Hill.

Ward, W. D., E. M. Cushing, et al. (1976). "Effective quiet and moderate TTS: Implications for noise exposure standards." The Journal of the Acoustical Society of America 59(1): 160-165.

Yang, Q. and J. G. Harris (2009). An Audibility Model for non-stationary sounds. The American Speech-Language-Hearing Association (ASHA) Annual Convention 2009.

Yang, Q. and J. G. Harris (2010). A higher-order spectro-temporal integration model for predicting signal audibility. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.

Zwicker, E., G. Flottorp, et al. (1957). "Critical band width in loudness summation." Journal of the Acoustical Society of America 29: 548-557.

BIOGRAPHICAL SKETCH

Qing Yang was born in Guangzhou, the largest city in South China, and grew up in Shenzhen, an immediate neighbor to Hongkong.

Qing received her bachelor's and master's degrees in electrical engineering from South China University of Technology (SCUT) in 2002 and 2005, respectively. She started her doctoral program at University of Florida in 2005 and joined the Computational Neuro-Engineering Laboratory (CNEL) in 2006. Since 2006, she has been working as a research assistant at CNEL under the guidance of Dr. John G. Harris. Her research interests include auditory perception, speech/audio signal processing, loudness measurement and control, dynamic range control, perceptual evaluation of sound quality. She worked as a software engineering intern at the iDen advanced technology group, Motorola Mobile Devices, Plantation, Florida in the fall semester of 2006 to optimize and implement a speech enhancement algorithm for Motorola mobile devices.