Incorporating Auditory Models in Speech/Audio Applications

by

Harish Krishnamoorthi

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2011 by the
Graduate Supervisory Committee:

Andreas Spanias, Chair
Antonia Papandreou-Suppappola
Cihan Tepedelenlioglu
Konstantinos Tsakalis

ARIZONA STATE UNIVERSITY

May 2011

ABSTRACT

Following the success in incorporating perceptual models in audio coding algorithms, their application in other speech/audio processing systems is expanding. In general, all perceptual speech/audio processing algorithms involve minimization of an objective function that directly/indirectly incorporates properties of human perception. This dissertation primarily investigates the problems associated with directly embedding an auditory model in the objective function formulation and proposes possible solutions to overcome high complexity issues for use in real-time speech/audio algorithms.

Specific problems addressed in this dissertation include: 1) the development of approximate but computationally efficient auditory model implementations that are consistent with the principles of psychoacoustics, 2) the development of a mapping scheme that allows synthesizing a time/frequency domain representation from its equivalent auditory model output.

The first problem is aimed at addressing the high computational complexity involved in solving perceptual objective functions that require repeated application of auditory model for evaluation of different candidate solutions. In this dissertation, a frequency pruning and a detector pruning algorithm is developed that efficiently implements the various auditory model stages. The performance of the pruned model is compared to that of the original auditory model for different types of test signals in the SQAM database. Experimental results indicate only a 4-7 % relative error in loudness while attaining up to 80-90 % reduction in computational complexity. Similarly, a hybrid algorithm is developed specifically for use with sinusoidal signals and employs the proposed auditory pattern combining technique together with a look-up table to store representative auditory patterns.

The second problem obtains an estimate of the auditory representation that minimizes a perceptual objective function and transforms the auditory pattern back to its equivalent time/frequency representation. This avoids the repeated application of auditory model stages to test different candidate time/frequency vectors in minimizing perceptual objective functions. In this dissertation, a constrained mapping scheme is developed by linearizing certain auditory model stages that ensures obtaining a time/frequency mapping corresponding to the estimated auditory representation. This paradigm was successfully incorporated in a perceptual speech enhancement algorithm and a sinusoidal component selection task.

To my parents.

ACKNOWLEDGEMENTS

I came to the US in Fall 2005 with the intent of pursuing a Masters and possibly a PhD degree. Half way through my Masters degree, I decided to continue with my PhD. The days turned into months, the months into semesters and soon all I could remember was the years that had gone by. Most of my friends had graduated and found jobs, and yet, the finish line for me looked further and further apart the nearer and nearer I thought I came to it. For some, who had seen the ”greener” side of life, it was difficult to understand the rationale behind going through such long periods of uncertainty that required tremendous patience and self-sustained motivation. In a way, this entire process can be compared to the journey that a person would undertake when hiking in an uncharted territory without a map. In the end, it was all one worthwhile adventure that was not only technically challenging but also tested one’s determination, patience and one’s ability to handle stressful situations. And I can now confidently say that this has been such a learning experience - thanks to the so many wonderful people who had been instrumental in keeping me going during good and difficult times.

First and foremost, I would like to thank my advisor Dr. Andreas Spanias for having faith in my abilities and providing me with the necessary resources to pursue the PhD degree. Also, many thanks to Dr. Spanias for providing me with a wonderful opportunity of being as a TA for the Digital Signal processing course for 5 good years. It was indeed a blessing in disguise for it not only strengthened my communication abilities and DSP fundamentals but repeatedly allowed me to become part of different students learning experiences. This was indeed a very satisfying and rewarding experience in itself and provided me with good insight into teaching techniques. Thanks to Dr. Spanias for funding the numerous IEEE ICASSP conferences, the SENSIP workshops and the AES convention trips. Each of these trips were memorable and provided me with a rich experience through the many interactions with other researchers, the tutorial sessions, not to forget the dinner/lunch get-together with former students. Many thanks to Dr. Spanias for helping me with crucial financial and administrative aspects that enabled me to spend an excellent summer at Philips Research Laboratories in Eindhoven, The Netherlands.

Thanks to Dr. Antonia Papandreou-Suppappola for being as a mentor at various stages of my PhD. I really consider myself blessed to have known you and for all the support and guidance received during my stay here. I would like to thank Dr. Cihan Tepedelenlioglu and Dr. Konstantinos Tsakalis for agreeing to be on my committee and for their valuable feedback. I would also like to thank Dr. Albertus C. den Brinker for honing my critical thinking and writing skills and for providing me an opportunity to be a part of the DSP group at Philips Research Laboratories amidst the several administrative hurdles that we faced.

I would like to thank Venkatraman Atti for all the encouraging words, initial guidance and help with the online TA course materials. Thanks for providing me with the code for excitation similarity weighting algorithm developed by Dr. Ted Painter. It served as the seed for my dissertation and eventually helped me identify research problems and provided a direction for my PhD. Thanks to Visar Berisha for guiding me during the initial phases of my PhD. We have had the most productive meetings and embarked on several projects. I still cannot forget our initiative to start a venture and apply for seed funding. Thanks to Venktesh Kalyanaraman (Venky) for helping me with every little thing in the first semester and for the small part that I could be in AID-Tempe.

I would like to thank Viju for being the “partner in crime” through the years. Many thanks to Mahesh, Jayaraman and Karthikeyan for all the research discussions, the paper reviews and not to mention the daily doses of enthusiasm that helped me stay sane through the many stressful times. Thanks to

Homin Kwon, Shibani Misra, Raghavendra Bhatt, Prasanna Sattigeri, Mohit Shah, Bharatan Konnanath, Tushar Gupta, Peter Knee, Robert Santucci and Henry. Thanks to my special friends Vinod, Sanjay, Viswesh, Dilip, Supraja, Shriya, Thripthi, Anusha, Shanta, Prashanth, Prathap, Hari Krishnan, Dinesh and Hariharan for all the memorable moments.

I would like to thank my parents for being patient and supporting me throughout the course of my PhD. All the above experiences and accomplishments wouldn't have been possible without that single effort from them. Thanks also to my sister for putting up with me.

Doing a PhD requires the right mix of patience, dedication and intelligence and I believe that each one of my friends, colleagues and professors have instilled in me these necessary ingredients at various times that has helped me to successfully complete my degree.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Chapter 1

Introduction

1.1 Perceptual Models: An Introduction

In the context of speech/audio processing, a perceptual model is one that takes into account the properties of human auditory system. The perceptual models are developed based on the results of numerous psychoacoustic experiments that study the relationship between the acoustic stimuli and the hearing sensations. A number of auditory models have been proposed in the literature [3–12]. These range from simple frequency/gain transformations to more elaborate filter-bank based perceptual models. The elaborate models characterize several aspects of the human auditory system such as their non-uniform frequency sensitivity, the notion of critical bands, the masking phenomenon, the response of the basilar membrane and neural receptors in the cochlea, the phenomenon of loudness among several others.

For example, the frequency weighting curves such as the A, B or C- weighting functions as shown in Figure 1.1 are derived from the equal-loudness contours model the non-uniform sensitivity of human auditory system [12]. These simple models do not account for masking and therefore perform poorly for transient and broadband sounds.

More elaborate models attempt to model the cochlea as a bank of auditory filters with bandwidths corresponding to critical bandwidths [3, 4, 13, 14]. These auditory filters are realized either in the time domain or in the frequency domain and depending on the mode of implementation, we have two broad classes of auditory models: i) time-domain models and ii) frequency-domain models. Some time-domain auditory filter implementations include the Gammatone filters [15], Gammachirp filters [10], dual-resonance nonlinear filter (DRNL) [11] among many others. Similarly, the rectangular filters [1] and rounded exponential (roex) fil-

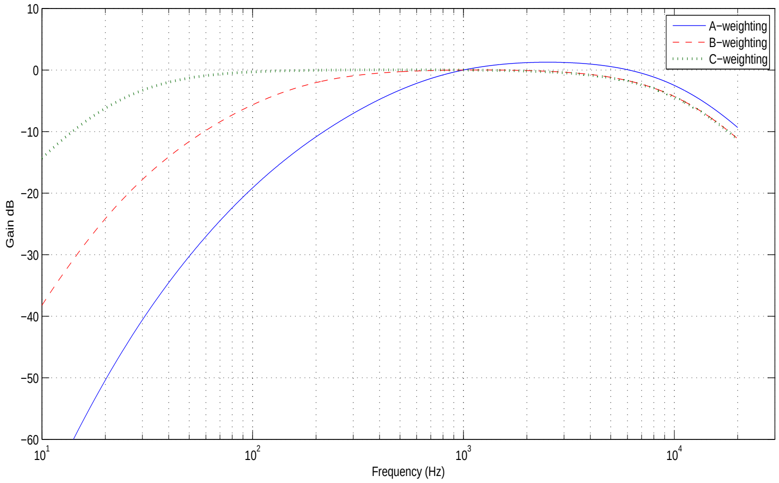

Figure 1.1: Different frequency weighting functions.

ters [16] represent the frequency-domain auditory filter implementations. In general, the frequency-domain auditory filter implementations are computationally less expensive than their time-domain counterparts. Some of these perceptual models [4,17] also estimate the instantaneous, short-term and long-term loudness associated with time-varying signals.

Another important class of perceptual models include those that generate a frequency dependent masked thresholds curve that characterizes the masking phenomenon but do not explicitly model the different stages of an auditory system. A hierarchical organization of perceptual models is shown in Figure 1.2. The introduction of several simple to more advanced auditory models have resulted in their widespread use in several speech/audio processing applications.

In the next section, an overview of how these perceptual models are incorporated in different speech/audio applications is described. This will highlight the current trends and limitations associated with including perceptual models in speech/audio applications.

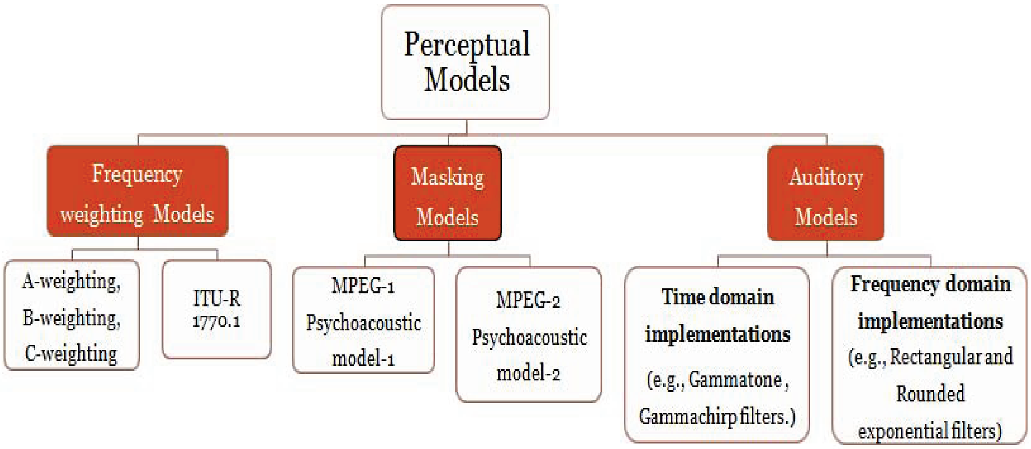
Perceptual Models
Frequency weighting Models
Masking Models
Auditory Models
A-weighting, B-weighting, C-weighting
ITU-R 1770.1
MPEG-1 Psychoacoustic model-1
MPEG-2 Psychoacoustic model-2
Time domain implementations (e.g., Gammatone, Gammachirp filters.)
Frequency domain implementations (e.g., Rectangular and Rounded exponential filters)

Figure 1.2: An overview of different types of perceptual models.

1.2 Current State-of-the-art Perceptual Algorithms: A Review

Perceptual models have been most widely used in audio coding algorithms. An excellent review of perceptual audio coding algorithms can be found in [2,18]. In perceptual audio coding algorithms, the objective is to quantize (or encode) the underlying signal with as few bits as possible while retaining a "transparent signal quality", i.e., the output audio should be indistinguishable from the original input. This is accomplished by making use of masking models (perceptual models) that calculates a global masking threshold for short segments of the input audio. The masked thresholds represents a signal dependent threshold of audibility curve wherein signal components falling below this threshold are rendered inaudible. This property of the human auditory system to mask components below a certain threshold is exploited by several state-of-the-art perceptual audio coders, such as ISO/IEC MPEG-1 Layer-3 (MP3) [19], the Dolby AC-2 and AC-3 standards [20], the MPEG-2 AAC [21] etc., to appropriately "hide" the quantization noise below the masked thresholds.

Following the success of perceptual models in several audio coding algorithms, their use in other audio/speech processing applications have been on the increase. Since the objective behind employing any perceptual model is to mimic the functioning of the human auditory system, they have been used to develop objective and hybrid measures that predict subjective quality [2]. For example, the PESQ (Perceptual Evaluation of Speech Quality) [22], the POM (Perceptual objective Measure) [23], the PERCEVAL [24] and the MBSD (Modified Bark Spectral Distortion) [25] represent some objective measures that make use of perceptual models to predict the perceived quality associated with a processed or a coded signal.

Another application where perceptual models have been widely employed is speech enhancement. Here, the objective is to improve the quality and intelligibility of a noise corrupted speech signal. The first step generally in these algorithms is to reduce/remove the noise from the degraded speech signal. In this context, several perceptual strategies have been employed. For example, perceptual weighting filters, derived from the LP (linear prediction) analysis of speech segments, are used to shape the residual noise so as to "hide" the noise in high energy spectral regions (i.e., formant peaks) and aggressively suppress noise near spectral valleys [26]. Similarly, in [27], masking thresholds are employed to adapt the parameters of the spectral subtraction based speech enhancement algorithm [27]. The parameters that usually control the tradeoff between the amount of speech distortion and residual noise are now adapted based on human auditory perception instead of energy based metrics. Similar perceptual strategies have also been incorporated in statistical model-based techniques [28] and subspace techniques [29, 30] for speech enhancement. Although most of these algorithms employ a tractable error criterion, the incorporation of perceptual constraints is usually done heuristically.

Other applications include sinusoidal analysis/synthesis algorithms that use masking models to iteratively extract sinusoids using a matching pursuit algorithm [31,32]. Also, hearing aid systems use loudness models to compensate for perception loss [13].

Finally, a few algorithms also make use of more sophisticated auditory models in order to directly optimize a perceptual distortion function. For example, sinusoidal component selection algorithms based on minimizing excitation pattern distortion [33] and loudness pattern distortions [34] select sinusoids by incorporating an auditory model (to generate the excitation/loudness patterns) during the error minimization process. Similarly, the bandwidth extension algorithm proposed in [35] makes use of auditory patterns to determine the perceptual importance of the different high-band sub-bands in order to reduce the amount of side-information bits. In [36], a perceptual linear prediction algorithm makes use of auditory patterns to estimate perceptual pole frequencies and thereby construct a perceptual all-pole LP filter. Recently, a fast algorithm to generate auditory patterns was proposed in [37,38] to reduce the computational complexity associated with generating auditory patterns in the above applications.

Limitations of Perceptual techniques

All of the above perceptual schemes can be roughly classified into the following two classes:

1. The first class of algorithms minimize a cost function $C(\mathbf{x}, \mathbf{x}')$ subject to a set of constraints, where $C(\mathbf{x}, \mathbf{x}')$ represents some cost function that measures the distance between the two time/frequency domain vectors $\mathbf{x}$ and $\mathbf{x}'$ respectively. This is illustrated in the left side of Figure 1.3 where the minimization process looks for solution vectors in the time/frequency domain. The objective function $C(\mathbf{x}, \mathbf{x}')$ does not directly embed a perceptual model in its formulation, instead additional constraints are placed to include properties of human perception. That is, suitable thresholds/weights are extracted from the output of the perceptual models to constrain the solution

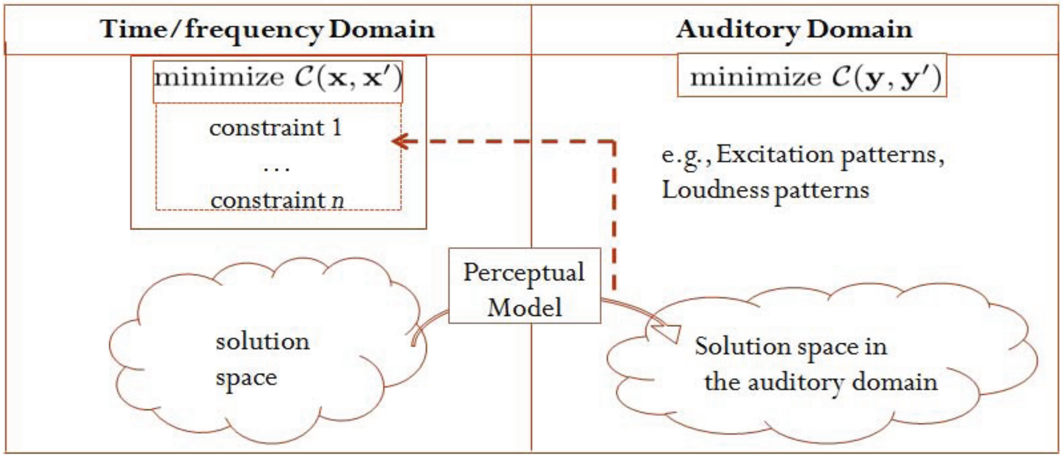
Time/frequency Domain
Auditory Domain
minimize $\mathcal{C}(\mathbf{x}, \mathbf{x}')$
constraint 1
...
constraint n
minimize $\mathcal{C}(\mathbf{y}, \mathbf{y}')$
e.g., Excitation patterns,
Loudness patterns
Perceptual Model
solution space
Solution space in the auditory domain

Figure 1.3: Illustrating the relationship between cost functions and their solution space in time/frequency domain and auditory domain.

obtained from the non-perceptual objective function as shown in Figure 1.3. Some examples of perceptual algorithms presented in Section 1.2 that belong to this class include the speech enhancement algorithms [26, 29, 30], the sinusoidal analysis/synthesis algorithms [31, 32] and the perceptual audio coding algorithms [19–21].

2. The second class of algorithms rely on minimizing a perceptual distortion function $C(\mathbf{y}, \mathbf{y}')$ rather than minimizing a time/frequency domain signal distortion $C(\mathbf{x}, \mathbf{x}')$. This involves directly embedding an auditory model in the objective function and carrying out the minimization in the auditory domain as illustrated in Figure 1.3. For example, instead of minimizing a mean square error between $\mathbf{x}$ and $\mathbf{x}'$ in the time/frequency domain, we can minimize the mean square error between their auditory representations $\mathbf{y}$ and $\mathbf{y}'$. Examples that belong to this class of algorithms include the bandwidth extension algorithm [35], the perceptual linear prediction algorithm [36] and the sinusoidal component selection algorithm [33, 34].

The first class of algorithms rely on indirect approaches in incorporating perceptual considerations. However, a number of limitations make this scheme not so attractive for incorporating perceptual characteristics:

a) Firstly, they do not attempt to minimize a perceptual distortion function, therefore there is scope for additional perceptual gains to be achieved with more direct approaches.

b) Secondly, in problems where the signal is corrupted by noise, the output thresholds/weights obtained from the perceptual models as shown in Figure 1.3 are in turn noisy and therefore become unreliable when included as constraints.

On the other hand, the second class of algorithms directly minimize a perceptual distortion function. However, they present difficulties in solving for their optimal solution. This is primarily due to the fact that the perceptual models contain nonlinear transformations during the various model stages. For example, due to the phenomenon of masking, two or more different signals can result in identical auditory perceptions. To illustrate this phenomenon, we can consider the following two test cases: one with only the masker signal and the other with the masker and a masked signal. By assumption, both signals are associated with the same perception since the maskee is inaudible. This many-to-one mapping presents additional problems as it is difficult to decide on a particular time/frequency domain vector that corresponds to the optimal auditory domain solution.

1.3 Statement of the Problem

In this dissertation, we address the problems associated with directly embedding an auditory model in a perceptual objective function. The minimization of perceptual objective functions is more consistent with processing signals according to human perception than that followed by minimizing an equivalent time/frequency domain error criterion. In general, the optimal solution can be obtained by following one of the two paradigms:

1. The first paradigm involves repeatedly employing the auditory model over the entire search space of candidate solutions $\mathbf{x}$ to obtain an optimal solution $\mathbf{x}^{\mathrm{opt}}$ that minimizes $\mathcal{C}(\mathbf{y}, \mathbf{y}')$. The computational complexity associated with this approach is combinatorial in nature and therefore very high for practical purposes. Alternatively, sub-optimal approaches are resorted to wherein $\mathcal{C}(\mathbf{y}, \mathbf{y}')$ is minimized using iterative optimization techniques similar to that followed in a matching pursuits approach. Although the iterative approach is associated with a lesser computational complexity than that of the exhaustive search procedure, it still requires repeated application of the auditory model stages in each of its iterations. The resulting computational complexity is still high and unsuitable for several real-time applications, most notably Internet streaming and telecommunication applications.

2. The second paradigm involves transforming time/frequency domain vectors $\mathbf{x}$ into their equivalent auditory representations $\mathbf{y}$ (by following the auditory model stages) and subsequently optimizing in the "auditory domain." This is different from the first paradigm in that it obtains an estimate $\mathbf{y}^{\mathrm{opt}}$ by minimizing $\mathcal{C}(\mathbf{y}, \mathbf{y}')$ whereas in the first paradigm, an estimate $\mathbf{x}^{\mathrm{opt}}$ is obtained by minimizing $\mathcal{C}(\mathbf{y}, \mathbf{y}')$. That is, the final obtained estimate in one case is in the time/frequency domain whereas in the other case, it is in the auditory domain. are in different domains. Although it overcomes the computational complexity bottlenecks associated with the first paradigm, it requires an inverse auditory mapping procedure that transform the auditory representation $\mathbf{y}^{\mathrm{opt}}$ back to its corresponding time/frequency representation $\mathbf{x}^{\mathrm{opt}}$. This inverse mapping procedure is not tractable due to the non-linearities, the many-to-one mappings and the dependence of the model parameters (such

as auditory filter shapes) on the input stimuli. That is, modifications done on the auditory model outputs (based on minimizing a distortion function) do not necessarily have a one-to-one correspondence with a time/frequency domain representation.

In this dissertation, we address the two problems described above and develop possible solutions to embed perceptual models into speech/audio applications in a straightforward and computationally efficient manner.

1.4 Motivation

There exist several motivating factors in developing solutions to directly embed perceptual models that follow either of the two paradigms described in Section 1.3. The following represents the most important ones:

- The need for computationally efficient techniques to solve perceptual objective functions.
- Development of elegant schemes to solve simple to more complex perceptual distortion criterion.
- Development of an inverse auditory mapping procedure to carry out the optimization in the auditory domain.
- Avoid the bottleneck with including perceptual methods in noisy conditions.

In this context, there exists related works that achieve either one or more of the above advantages when incorporating perceptual models. For example, in [39], the authors show from an information theoretic standpoint that that no information is lost during the different processing stages of an auditory model and therefore it is possible to develop an inverse procedure to synthesize a time/frequency domain signal from its auditory representation. In [40, 41], speech coding is carried out in perceptual domain. Here, the auditory model outputs were quantized before transmission. At the decoder, an inverse auditory mapping was proposed that synthesizes the time-domain signal from the auditory model outputs. This is in contrast to existing perceptual speech/audio coding algorithms that quantize spectral components based on masking thresholds. Similarly, in [42], the authors proposed a general framework to embed advanced auditory models in perceptual distortion functions. This was accomplished by developing a sensitivity matrix approach that approximates the auditory model reasonably well particularly in cases where small distortions are observed.

All of the above perceptual processing trends motivated us to further develop new solutions that enable one to integrate auditory models directly in perceptual distortion functions.

1.5 Contributions

In this dissertation, we provide possible solutions to address the high computational complexity and the inverse mapping problems by developing computationally efficient algorithms to existing auditory model implementations. Furthermore, a constrained mapping technique is developed that obtains a time/frequency domain vector while simultaneously minimizing a perceptual distortion function. This overcomes the need for an inverse mapping strategy. The constrained mapping scheme is incorporated in a speech enhancement and a sinusoidal component selection task. These developments have led to the following contributions during the course of this research:

A frequency/detector pruning approach for loudness estimation [38, 43]: The proposed frequency and detector pruning approach provides a framework for lowering the computational complexity associated with the auditory model evaluation stages. The main idea here is to prune the number of frequency components and detector locations in a perceptually relevant manner, i.e., the deviations of the estimated auditory model outputs from the true auditory model outputs (e.g., excitation/loudness patterns) should be minimal. To that end, two different algorithms have been proposed for the purpose of frequency/detector pruning. The first algorithm prunes the frequency components by uniformly approximating the spectral energy in each critical band by a single component. It then jointly estimates the best frequency location for the approximated component and the pruned detector locations by taking into account the shape of the auditory filters. The second algorithm is more efficient than the first algorithm and relies on a tone/noise classification in individual critical bands for frequency pruning. This is accomplished by obtaining an auxiliary pattern that is subsequently used for both frequency pruning and detector pruning. Experimental results indicate only a 4-7 % relative error in loudness while attaining up to 80-90 % reduction in computational complexity.

An hybrid algorithm for loudness estimation [44] The hybrid algorithm is developed specifically for use with sinusoidal signals and employs the proposed auditory pattern combining technique together with a look-up table to store representative auditory patterns. The hybrid algorithm evaluates the auditory pattern associated with a mixture of sinusoidal components in a computationally efficient manner. The main idea here is to store the representative auditory patterns in a look-up table and exploit the frequency separation between different sinusoidal signals. That is, for frequency separations less than a critical band, the masking phenomenon plays a major role and therefore all the stages of an elaborate auditory model are employed. For larger frequency separations between two sinusoids, the envelope of the individual auditory patterns are combined using the auditory pattern combining technique. The proposed hybrid scheme was further incorporated in a sinusoidal component selection task where it resulted in $80-90\%$ reduction in the computational complexity while maintaining a sinusoidal selection accuracy of 90 % compared to that of the iterative greedy algorithm.

Speech Enhancement using auditory patterns [45] Speech enhancement algorithms minimize a suitable error criterion in the time or spectral domain and include perceptual properties such as masking thresholds, non-uniform sensitivity of the auditory system only in a heuristic manner. The main idea here is to explicitly minimize the error between the auditory representations associated with the original signal and that associated with the enhanced speech signal. A constrained optimization problem that measures the distortion in the auditory domain is formulated and solved using interior point methods. Simulation results suggest that incorporating auditory models is beneficial particularly at low signal-to-noise ratios contrary to what is possible with current perceptual speech enhancement algorithms. Moreover, the proposed approach overcomes estimation of perceptual quantities such as masked thresholds from the noisy signal.

Sinusoidal Component Selection using auditory patterns [46] A series of techniques that pose the problem of selecting perceptually relevant sinusoids as a convex optimization problem are proposed. The proposed techniques maximizes the matching between the auditory excitation pattern associated with the original signal and that associated with a modeled version (represented by a small set of sinusoidal parameters) of the same signal. In particular, we propose three techniques that are not only computationally efficient but also result in similar levels of performance as compared to the greedy approaches.

1.6 Organization of Dissertation

The rest of the dissertation is organized as follows. In Chapter 2, we review the physiological and functional aspects of the human auditory system and the principles of psychoacoustics. In Chapter 3, a review of various auditory modeling techniques are presented and a detailed description of the various stages in the Moore and Glasberg auditory model [3] together with their computational complexity is presented. Chapter 4 presents the proposed frequency/detector pruning approach for a low-complexity loudness estimation procedure. Chapter 5 describes the proposed auditory-domain based speech enhancement algorithm. In Chapter 6, different perceptual strategies for the sinusoidal component selection task are described. Conclusions and directions for further research are presented in Chapter 7. Finally, Simulink implementations of Moore & Glasberg auditory model are presented in Appendix A. Simulink demos based on incorporating loudness measures are presented in Appendix B.

Chapter 2

Human Auditory System, Principles of Psychoacoustics and Auditory Models

In this chapter, an overview of the physiological and functional aspects of the human auditory system are provided. The physiological aspects include a description of the physical structure of the outer ear, the middle ear and the inner ear. The functional aspects help us to understand the different mechanisms employed by the human auditory system towards creating a perception. The field of psychoacoustics has been instrumental in understanding and characterizing the various mechanisms employed by the human auditory system. Therefore, an overview of the different psychoacoustic principles and the underlying psychoacoustic experiments are described in order to gain an understanding of the functional aspects of the auditory system.

2.1 Human Auditory System

The human auditory system is generally divided into three major parts namely the outer ear, the middle ear and the inner ear. A schematic layout of the human ear with labeled parts is shown in Figures 2.1 and 2.2. The physiological aspects are discussed next.

Outer ear

The outer ear consists of the pinna and the ear canal which are together responsible for collecting the acoustic stimuli and directing it toward the ear drum. The ear drum is present at the end of the ear canal. The ear canal acts as a 2 cm long open pipe which is resonant close to 4 kHz. This can also be observed from the "Threshold in Quiet" curve shown in Figure 2.4 which exhibits a minimum (due to maximum sensitivity) between the $2 - 4$ kHz region.

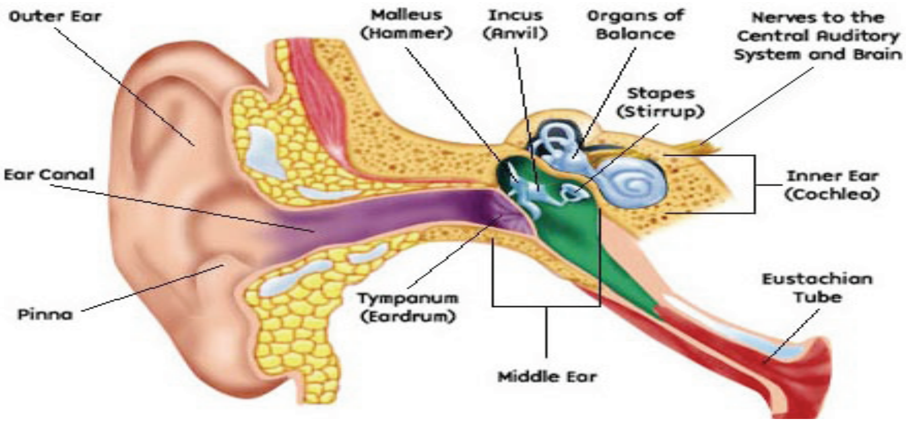

Figure 2.1: Structure of the human ear.

Middle Ear

As the name suggests, the middle ear is connected to the outer ear at the ear drum on one side and to the inner ear at the oval window on the other side. The middle ear consists of three bones referred to as the malleus, the incus and the stapes. These bones collectivity act as an impedance matching unit between the outer ear activity (consisting of air movement) and the inner ear activity (consisting of fluid movements). The best impedance match is however obtained roughly at a frequency of about 1 kHz. Therefore, the middle ear is responsible for transforming the air vibrations at the ear drum into fluid motions inside the inner ear.

Inner Ear

The inner ear is the most important part which is responsible towards human perception. The Cochlea is the primary organ present in the inner ear where the acoustic signal is processed to create hearing sensations. The Cochlea is a snail shaped structure which is wound two and a half times around itself thus forming a spiral structure. The spiralled version is shown in Figure 2.1 and an unwound version is shown for its details in Figure 2.2. The Basilar Membrane (BM) runs along the length of the Cochlea and separates the cochlear canal into two fluid

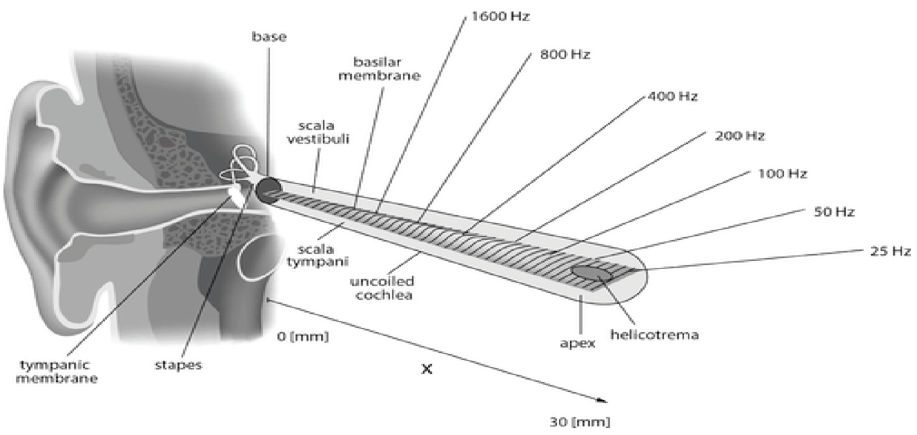

Figure 2.2: Frequencies points along basilar membrane.

filled regions known as the Scala Vestibuli and Scala Tympani. The oval window represents the base or the start of the cochlea and the apex represents the inner tip of the cochlea after about two and a half turns of the cochlea.

When the oval window is set in motion due to the movement of stapes, the fluid inside the inner ear is also set in motion and causes the BM to move. The response of the BM is responsible towards creating a perception. Therefore, several studies that characterized the functioning of the basilar membrane were reported in the literature. In the early 20^{th} century, Helmholtz postulated that the basilar membrane is composed of a series of separately tuned frequency resonators [47]. Later, in 1960, Von Bekesy provided evidence that there are a continuum of tuned frequency resonators along the basilar membrane rather than a set of fixed frequency resonators [48].

Both experiments revealed that the lower frequencies cause the apical end or the inner tip of the basilar membrane to vibrate, whereas the higher frequencies excite the basal end of the membrane. That is, each point on the Basilar Membrane is sensitive to a specific frequency. This is illustrated in Figure 2.2. Therefore, the higher the frequency, the lesser it travels along the membrane before it reaches its point of maximum response. This suggests that the BM is associated with different temporal delays corresponding to different frequency components, i.e., lower frequencies take longer time to travel along the BM before they reach their point of maximum response. This results in higher response times for low frequency components and vice versa for high frequency components.

The organ of corti with its sensory cells (also known as the hair cells) are spread throughout the length of the basilar membrane. The hair cells transform the basilar membrane's mechanical oscillations induced due to the action of the fluids into electrical nerve pulses which are sent to the brain through the attached nerve fibers. The nerve fibers maintain a spatial relationship with one another based on its originating location on the basilar membrane. These nerve fibers are fanned out from the auditory nerve which carries the nerve impulses to the brain where a sense of perception associated with that acoustic stimuli is created.

2.2 Principles of Psychoacoustics

As it is not possible to directly measure the hearing sensations produced by the human auditory system, the functional aspects of the human auditory system are studied by resorting to indirect methods of analysis. That is, a series of carefully designed experiments are carried out to study the mechanisms employed by the human auditory system in creating the corresponding hearing sensations. These experiments are termed psychophysical or psychoacoustic experiments, since they involve generating a physical stimuli (i.e., an acoustic signal) and recording the corresponding response provided by a human listener (without actually measuring the hearing sensations). For example, in one experiment, the human subject can be asked to rate "how loud a set of tones are" on a relative scale. Another example is to ask the subject to identify "when a particular tone is masked." The response of the subjects are used to understand the functional aspects of the auditory system. This has led to the field of Psychoacoustics.

The field of Psychoacoustics is concerned with studying the relationships between the acoustical stimuli presented to the ear and the hearing sensations that they correspond to. A brief overview of the general principles of psychoacoustics and the underlying psychoacoustic experiments are provided in this section. The absolute threshold of hearing, the masking phenomenon and the concept of critical bands constitute the fundamental principles of psychoacoustics. Other psychoacoustic principles include the spread of masking, the asymmetry of masking, simultaneous and temporal masking.

Almost all auditory models exploit the principles of psychoacoustics in modeling the human auditory system. For example, a simplest form of an auditory model is the absolute threshold of hearing curve that characterizes only the non-uniform sensitivity of the auditory system. More sophisticated auditory models exploit the frequency selectivity property that can be described in terms of the concept of critical bands and the phenomenon of masking.

The physical sound stimuli is measured in decibel units of the Sound Pressure Level (SPL) and is expressed in units of dB SPL. A dB SPL is defined as $SPL(dB) = 10log_{10}(I/I_0)$, where I and $I_0 = 10^{-12}$ denote the sound intensity (in watts/meter^2) associated with the acoustic stimuli and the reference stimuli respectively.

The Absolute Threshold of Hearing

The absolute threshold of hearing (ATH) is defined as the smallest intensity level (in dB SPL) of a pure tone that is just audible in a quiet surrounding. It describes the ability of the auditory system in detecting weak sounds. The threshold curve as a function of frequency is shown in Figure 2.4. A good approximation to the absolute threshold of hearing is given by the following non-linear function [49]:

$$ATH \text{ dB SPL} = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 \qquad (2.1)$$

where f denotes the frequency in Hz and ATH denotes the corresponding audibility threshold in dB SPL.

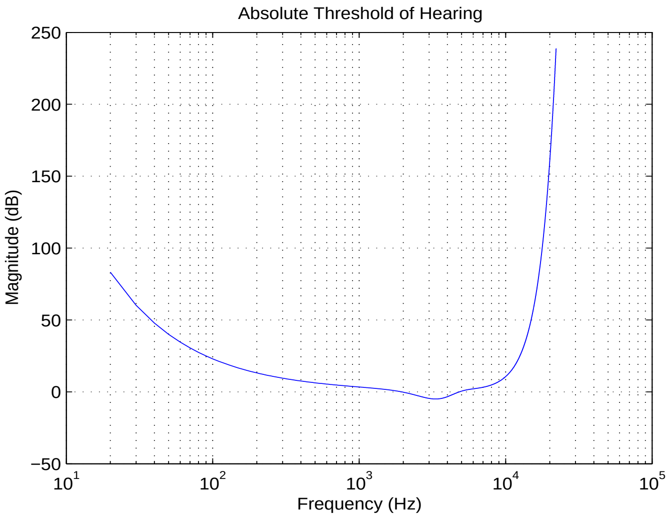

Figure 2.4: Absolute threshold of hearing.

Figure 2.4 represents the average threshold of a person with “normal” hearing ability. It can be observed that the threshold is not constant and represents the non-uniform sensitivity of the human auditory system across different frequencies. The combined effects of the outer and middle ear are greatly responsible towards creating the non-uniform sensitivity. Transmission is efficient for mid range frequency and drops off at high and low frequencies.

There exists variations of the absolute threshold of hearing curve depending on the method of measuring the intensity level. Two important variations are the minimum audible field (MAF) and the minimum audible pressure (MAP) [13]. The MAP threshold is obtained by measuring the sound pressure at some point close to the ear drum along the ear canal using a small probe microphone. The sound is usually delivered through headphones in this case. On the other hand, the MAF threshold is obtained by measuring the sound pressure at the center of the listener’s head after the head is removed from that position. In this case,the sound is usually delivered through loudspeakers in an anechoic chamber. MAP thresholds represent monaural listening conditions whereas MAF thresholds represent binaural listening conditions. On average, the binaural thresholds are about 2 dB lower than the monaural thresholds.

Applications : The absolute threshold of hearing has important consequences in many speech/audio applications. The following examples illustrate this:

- In coding applications, the bit-rate can be adjusted until the quantization noise level falls below this threshold of hearing.
- Secondly, the energy corresponding to 1-bit can be made to correspond to a minimum audible level, i.e., the intensity level of the lowermost point of the threshold curve (4 kHz).
- Thirdly, the headphones or the loudspeaker mode of presentation can decide whether the MAP or the MAF based thresholds should be incorporated in the design of the algorithm. For example, in the case of hearing aid devices that are matched to a particular ear, the MAP based thresholds can be employed.

Limitations : It should be noted that the thresholds represent minimum audibility levels only for tonal sounds and do not correspond to sounds that have a complex spectrum. This fundamental assumption should be considered during the design of any speech/audio algorithm that exploits this phenomenon. Moreover, these thresholds show variability in their shape across different age groups.

Critical Bands

In 1940, Harvey Fletcher [1] conducted a series of experiments to study the human hearing mechanism and suggested a model of highly overlapping bandpass filters with bandwidths equal to critical bandwidths (described later) for modeling the human auditory system. Fletcher's experiments consisted of detecting a pure tone in the presence of a noise band centered at the same frequency as that of the tone. He measured the detection threshold of the pure tone as a function of the noise bandwidth. In doing this, the noise power density I_f was held constant, i.e., the noise power increased as the bandwidth increased. The results of his experiments are summarized in Figure 2.5. It can be observed that the detection threshold or the intensity I_m at which the tone is detected increases until a certain noise bandwidth is reached and remains constant there after. Critical bandwidth then corresponds to that width of the noise band at which the detection threshold associated with the tone ceases to increase, i.e., any further increase in the noise bandwidth has no effect on the detection threshold (audibility) of the pure tone. His experiments also concluded that critical bandwidths are not constant across all frequencies and changes as a function of the center frequency. In Figure 2.5, each horizontal lines corresponds to a different center frequency where the tone and the noise band are centered. This horizontal line intersects the 45-degree line at different points which correspond to a different critical bandwidths.

Fletcher offered the following explanation to account for this phenomenon. He suggested that the basilar membrane can be thought of as a bank of overlapping bandpass filters. These bandpass filters, now known as auditory filters, are assumed to span the length of the basilar membrane. Hence any point on the basilar membrane reacts only to a narrow band of frequencies, which is responsible for the frequency-to-place transformation observed along the membrane. He explained that the detection threshold increased with the noise bandwidth as long as the noise bandwidth falls within the pass band of the filter, thereby masking the tone. When the noise bandwidth grows beyond the bandwidth of the filter, there is no additional masking as the components falling outside the passband are filtered out. This explains the horizontal lines seen in Figure 2.5. Fletcher assumed that the shape of the auditory filter is rectangular with bandwidths corresponding to the critical bandwidths.

Following Fletcher, other experiments [50–52] also established the notion

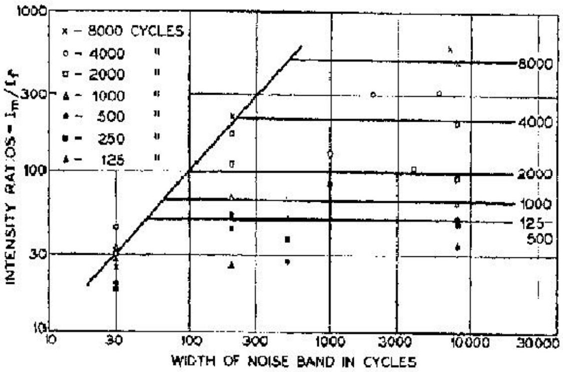

Figure 2.5: Plot of detection threshold as a function of noise bandwidth (from [1]).

of critical bands and their associated bandwidths. In [13], Zwicker and Fastl derived the following analytical expression that describes the dependance of critical bandwidth, $C_{BW}(f)$, on the center frequency, f:

$$C_{BW}(f) = 25 + 75[1 + 1.4(f/1000)^2]^{0.69} \text{ Hz}. \tag{2.2}$$

However, it should be noted that the above estimates of critical bandwidths are based on the assumption that the auditory filter shapes are rectangular in shape. Recent estimates based on more direct measures such as notched-noise experiments [16,53] suggest that the auditory filters are not rectangular in shape. In notched-noise experiments, a noise masker with a band-stop or notch is considered and the signal frequency is centered in the notch. This prevents detection of the signal due to the occurrence of beats. In [16], Patterson described a method of estimating auditory filter shapes from notched-noise experiments. He suggested a rounded top for the pass-band and exponential fall off in the stop-band of the auditory filter; it is now known as the rounded exponential or the "roex" model of auditory filter shapes. Based on the rounded exponential model [16], Glasberg and Moore estimated the auditory filter shapes at different center frequency and intensity level [54]. Critical bandwidths then correspond to the "effective bandwidth" of these auditory filters which are now referred to as equivalent rectangular

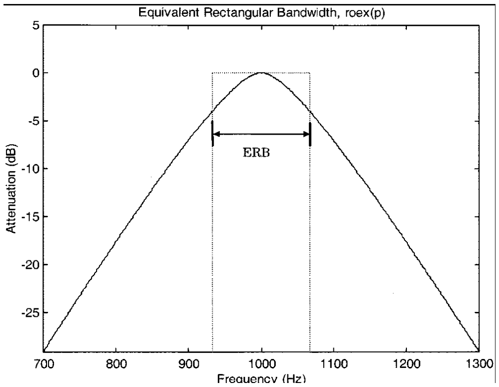

Figure 2.6: Auditory filter shape and its equivalent rectangular bandwidth (from [2])

bandwidths (ERB). This is illustrated in Figure 2.6. In [52], the authors give the following analytic expression to calculate the equivalent rectangular bandwidth:

$$ERB(f) = 24.7(4.37f/1000 + 1) \text{ Hz} \tag{2.3}$$

where f, $ERB(f)$ denote the frequency and equivalent rectangular bandwidth in Hz respectively.

Figure 2.7 shows a comparison plot between critical bandwidths defined according to (2.2) and equivalent rectangular bandwidths defined according to (2.3).

Following the notion of critical bandwidths, a scale more closely related to the way the human ear analyzes sound was developed. This scale is known as the

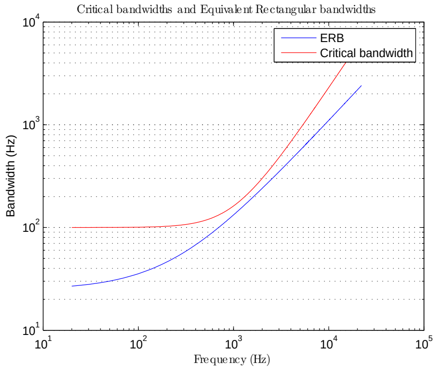

Figure 2.7: Critical Bandwidth vs. ERB

critical band-rate scale and is obtained by stacking individual critical bandwidths one next to the other such that the upper end of one critical band coincides with the lower end of the next critical band. The crossover points, which correspond to certain fixed frequencies, are tabulated against the number of critical bands that are present below that frequency as shown in Table 2.1. A unit of "Bark" (in honor of Barkhausen, who introduced the concept of loudness level) was proposed by Zwicker and his co-workers to measure distances along the critical band-rate scale [13]. For example, a difference of 1-Bark on this scale represents one critical bandwidth irrespective of the width of the band. The critical band-rate scale is taken to represent unit length distances along the basilar membrane and helps us in mapping frequency onto linear distances along the basilar membrane.

For the critical bandwidth definition introduced by Zwicker [51], the authors calculate the critical band-rate using the following relation:

$$Z(f) \text{ (in Bark units)} = 13 \arctan(0.76f/1000) + 3.5 \arctan\left(\frac{f}{7500}\right)^2 . \qquad (2.4)$$

Similar to the Bark scale in (2.4), an ERB scale measures the number of equivalent rectangular bandwidth auditory filters that can be fitted below any given frequency, f. An analytical expression relating the ERB number (also referred to as ERB units) to the frequency is given by [52]:

$$p \text{ (in ERB units)} = 21.4 \log_{10}(4.37f/1000 + 1) \qquad (2.5)$$

where p denotes the ERB number corresponding to the frequency f in Hz.

The Masking Phenomenon

The absolute threshold of hearing represents the audibility of single pure tones presented in a quiet surrounding under steady-state conditions. Steady-state conditions refers to a sound stimuli that exists for at least 200 ms duration. However, real-life signals (such as speech, audio or noise) are composed of a complex spectra containing several different frequency components as opposed to single tones. In

Table 2.1: Band edges and center frequencies for a collection of 25 critical bandwidth auditory filters.

Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz) (Hz)
1	50	$\cdots - 100$	14	2150	2000-2320
2	150	100-200	15	2500	2320-2700
3	250	200-300	16	2900	2700-3150
4	350	300-400	17	3400	3150-3700
5	450	400-510	18	4000	3700-4400
6	570	510-630	19	4800	4400-5300
7	700	630-770	20	5800	5300-6400
8	840	770-920	21	7000	6400-7700
9	1000	920-1080	22	8500	7700-9500
10	1175	1080-1270	23	10500	9500-12000
11	1370	1270-1480	24	13500	12000-15500
12	1600	1480-1720	25	19500	$15500 - \cdots$
13	1850	1720-2000			

this case, it is not straightforward to predict the audibility of a particular tone due to the presence of other neighboring frequency components. Moreover, the frequency components that constitute the complex spectrum can have arbitrary relative phase offsets with respect to each other and can exhibit different intensity levels in real-life. These variabilities make it difficult to predict their audibility directly from the absolute threshold curve. In such cases, it is required to somehow estimate the threshold of audibility by exploiting the mechanisms of the human auditory system.

Masking refers to the phenomenon where one sound is rendered inaudible in the presence of another sound. Masking phenomenon can be explained in terms of the auditory filter analogy described in previous section. For example, consider a tonal signal centered in a noise-band signal where the tonal signal is to be masked and the noise-band acts as the masker. The detection of the tonal signal (maskee) in presence of a noise-band (masker) depends on the amount of noise power that falls within the pass-band of the auditory filter that is centered at the tone frequency. This auditory filter is used to "listen" to the tonal input.

Masking occurs when this tonal stimuli cannot be detected due to the noise power exceeding a certain threshold power within the pass-band of the auditory filter. This threshold is referred to as the masking threshold and is widely employed in several audio coding algorithms [2]. In other words, the masker (loud sound) creates a certain degree of excitation along the basilar membrane that prevents the detection of a weaker excitation created by the maskee (soft sound).

The phenomenon of masking is widely studied in the psychoacoustic literature and provides a means to understand the frequency selectivity property of the human auditory system. Both the masking phenomenon and the notion of critical bandwidths characterize the frequency selectivity property of the human auditory system. In this section, we will review the different types of masking and the underlying psychoacoustic experiments that characterize the phenomenon of masking. A tutorial treatment of the masking phenomenon can be found in [2, 55, 56].

Depending on the order of occurrence of masker and maskee, the phenomenon of masking can be classified into the following two types:

1. Simultaneous Masking or Frequency Masking

2. Non-simultaneous Masking or Temporal Masking

Simultaneous Masking

As the name indicates, Simultaneous masking occurs when the masker and maskee are presented simultaneously, i.e., the frequency components associated with the masker and maskee occur simultaneously. For example, in Figure 2.8, two tonal components are presented simultaneously. The stronger tone represents the masker and the weaker tone represents the masked tone. The dotted lines in Figure 2.8 indicate that the threshold of hearing is raised in the vicinity of the masker. This modified threshold of audibility is known as the Masked thresholds. In general, it is usually sufficient to consider the following four simultaneous

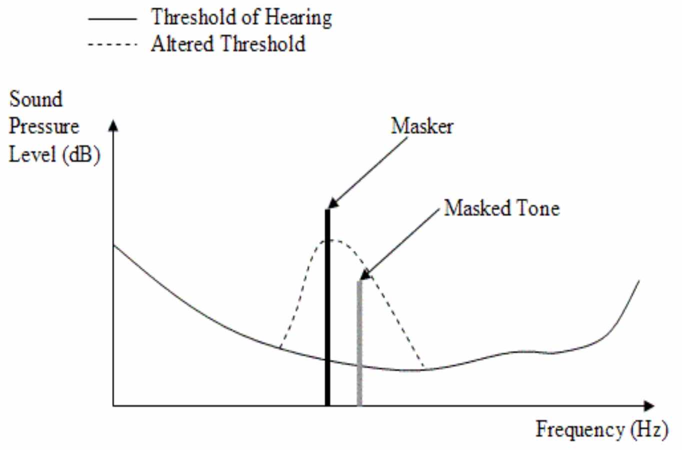

Figure 2.8: Illustration of Frequency Masking.

masking scenarios:

Noise – Masking – Tone : In this scenario, a tone that is to be masked is centered in the same critical band as that of a narrow-band noise masker. The masking ability of a signal (masker) at neighboring frequencies is measured by means of a Signal-to-Mask Ratio (SMR). A minimum signal-to-mask ratio is obtained when the masked tone is close to the center of the noise-band; the masking is most effective at this frequency. Psychoacoustic experiments [6,13] reveal that the minimum SMR is of the order of $1 - 5$ dB for the noise-masking-tone scenario. This suggests that noise is a better masker since the masked tone needs to have a much higher intensity in order to be audible in the presence of noise.

Tone – Masking – Tone : In a tone-masking-tone scenario, the masker and maskee are both tones. Measurement of masking thresholds in such experiments presents more difficulties due to the occurrence of beats [57]. The beats lead to the detection of an additional component created due to the action between the masker and the maskee signals. This can make an otherwise inaudible maskee signal audible. In [13], the masker and maskee were set 90 degrees out of phase, in the region where beating was observed. A minimum SMR of roughly 15 dB is observed in this case.

Tone – Masking – Noise : In the tone-masking-noise scenario, the tonal masker is present at the center of the critical band and the noise maskee of bandwidth smaller than one critical bandwidth is used. The minimum signal-to-mask ratio in this case usually ranges between $20 - 30$ dB [2].

Noise – Masking – Noise : In this scenario, a narrow-band noise signal masks another narrow-band noise signal. The masked thresholds for this scenario have been difficult to characterize due to the phase relationships that exist between individual frequency components of the masker and maskee noise signals [2].

Non-simultaneous Masking

Masking can also extend in time when the masker and maskee are presented in succession and not together as in the case of simultaneous masking. This type of masking is referred to as Temporal Masking. They can be classified into Premasking and Post-masking depending on the relative onset of the maskee with respect to the masker. Pre-masking lasts about $10 - 20$ ms whereas post-masking is a stronger effect and lasts longer for about $100 - 150$ ms as illustrated in Figure 2.9.

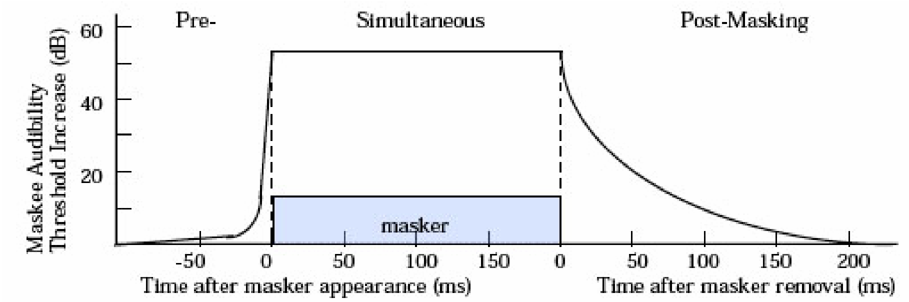

Figure 2.9: Illustration of Temporal Masking.

Pre – masking : Pre-masking refers to the situation where the masking occurs even before the onset of the masker. It is a poorly understood phenomenon as the maskee signal is rendered inaudible even before the onset of the masker. One possible explanation is that the ear is associated with a certain integration time to create a perception associated with any sound. When the masker and maskee are presented close in time such that they fall within this integration time, the perception associated with the masker builds up faster than that of the maskee and is strong enough to render the maskee inaudible. This phenomenon can be characterized by means of a pre-masking threshold wherein the maskee is rendered inaudible whenever the intensity level of the maskee falls below this threshold.

Pre-masking is exploited in the design of audio coders in an attempt to mask the pre-echo distortion that results when the energy of the coded signal falls prior to its actual onset. This occurs frequently in transient segments with abrupt or sudden bursts of energy. Window switching was a popular technique that was used to overcome pre-echo distortion [19]. It relies on switching the shape and length of the window to control the spread of the pre-echo distortion resulting from coding that segment. More recently, Temporal Noise Shaping [58, 59] is used to control the pre-echo distortion that arises from coding abrupt changes in audio segments.

Post – masking : Post-masking refers to the case, where the masking phenomenon is observed after the masker is switched off. Unlike pre-masking, post-masking is a better understood phenomenon. Experimental results have shown evidence that post masking depends on the stimulus frequency, masker intensity and signal delay [60]. Several audio coding algorithms that exploit post masking phenomenon have been proposed in the literature [2].

Asymmetry and Spread of Masking

The tone-masking-noise and the noise-masking-tone scenarios show a marked difference in their minimum signal-to-mask ratios. The minimum SMR for a tone-masking-noise is about $20-30$ dB whereas the minimum SMR for noise-masking-tone is about $1-5$ dB indicating that noise is a better masker compared to tones. There exists asymmetry in the masking powers of the tone as opposed to the noise. This is referred to as the "The asymmetry of masking." This has implications in the design of audio coders as the quantization noise introduced can now be shaped to fall under the the noise bands rather than the tonal bands in order to render the quantization noise inaudible.

Spread of Masking refers to the ability of a masker to not only mask frequency components present within the critical band but also influence the detection thresholds of frequency components present in neighboring critical bands. That is, the masking effect extends to neighboring critical bands as well. This effect is usually modeled by means of a spreading function in several psychoacoustic models [19].

Chapter 3

Auditory Models for Loudness Estimation

3.1 Introduction

Based on the principles of psychoacoustics, several different auditory models [3,8, 10,11,13,17,61] have been developed over the years to mimic the functioning of the human auditory system. One attribute of human perception that is considered by these auditory models is Loudness. Loudness is a subjective phenomenon which represents the magnitude of perceived intensity, i.e., it is a measure of the magnitude of neural activity that corresponds to the hearing sensations. It is measured in units of sones (to be described later) which is different from measuring the signal intensity in dB SPL units.

The loudness estimation algorithms can be classified into two broad categories:

- Algorithms that measure the loudness level (in phons)
- Algorithms that measure the loudness (in sones)

In the subsequent sections, we will highlight the difference between loudness level and loudness and also present an overview of algorithms that fall into these two categories.

3.2 Early Loudness Estimation Techniques: A Review

Since loudness is a subjective quantity, it cannot be measured directly. Early attempts to solve this problem were addressed by the magnitude estimation and magnitude production techniques [62]. Magnitude production requires the subjects to adjust the level of a test sound until the test sound has a certain loudness relationship with the loudness of a reference sound (usually a 1 kHz tone). For example, the subject can be asked to judge when a test sound is half as loud or twice as loud as the reference tone. On the other hand, in magnitude estimation

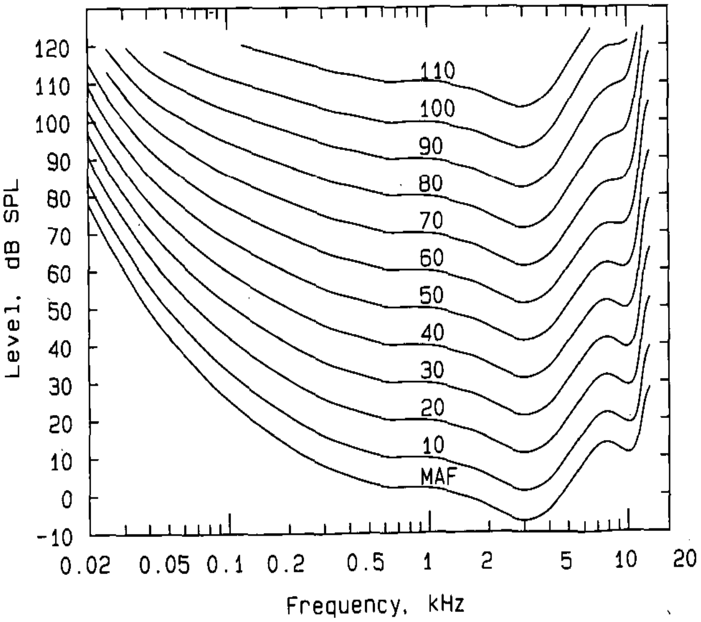

Figure 3.1: Equal Loudness Contours [3].

technique, the user is presented with sounds of different intensity levels and instructed to assign a number to each of them according to their perceived loudness. Unlike the magnitude production technique, there is no adjustment of test sound intensity in this case. The users can also be asked to rate the perceived loudness of each sound relative to a reference stimuli.

Limitations : The above techniques are time consuming to carry out and therefore are not practical. Furthermore, they are prone to errors in judgment. Therefore, such methods of loudness estimation cannot be incorporated into automated signal processing algorithms.

Equal Loudness Contours

In order to overcome the difficulties associated with the manual estimation procedure, it is necessary to understand the functioning of human auditory system and emulate its behavior. In 1933, a first step towards this was made by Fletcher and Munson who studied the dependence of loudness on frequency and intensity of individual tones in [5]. They characterized the sensitivity of human hearing at different frequencies and set the loudness of a 1 kHz tone as a reference for comparative purpose. In particular, they adjusted the intensity level of individual tones until they are perceived equally loud as that of the 1 kHz tone. Here, the 1 kHz tone is presented at a fixed intensity level. The results of the experiment are summarized in Figure 3.1 and are collectively known as the Equal Loudness Contours. That is, the points along any contour represent points of equal loudness. Therefore, two tones of different frequencies can sound equally loud even if they do not have the same intensity level. This indicates that the dB measure does not correspond to the actual loudness perception.

Therefore, a different scale related to human perception is needed to measure the loudness where tones with different intensity but same loudness are represented with the same numerical magnitude on this scale. To that end, a loudness level scale was adopted and is measured in units of phons.

Loudness level is defined as the intensity level of a 1 kHz tone that is perceived as loud as the sound under consideration. Since the frequency of 1 kHz tone is chosen as reference the loudness level of a 1 kHz tone is equal to its sound pressure level in dB SPL. That is, a 1 kHz tone at 50 dB SPL has a loudness level of 50 phons. The following observations can be made from Figure 3.1.

- The lowest curve represents the threshold in quiet or the absolute threshold of hearing (also shown in Figure 2.4) and corresponds to a loudness level of 3 phons.

- At low loudness levels, the equal loudness contours are almost parallel to the absolute threshold curve. However, they become flatter at higher loudness levels. Therefore, the growth of loudness is different at different frequencies. For example, it can be observed from Figure 3.1 that the intensity levels of 100 Hz and 1000 Hz tone at absolute threshold (i.e., loudness level of 3 phons) are 26 dB and 2 dB respectively. At a loudness level of 100 phons, the intensity levels of the two tones are 106 dB and 100 dB respectively. This indicates that the intensity level of the 100 Hz tone must be increased by 80 dB while that of the 1000 Hz tone must be increased by 98 dB to get the same increase in their loudness level. Therefore, the growth of loudness is higher at lower frequencies compared to the mid and high frequency regions.

Applications : The equal loudness contour properties are exploited in the following applications:

1) Loudness Controls: Equal loudness contours have been incorporated in many loudness control circuits. The main objective of a loudness control circuit is to control the overall loudness level despite the fluctuations in the intensity levels of the input stimuli. In doing so, the loudness control circuits also compensate for the uneven growth of loudness across different frequency regions. For example, at low listening levels, they boost the bass and the treble frequencies relative to the mid frequencies to compensate for the lower contribution of loudness from these frequency regions.

2) Loudness Meters: On the other hand, loudness meters measure the loudness level associated with any complex sound. They make use of a modified version of equal loudness contours for this purpose. These modified versions are knows as the frequency weighting functions. There are several variants of frequency weighting functions and an overview of them is provided in the next section.

Frequency weighting functions

The frequency weighting function is a popular technique used for the purpose of loudness level estimation. They are derived from the equal loudness contours and evaluate the loudness level associated with an input stimuli with arbitrary frequency spectrum.

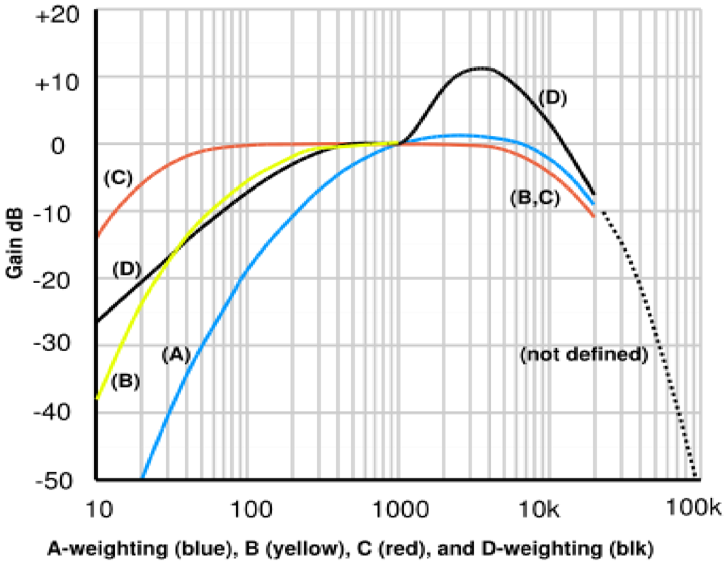

Figure 3.2: A, B, C Weighting curves.

To obtain the loudness level of a signal with complex spectral shape, the individual frequency components are scaled according to a predetermined function so that the scaled levels correspond to the loudness level in phons. Finally, these scaled intensities are summed to obtain the overall loudness level associated with the complex sound.

There are several variants of the frequency weighting functions. The most popular ones include the A, B, and C-weighting functions [12]. In Figure 3.2, a plot of the three frequency weighting functions are shown.

- The A-weighting is based on the 30-phon equal loudness contour and is a good approximation for sounds presented at low intensity levels. At low sound levels, the ear is insensitive to low frequency components, i.e., the low frequencies contribute little towards the total loudness of the sound. Hence, the A-weighting function attenuates the low frequency components such that their contribution to measurement of loudness level is reduced.

- The B-weighting is used for intermediate sound intensity levels and can be obtained by inverting the 70-phon equal loudness contour.

- The C-weighting is used at high intensity levels. At higher sound levels, the equal-loudness contours are almost flat.

In practice, sound level meters express the measured level in dB along with the particular weighting scheme used. For example, 40 dBA implies that the total loudness level corresponds to 40 phons by making use of the A-weighting function. However, a few limitations associated with this approach present difficulties in measuring the actual loudness perception.

Drawbacks : In practice, the sound spectrum can span a wide dynamic range of intensities between the threshold of hearing and the threshold of pain. In such cases, using a particular weighting scheme will lead to significant errors in the loudness level estimates since each weighting scheme is tuned for a particular range of intensities. The A-weighting is suited for low listening levels, whereas the B-weighting and C-weighting functions are suited for the medium to high listening levels. Moreover, the phenomenon of masking is not captured by these weighting functions thereby resulting in poor modeling of the human auditory system.

3.3 The Phon vs. the Sone scales

In the previous section, a subjective scale that measures the loudness level in phons was introduced. Furthermore, several frequency weighting functions were introduced to estimate the loudness level. However, the loudness level measurement technique represents only an indirect method to map the intensities to the loudness perception associated with the 1 kHz tone.

Limitations of the phon scale

One of the primary limitations of the phon scale is that they do not correspond to the actual subjective scale of loudness perception; they only correspond to the dB SPL level of a 1 kHz tone that is equally loud as the sound under consideration. For example, consider a 1 kHz tone presented at 40 dB SPL. Increasing the intensity of this tone to 50 dB results in a doubling of the actual loudness perception. However, in terms of the measured loudness level, it does not correspond to a doubling in the number of phons. The loudness level of the original tone presented at 40 dB is 40 phons and that presented at 50 dB is 50 phons. In this case, a 10-phon increase in loudness level of a 1 kHz tone actually sounds twice as loud as the original signal. Hence, it is not straightforward to judge the loudness relationship of how loud one stimuli is with respect to the other on the phon scale. This necessitates the development of a true subjective scale that measures the loudness perception similar to that of the human auditory system. In the next section, we describe another scale for measuring loudness perception.

The Sone Scale

The development of a scale of loudness was pioneered by S. Stevens and described in the classic paper [63] published in 1936. Stevens proposed the unit of sone to express the loudness of any given sound. The validity of this scale of loudness was further confirmed in a number of studies reported in [1].

One sone is defined as the loudness of a 1000 Hz tone presented at 40 dB SPL presented binaurally from a frontal direction in free field conditions. Since, a 1000 Hz tone at 40 dB SPL corresponds to a loudness level of 40 phons, a loudness measure of one sone also corresponds to a loudness level of 40 phons. There also exists a direct relationship between the loudness level (in phons) and loudness (in sones) and is given by:

$$S \text{ (in sones)} = 2^{\frac{L-40}{10}}, \text{ if } L \geq 40 \tag{3.1}$$

where L represents the loudness level in phons and S represents the loudness in sones. This relation does not hold for sounds with loudness level below 40 phons.

According to (3.1), a tone with a loudness level of 50-phon will have twice the loudness as that of a tone with a loudness level of 40-phon. On the sone scale, a sound stimuli with a loudness of 2 sones will actually be twice as loud as that of a stimuli whose loudness is 1 sone. Therefore, this loudness scale is consistent with human auditory system's mechanism or arranging sounds according to their loudness relationship. This loudness level to loudness mapping in (3.1) can also be used in loudness meters to convert the estimated loudness level into loudness measure. However, it should be noted that the errors made in loudness level measurements will be carried over to the loudness measures when the mapping in (3.1) is used.

It should be noted that the loudness level and loudness are two different quantities and are expressed in units of phon and sone respectively. The phon scale expresses the level (in dB SPL) of a 1 kHz tone that is equally loud as the test stimuli, whereas the sone is an absolute scale of loudness and is taken to represent true subjective perception.

The Scaling of Loudness

Following the loudness level estimation techniques, a first model of loudness that estimates the loudness directly from the intensity level was proposed in 1961 by Stevens in [7]. This was commonly referred to as the ”power law of loudness” in the literature and can be expressed according to:

$$S = KI^{0.3} \tag{3.2}$$

where the loudness S denotes the loudness in sones, I represents the intensity in linear power units (watts/meter^2) and k is a constant that depends on the subject.

3.4 Neural Activity based Loudness Estimation Algorithms

Although, Steven's power law relationship specifies a general model relating the loudness perception to the intensity of input stimuli, it does not provide a comprehensive view of the different processes happening in the auditory system. It is now well known from the psychoacoustic experiments [13, 62] that the phenomenon of masking and the notion of critical bandwidths are exploited by the auditory system. It is therefore necessary to develop auditory models that mimic these different auditory mechanisms more precisely.

Therefore, several new auditory models [3, 8, 10, 11, 61] proposed in the recent past are based on the idea of obtaining internal auditory representations corresponding to an input stimuli. The underlying idea behind all these models is that loudness is proportional to the amount of neural activity evoked by an input stimuli along the length of cochlea. Therefore, these models try to obtain an accurate representation of the neural activity pattern. The loudness is then calculated as the area under this neural activity pattern. The general flow of the processing stages in such models is shown in Figure 3.3. In this section, an overview of these elaborate auditory models will be presented.

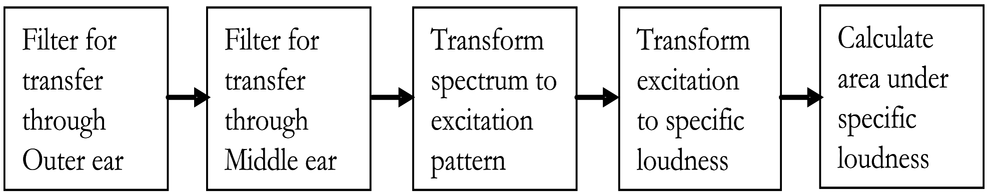
Filter for transfer through Outer ear
Filter for transfer through Middle ear
Transform spectrum to excitation pattern
Transform excitation to specific loudness
Calculate area under specific loudness

Figure 3.3: Basic structure of Loudness estimation algorithms.

In [5], Fletcher and Munson proposed a method to measure the neural activity by combining the masking patterns associated with individual tones. Although this method worked well for tones that are spaced far away, it did not accurately predict the loudness when a large number of components are closely spaced (e.g., noise bands). This is due to the fact that for closely placed components, the loudness of one component is largely determined by the masking effects of the other neighboring components. In [6], Fletcher and Munson studied the relationship between loudness and masking phenomenon in order to obtain better loudness predictions for closely spaced components. Their model involved obtaining a masking audiogram from the intensity spectrum of the sound. The masking audiogram specifies the number of dB units that the intensity of a pure tone should be raised from its threshold in quiet to be just audible. In other words, it calculates the difference between the masked thresholds and the threshold in quiet. Since, the internal auditory representations are difficult to directly measure, the masking audiogram was taken to represent an indirect measure of the magnitude of internal auditory representation. The loudness estimates were obtained by calculating the area under the masking audiogram. Although this method performed satisfactorily for noise bands, they performed poorly for tone complexes. One reason for this behavior was the fact that it is difficult to obtain masking audiograms for tone complexes due to the occurrence of beats.

In [61], Zwicker and Scharf proposed a more generic model that was applicable to both tonal and noise-like signals. Rather than blindly obtaining the masking patterns of any complex sound from noise bands as suggested in [6], an accurate representation of the masking patterns is obtained by employing a bank of band-pass auditory filters. These auditory filters accounts for the critical band nature and the masking phenomenon of the auditory system. In fact, each auditory filter has a bandwidth corresponding to the critical bandwidth and their pass-band and stop-band characteristics models the extent of masking. The masking pattern obtained in this manner is converted to an excitation pattern that represents the excitation level of the basilar membrane along its length. Following this, a loudness pattern is obtained from the excitation pattern by following a procedure similar to Stevens power law. The specific loudness pattern represents the loudness density or loudness per unit Bark. Finally, the total loudness is obtained as the integral of the specific loudness pattern across the entire critical band-rate scale.

In [3, 64], Moore and Glasberg presented several modifications and extensions to the one proposed by Zwicker in [61]. Two important differences are the excitation pattern calculation and shape of the auditory filters employed. The revised model calculates the excitation pattern directly from the intensity using analytic expressions rather than following the two-step approach employed in [61]. In the two-step approach, a masking audiogram is initially obtained followed by its transformation to obtain the excitation pattern. Secondly, recent evidence from notched-noise experiments reported in [52, 54] suggest a rounded exponential shape for the auditory filters rather than a rectangular shape for auditory filters as assumed in [61]. Therefore, the revised model takes into account this change in the auditory filter shapes when calculating the excitation pattern. Other minor improvements in the model can be found in [3, 64].

In [8, 9], an auditory model that simulates the detection characteristics of the human auditory system is described. That is, not all the changes in input signal can be detected by the auditory system. It was hypothesized in [8] that if the mean difference between two auditory representations exceeds a certain threshold, they are detected by the auditory system with a high probability. Therefore, the objective behind the design of the detector is to discriminate between two auditory representations much like that of the human auditory system. In general, the model includes several pre-processing and nonlinear processing stages followed by the design of an optimal detector as a decision device. It also accounts for both simultaneous and non-simultaneous masking and allows tuning of other intermediate stages based on feedback from the detector.

In addition to the above auditory modeling techniques, several different auditory filter shapes [10, 11, 15, 16] have been experimented by researchers in the past. For example, Patterson considered the use of Gammatone filters for modeling the auditory filter responses from notched-noise experiments [15]. The

Gammatone filter [15] has the following impulse response function:

$$h(t) = kt^{(n-1)} \exp(-2\pi Bt) \cos(2\pi f_c t + \varphi) u(t) \tag{3.3}$$

where n is the order of the filter, B is its bandwidth, f_c is the center frequency of the filter, φ is the phase offset and k is a gain parameter. When the order of the filter is in the range $3-5$, the magnitude characteristic of the Gammatone filter is very similar to that of the "rounded exponential" roex(p) model [16] of auditory filters. The rounded exponential model was also developed by Patterson in 1986. The Gammatone filters are symmetric, linear and level independent. That is, the filter shapes do not depend on the intensity level of the input stimuli.

This is in contrast to the data obtained from psychophysical experiments which indicate that the auditory filters are asymmetric, non-linear and level-dependent. To overcome these limitations, Irino and Patterson proposed a Gammachirp filter bank in [10]. Similarly, in [11] a dual resonance nonlinear (DRNL) filter was proposed as an alternative to the widely employed Gammatone filterbanks. The DRNL filter is implemented as a cascade of several first-order Gammatone filters.

These advances in modeling the auditory system has led to the development of several software packages. The following packages are popular

i) HUT EAR Package [65],

ii) Auditory Image Model (AIM) [66].

In the next section, a detailed description of the Moore & Glasberg auditory model described in [3,4] is presented. In addition, an analysis of the computational complexity in the different processing stages is also presented.

3.5 Moore & Glasberg Model of Loudness Estimation

Although there exists several auditory models, we employ the Moore and Glasberg model of loudness estimation in this dissertation. A number of reasons motivate the selection of this model:

1. It has been standardized by ANSI in 2005 as a new loudness standard [67].

2. This newly adopted ANSI standard performs satisfactorily for tonal as well as broadband spectral content [67]. Therefore, loudness predictions for general speech and audio signals with arbitrary spectral content can be obtained.

3. It also predicts loudness reasonably well for sounds at or below 40 phons (i.e., absolute threshold of hearing).

4. It incorporates recent results from psychoacoustic research regarding critical bandwidths and the possible shape of auditory filters [13]. In particular, critical bandwidths are now based on an equivalent rectangular bandwidth (ERB) measure and a new auditory scale in terms of the ERB unit was further developed.

5. Finally, it makes use of the roex(p) model of auditory filter shapes to characterize the magnitude response of filter frequency response. The roex model presents a computationally efficient alternative to the Gammatone auditory filters generally used in earlier auditory models. The Gammatone auditory filters are level independent whereas the rounded exponential filter model are level dependent and consistent with the observed psychophysical data.

Therefore, it remains one of the most sophisticated auditory model currently used by several researchers across the world to incorporate auditory mechanisms. The block diagram of the Moore and Glasberg loudness estimation process is shown in Figure 3.4. The Moore & Glasberg model [3] consists of the following processing stages:

1. Spectral analysis of the incoming audio.

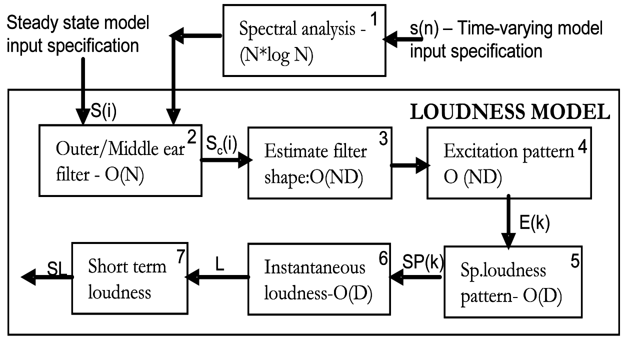
Steady state model input specification
Spectral analysis - (N*log N)
1
s(n) – Time-varying model input specification
LOUDNESS MODEL
S(i)
Outer/Middle ear filter - O(N)
2
S_c(i)
Estimate filter shape:O(ND)
3
Excitation pattern O (ND)
4
E(k)
Sp.loudness pattern- O(D)
5
SP(k)
Instantaneous loudness-O(D)
6
L
Short term loudness
7
SL

Figure 3.4: Block diagram of the Moore and Glasberg model [3, 4] of loudness estimation.

2. Fixed filter representing the transmission characteristics of the outer and the middle ear.

3. Estimation of the auditory filter parameters (e.g., filter slopes and bandwidths).

4. Calculation of an excitation pattern.

5. Transformation of the excitation pattern to a specific loudness pattern.

6. Calculation of total loudness.

7. Short-term and Long-term loudness calculation.

Transmission through outer ear

The transmission characteristics through the outer ear models the transformation that the sound undergoes as it reaches the ear drum. The transfer function is defined as the ratio of the free-field sound pressure measured at a position corresponding to the listener's head to the eardrum sound pressure. It should be noted that the transmission characteristics change with the type of incidence (free-field or diffuse-field) and the angle of incidence of the incoming sound. Typically, a linear filter is used to model the outer ear transmission characteristics.

Transmission through middle ear

Zwicker and Fastl studied the transmission characteristics of the middle ear and assumed that the inner ear is equally sensitive to all frequencies below 2000 Hz [13]. That is, tones of different frequencies having equal intensities at the ear drum results in an equal magnitude of sensation along the basilar membrane. On the other hand, the absolute threshold of hearing at these frequencies is not uniform thereby suggesting that the non-uniform behavior be attributed to the middle ear transmission characteristics. However, instead of modeling the middle ear filter based on the shape of the absolute threshold of hearing, Zwicker assumed that the transmission was uniform below 2000 Hz [13] and attributed the rise in absolute threshold at these low frequencies to an increased internal noise at these frequencies.

However, recent evidence [13] suggests that the increase in the absolute threshold of hearing at low frequencies can only be partly attributed to internal noise and partly to the middle ear transmission characteristics. To account for the rise in absolute thresholds not captured by both these factors, it was assumed that an increase in the level of the internal excitation at absolute threshold is required at these low frequencies. This is equivalent to saying that the inner ear is less sensitive at these low frequencies, i.e., the cochlea has a lesser gain at these low frequencies. Biologically, this mechanism may have evolved to give less gain to the internal noise present at the low frequencies, otherwise the noise will be amplified.

In view of the above considerations, Moore and Glasberg assume that the inner ear is equally sensitive to all frequencies above 500 Hz in their revised model described in [3]. The middle ear filter therefore corresponds to an inverted shape of the absolute threshold curve at these frequencies. The reduced cochlear gain at frequencies below 500 Hz is accounted for by defining a minimum excitation level required for a tone to be at detection threshold. The middle ear filter is then designed at all the frequencies such that they gave correct prediction of the absolute threshold as specified in ISO $389-7$ [68] (particularly below 500 Hz).

The combined transmission characteristics of the outer and middle ear can be modeled with a single filter whose frequency response is shown in Figure 3.5.

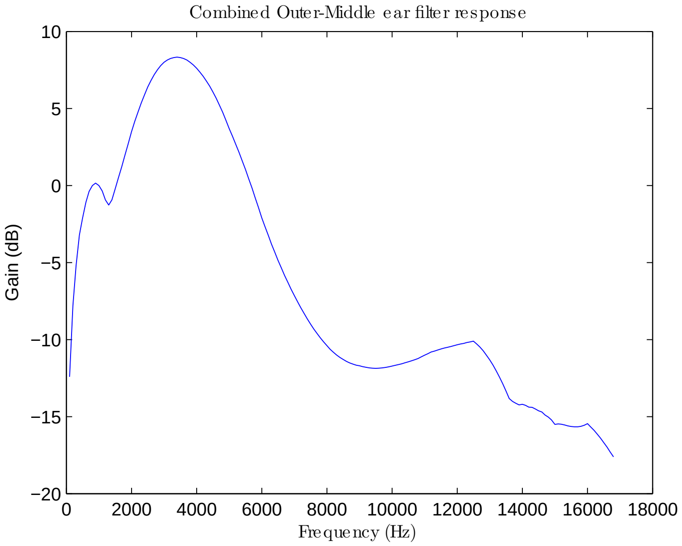

Figure 3.5: Combined outer and middle ear filter response.

The input signal $x(n)$ is referenced to an assumed sound pressure level (SPL) of P dB. Let $S_x(\omega_i)$ denote the power spectrum of $x(n)$ where $\omega_i = e^{\frac{j2\pi f_i}{f_s}}$ and f_s denotes the sampling frequency. If $|M(\omega_i)|$ denotes the frequency response of the outer/middle ear filter, then the effective power spectrum reaching the inner ear is $S_x^c(\omega_i) = |M(\omega_i)|^2 S_x(\omega_i)$.

Excitation pattern calculation

The excitation pattern represents the magnitude of the basilar membrane vibrations, i.e., it corresponds to the sensation level observed along the basilar membrane. The excitation pattern is calculated from the effective spectrum reaching the inner ear after transmission through the outer and middle ear stages. Different points along the basilar membrane are tuned to different frequencies and therefore react to a narrow band of incoming frequency components. This process is modeled using an overlapping bank of band-pass filters called auditory filters. The excitation pattern is then evaluated as the output of these auditory filters to the effective spectrum reaching the inner ear.

The steps associated with evaluation of the excitation pattern is described next. The frequency scale is first transformed into an auditory scale that is measured using an equivalent rectangular bandwidth (ERB) number and is calculated according to:

$$d \text{ (in ERB units)} = 21.4 \log_{10}(4.37f/1000 + 1) \tag{3.4}$$

where d represents the ERB number and f denotes frequency in Hz. The ERB number represents the number of equivalent rectangular bandwidth auditory filters that can be fitted below any frequency.

Let $L_r = \{d_k || d_k - d_{k-1}| = 0.1, k = 1, \cdots, D\}$ denote the reference set of D detector locations, such that they are uniformly spaced at 0.1 ERB units along the auditory scale. These detectors represent discrete sample locations where the excitation pattern is evaluated. Each detector d_k further represents the centers of the auditory filters used during the loudness estimation process. Let $\{cf_k\}_{k=1}^{D}$ denote the center frequencies (in Hz) corresponding to the centers $\{d_k\}_{k=1}^{D}$ (in ERB units) of the auditory filters.

The excitation pattern, $EP(k)$, is now evaluated as the output of these auditory filters to the effective spectrum reaching the inner ear, and is calculated according to:

$$EP(k) = \sum_{k=1}^{N} (1 + p_k g_{k,i}) \exp(-p_k g_{k,i}) S_x^c(\omega_i), \text{ for } 1 \le k \le D, \tag{3.5}$$

where p_k denotes the slope of the auditory filter centered at detector d_k and $g_{k,i} = |(f_i - cf_k)/cf_k|$ denotes the normalized deviation of f_i from cf_k. Here, f_i's denotes the frequencies (in Hz) corresponding to the spectral components of the input signal and $S_x^c(\omega_i)$ denotes the effective spectrum reaching the inner ear.

Furthermore, it should be noted that although (3.5) appears linear in $S_x^c(\omega_i)$, the values p_k change as a function of $S_x^c(\omega_i)$. Therefore, the slopes $\{p_k\}_{k=1}^D$ in (3.5) have to be evaluated every time there is a change in the intensity level of of the effective spectrum $S_x^c(\omega_i)$ reaching the inner ear. The dependance of the auditory filter slopes on the intensity level of incoming audio is described next.

Intensity Pattern and Auditory filter slope evaluation

The auditory filter slope evaluation depends on the intensity level of the effective spectrum reaching the inner ear. More specifically, it depends on an intermediate quantity known as the Intensity pattern. The Intensity pattern, $I(k)$, represents the total power within one ERB unit surrounding the detector d_k and is given by,

$$I(k) = \sum_{i \epsilon A_i} S_x^c(\omega_i), \text{ where } A_i = \{i | d_k - 0.5 < f_i^{erb} \leq d_k + 0.5\} \tag{3.6}$$

where f_i^{erb} denotes the ERB number corresponding to the input frequency f_i (in Hz) obtained using (3.4).

Each auditory filter has a rounded top and an upper and lower skirt (slope) parameter. In [52], it was assumed that the upper skirt parameter is fixed and does not change with intensity of the incoming audio. However, the lower skirt parameter still changes as a function of the intensity level. The upper and lower skirt parameters are given by [52]:

$$p^l = p^{51} - 0.38(p^{51}/p^{51}_{1000})(I(k) - 51), \tag{3.7a}$$

$$p^u = p^{51}. \tag{3.7b}$$

where p^{51} and p^{51}_{1000} are constants and can be calculated according to:

$$p^{51} = 4cf_k/CB(cf_k), \tag{3.8a}$$

$$p^{51}_{1000} = 4cf_k/CB(1000). \tag{3.8b}$$

In (3.8a)-(3.8b), the critical bandwidth $CB(f)$ represents the critical bandwidth (in Hz) associated with a center frequency f (in Hz) and is given by [13],

$$CB(f) = 24.67(4.368\frac{f}{1000} + 1). \tag{3.9}$$

In (3.5), the upper or lower skirt parameter is selected based on the sign of the normalized deviation $g_{k,i}$, i.e.,

$$p_k = \begin{cases} p^u & \text{if } g_{k,i} \geq 0, \\ p^l & \text{if } g_{k,i} < 0. \end{cases} \tag{3.10}$$

That is, the appropriate filter slope is selected based on the frequency location of the spectral components.

Specific loudness pattern calculation

The specific loudness pattern represents the action of the cochlea on the basilar membrane vibrations (i.e., the excitation pattern). It gives a measure of the neuron firing rate along the length of the cochlea and represents the loudness density, i.e., loudness per ERB.

The specific loudness pattern, $SP(k)$, is usually obtained through a non-linear transformation of the excitation pattern $EP(k)$ similar to the power law proposed by Stevens [7]. For moderate sound level intensities between $30 - 100$ dB, the transformation can be expressed mathematically as [3]:

$$SP(k) = c((EP(k) + A(k))^{\alpha} - A(k)^{\alpha}), \text{ for } k = 1, \cdots, D \tag{3.11}$$

where $c = 0.047$ and $\alpha = 0.2$ and $A(k)$ is a frequency dependent constant which is assumed to be equal to twice the peak excitation produced by a sinusoidal signal at absolute threshold for frequencies greater than 500 Hz. i.e., $A = 2E_{THRQ}$. For frequencies less than 500 Hz, the gain applied by the cochlear amplifier decreases and hence the excitation needed at threshold increases. To model this effect, an additional term G is introduced in (3.11), i.e.,

$$SP(k) = c((GEP(k) + A(k))^{\alpha} - A(k)^{\alpha}), \text{ for } k = 1, \cdots, D \tag{3.12}$$

where G represents the low-level gain of the cochlear amplifier at a given frequency.

The following important differences from previous loudness estimation algorithms [61, 64] are worth mentioning. The loudness of any signal predicted by (3.12) is never zero, even for signals below threshold. Hence subthreshold amount of loudness may add up across the entire frequency range and render a broadband sound audible, which corresponds well with the physical phenomenon observed.

However, the rate of decrease of specific loudness is higher for sounds below absolute thresholds than predicted by (3.12). To account for this, an additional factor is introduced as shown in (3.13) so that the rate of decrease is consistent with the observed psychophysical measurements for sub threshold signals, $EP(k) < E_{THRQ}(k)$.

$$SP(k) = C\left(\frac{2EP(k)}{2EP(k) + 2E_{THRQ}(k)}\right)^{1.5} [(GEP(k){+}A(k))^{\alpha}{-}A^{\alpha}],\ EP(k) < E_{THRQ}(k) \tag{3.13}$$

Similarly, the rate of increase of specific loudness for sounds above 100 dB is higher than that predicted by (3.12). At high intensity levels, the following expression is used for specific loudness calculation:

$$SP(k) = C\left(\frac{EP(k)}{1.04 \times 10^{6}}\right)^{0.5} \tag{3.14}$$

where the constant 1.04×10^{6} is used to make the specific loudness function continuous at $EP(k) = 10^{10}$.

Therefore, depending on the excitation level observed at a particular detector d_j, one of equations (3.12)-(3.14) should be used.

Total loudness computation

Finally, the area under the specific loudness pattern $SP(k)$ is calculated to obtain the total instantaneous loudness L. This represents the monaural loudness. The binaural loudness is obtained by summing the specific loudness pattern associated with each ear. If the same sound is presented to both ears, then the binaural loudness is just twice the monaural loudness.

Short-term and Long-term Loudness

The previous section described a model for estimating the loudness associated with steady sounds. However, real-life signals are time-varying in nature and do not exhibit a steady-sound behavior. Therefore, they exhibit temporal masking in addition to the simultaneous masking that is observed in the case of steady state sounds. Next, we present an overview of the time-varying loudness estimation algorithm.

In [4], a model of loudness estimation for time-varying sounds was built on top of the steady sounds model to account for the temporal masking phenomenon. The first step involved estimating power spectral density on a continual basis to capture the variations over time. This was accomplished using six parallel FFTs (fast Fourier transform) on hanning windowed segments of 2, 4, 8, 16, 32, and 64 ms duration signals. Select frequency components are extracted from each FFT spectrum in order to obtain the best tradeoff between time and frequency resolution. This spectrum was updated at the rate of 1 ms (i.e., a frame hop size of 1 ms is used).

The subsequent stages of outer and middle ear filtering, excitation pattern evaluation, loudness pattern evaluation and instantaneous loudness evaluation are similar to that described in the steady state model proposed in [3]. Following this, a short-term loudness and a long-term loudness are calculated based on attack and release parameters.

For speech signals, the short-term loudness can be thought of as the loudness impression created by a specific syllable or word. The long-term loudness on the other hand corresponds to the overall loudness created by the entire sentence.

As the name implies, long-term loudness has a higher memory effect compared to that of the short-term loudness. If $L(j)$ denotes the instantaneous loudness in the j^{th} frame, then the short-term loudness at the j^{th} frame segment is given by:

$$SL_j = \begin{cases} \alpha_t L(j) + (1-\alpha_t)SL(j-1), & L(j) > SL(j-1) \\ \alpha_r L(j) + (1-\alpha_r)SL(j-1), & L(j) \leq SL(j-1) \end{cases} \tag{3.15}$$

where $SL(j-1)$ is the short-term loudness in the $j-1^{th}$ frame, α_t and α_r denote the attack and release parameters respectively and can be calculated as follows:

$$\alpha_t = 1 - e^{-T_i/T_a}, \tag{3.16}$$

$$\alpha_r = 1 - e^{-T_i/T_r}. \tag{3.17}$$

In (3.16)-(3.17), T_i denotes the time interval between successive frame segments, T_a and T_r are the attack and release time constants. In [4], the values of T_a and T_r were chosen to be 0.045 and 0.02 respectively for a hop duration of $T_i = 1\ ms$. The long-term loudness is computed in a similar manner with longer attack and release time constants to model the long-term memory effect.

3.6 Complexity analysis of the Moore and Glasberg algorithm

As described in the previous section, the input signal $x(n)$ is first referenced to an assumed sound pressure level of P dB. The specifications of the input power spectral components $\{S_x(\omega_i)\}_{i=1}^{N}$ are obtained either directly or through a spectral analysis stage, where $\omega_i = e^{\frac{j2\pi f_i}{f_s}}$ and f_s denotes the sampling frequency.

1. If $|M(\omega_i)|$ denotes the frequency response of the outer/middle ear filter, then the effective power spectrum reaching the inner ear is $S_x^c(\omega_i) = |M(\omega_i)|^2 S_x(\omega_i)$. This stage has an $O(N)$ computational complexity (N additions), where N represents the number of spectral components.

2. The next stage involves evaluation of the excitation pattern $\{EP(k)\}_{k=1}^{D}$ associated with the sound reaching the inner ear. The excitation pattern $EP(k)$ at any detector d_k is calculated as the sum of the response from the different auditory filters according to (3.5) [52]. This stage is associated with an $O(ND)$ computational complexity.

3. The slopes of the auditory filters, $\{p_k\}_{k=1}^{D}$, in (3.5) have to be evaluated for each pattern, since they change as a function of the center frequency and the total intensity level $\{I(k)\}_{k=1}^{D}$ [52,54]. This has an $O(D)$ computational complexity.

4. The Intensity pattern, $\{I(k)\}_{k=1}^{D}$, calculates the total power within one ERB unit surrounding the detector d_k. This process is associated with an $O(D)$ computational complexity.

5. Next, for each auditory filter, the magnitudes, $W(k,i) = (1+p_k g_{k,i}) \exp(-p_k g_{k,i})$, have to be evaluated for all N frequency components. This operation is associated with an $O(ND)$ complexity.

6. The excitation pattern $EP(k)$ is transformed to a specific loudness pattern $SP(k)$ according to the procedure described in [3]. Therefore with D detectors this stage has an $O(D)$ complexity.

7. The final stage involves calculation of the area under the specific loudness pattern $SP(k)$ in order to obtain the total instantaneous loudness L. This stage is associated with an $O(D)$ complexity.

It can be observed that the excitation pattern and auditory filter evaluation stages are associated with the highest complexity, i.e., $O(ND)$. In the next section, a frequency and detector pruning approach is proposed that implements the stages of the auditory model in a computationally efficient manner.

3.7 Summary

In this chapter, an overview of two different classes of auditory models used for the purpose of loudness estimation were described. The simple models were based on the shape of equal loudness contours whereas the more elaborate auditory models made use of highly overlapping bank of bandpass auditory filters. The shapes and bandwidths of these auditory filters accounted for the frequency selectivity property of the auditory system. Following this, a detailed description of the Moore & Glasberg auditory model and an analysis of the model's computational complexity was described.

Chapter 4

A Frequency/Detector Pruning approach for auditory models

4.1 Introduction

In the previous chapter, it was observed that the auditory filter evaluation and excitation pattern evaluation stages are associated with the highest computational complexity. The computational complexity of both these stages are dependent on the number of frequency components N and the number of detector locations D. In this chapter, a computationally efficient alternative to evaluate the auditory model stages is described. The proposed algorithm is based on a frequency pruning and detector pruning approach that obtains fast estimates of the excitation pattern, the loudness pattern and the total loudness quantities.

The objective behind the proposed frequency and detector pruning approach is to prune the number of frequency components N and the number of detector locations D in a manner consistent with human perception. It now remains to decide what frequency components $f_i's$, where $i \in \{1, 2, ...N\}$ and detector locations $d_k's$ where $k \in \{1, 2, ...D\}$ to choose in order to evaluate the model stages.

The frequency pruning approach approximates the spectrum with a few spectral components such that the total neural activity is preserved. The approximation is carried out in a perceptually relevant manner by exploiting the principles of psychoacoustics. The detector pruning algorithm selects the detectors in a nonuniform manner such that the general shape of the excitation or the loudness pattern is captured. The performance of the proposed algorithm is compared to the Moore and Glasberg process. Simulation results indicate that the differences in loudness estimates are minimal when tested on a representative audio corpus from the SQAM database [69]. Additionally, the corresponding high resolution patterns can be obtained by linearly interpolating the low resolution

EP.

The proposed frequency and detector pruning approach can be embedded into the original Moore & Glasberg auditory model without changes to the model parameters. The new model operates at a much lower computational complexity and is aimed at solving perceptual objective functions in a computationally efficient manner. In particular, it reduces the computations involved in repeatedly employing the auditory model stages to test candidate solutions when solving perceptual objective functions. In this chapter, we describe the proposed low-complexity loudness estimation algorithm [38, 43] applicable to both steady and time-varying sounds.

The block diagram of the proposed model is shown in Figure 4.1.

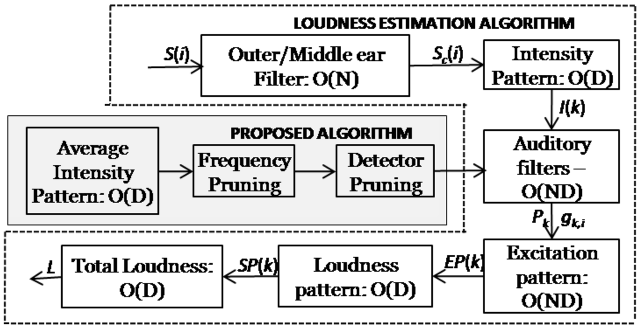

Figure 4.1: Reference and proposed loudness estimation scheme.

4.2 Frequency Pruning: Problem Statement

The objective of the frequency pruning algorithm is to reduce the number of frequency components (N) in a manner consistent with human perception, i.e., decide what frequency components f_i's to choose, where $i \epsilon \{1, \cdots, N\}$, such that the excitation pattern, loudness pattern and loudness estimates are preserved. Firstly, it is known from the masking phenomenon [18] that the masked components are inaudible and hence do not contribute towards a loudness perception. Therefore, only the limited set of perceptually relevant unmasked components that contribute towards a loudness perception need be considered. However, determining these unmasked components requires computationally demanding algorithms that estimate masked thresholds. Secondly, it is known that white noise bands, falling within the same critical band, will have the same instantaneous loudness as any individual component with their combined sum of intensities [62], i.e., the loudness depends only on the total neural activity evoked and not on the intensity distribution of the frequency components. Although this property can be exploited to approximate the spectrum with fewer components, in practice, the spectrum of an audio segment has a complex structure and does not have perfect white noise bands. In such cases, spectrum approximation not only distorts the actual shape of the excitation and specific loudness patterns but also the final loudness estimates.

4.3 Detector Pruning: Problem Statement

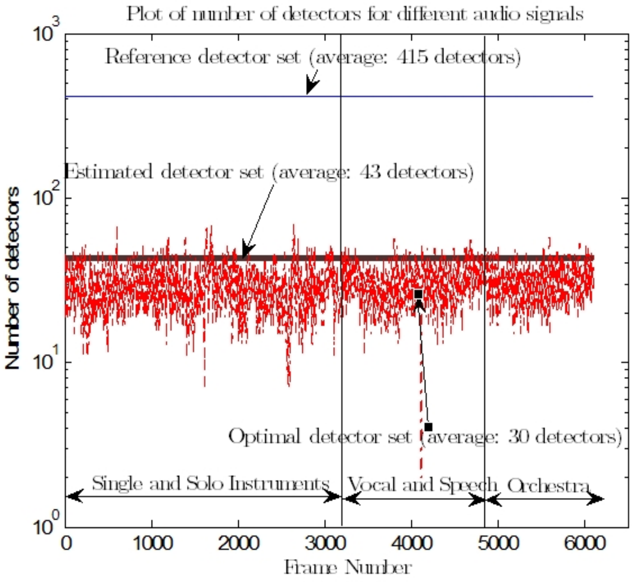

Figure 4.2: Plot showing cardinality of optimal detector set L_o compared with reference detector set L_r, and estimated detector set, L_e.

Unlike most existing methods for generating excitation patterns that place detectors uniformly along the basilar membrane, the objective behind the proposed detector pruning approach is to non-uniformly sample the excitation pattern at a sufficient number of points in order to capture its general shape. This is motivated by the following two analysis [43]: Firstly, a fast Fourier transform of the reference excitation pattern corresponding to a spectrally complex music signal (a worst case scenario) shows that 99% of energy is concentrated in the first 10% of the spectrum, indicating that the excitation pattern is slowly varying and can be sampled accordingly. Let $L_r = \{d_k| \, |d_k - d_{k-1}| = 0.1, k = 1, 2, \ldots D\}$ denote the reference set of detector locations expressed in ERB units, such that they are uniformly spaced at 0.1 ERB units. Let $L_o = \{d_k | \partial EP(k)/\partial k = 0, k = 1, \cdots, D\}$ denote the "optimal" set of detector locations such that they correspond to the extrema of EP. Secondly, a search for the set L_o, carried out on the reference excitation pattern for different types of audio indicates that the cardinality of the set L_o is of the order O (number of ERB units) that is spanned by the input audio spectrum [43]. In Figure 4.2, we plot the cardinality of the reference set of detectors (L_r), the optimal set of detectors (L_o), and the estimated set of detectors (L_e). Comparing the reference set with the optimal set shows that the excitation pattern can be generated using significantly fewer detectors. Therefore, it is sufficient to evaluate the EP at its maxima and minima to capture its shape. Since, the EP is unavailable to us and therefore the set L_o of its maxima and minima, the problem now reduces to obtaining an estimate of L_o.

4.4 Proposed Algorithm 1: Implementation and Results Estimating pruned frequency components

The proposed frequency pruning algorithm exploits the nature of the intensity pattern to prune the frequency components in a computationally efficient manner. The intensity pattern, $I(k)$, is subject to a simple averaging operation. The difference equation representing the filtering operation is given by

$$Y(k) = \frac{1}{11} \sum_{m=-5}^{5} I(k-m) \text{ for } k = 1, \cdots, D \tag{4.1}$$

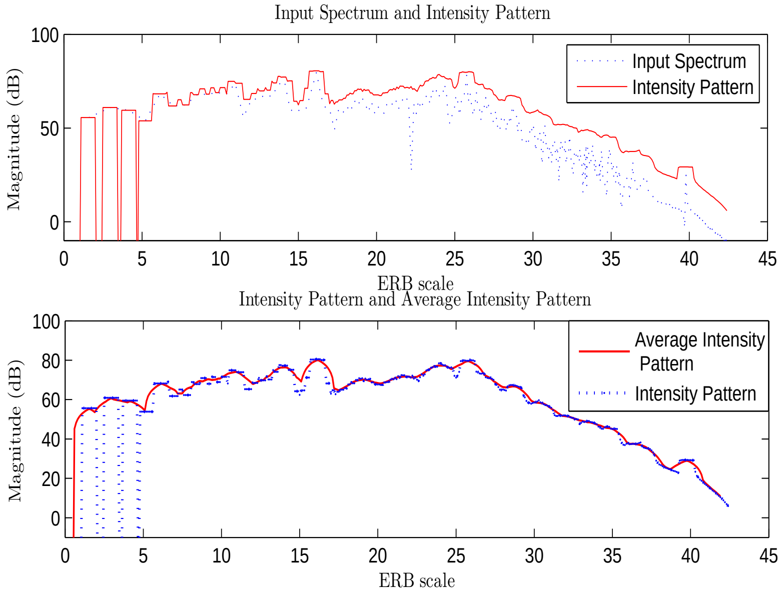

Figure 4.3: Top: Plot of corrected spectrum, $S_x^c(i)$ and intensity pattern,$I(k)$. Bottom: Plot of average intensity pattern, $Y(k)$, and intensity pattern, $I(k)$.

where $Y(k)$ represents the “average intensity pattern" (i.e, average intensity per ERB) surrounding the detector d_k. We further note that the filtering operation in (4.1) can be realized in a computationally efficient manner with fewer additions by realizing the filter's transfer function, $H(z)$, as,

$$H(z) = \frac{1}{11}\frac{z^5 - z^{-5}}{1 - z^{-1}}. \tag{4.2}$$

Tonal bands

From Fig. 4.3, we can observe that the intensity pattern $I(k)$ remains approximately flat in certain critical bands. Let R_j denote the j^{th} continuous subset of detectors over which the intensity pattern is observed approximately constant. Since, $I(k)$ is obtained from (3.6) as a sum of components, the flat structure of $I(k)$ implies that this sum remains approximately constant for all $k\epsilon R_j$ which is possible only when a strong component is present in the midst of much weaker components. This indicates the strong tonal nature of the critical band.

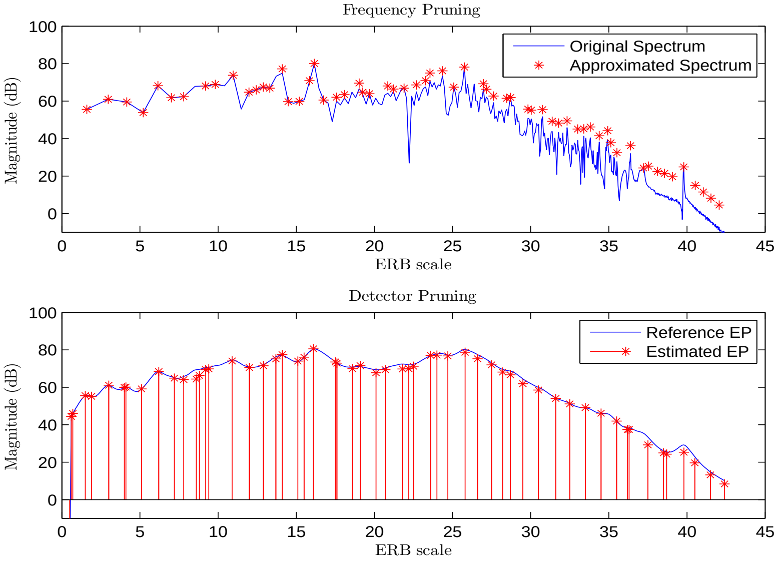

Figure 4.4: Top: Frequency pruning, Bottom: Detector pruning.

As a consequence, the average intensity pattern $Y(k)$ exhibits a peak corresponding to the tonal component in that critical band. This behavior is shown in Fig. 4.3 where the corrected input spectrum, $S_x^c(i)$, the intensity pattern, $I(k)$, and the average intensity pattern, $Y(k)$, are plotted. Therefore, peaks in $Y(k)$ can be used to identify tonal bands and the tonal components within. Furthermore, the average intensity pattern detects only the peaks corresponding to the tonal components and filters out the other spurious peaks from $S_x^c(i)$ that correspond to the noise-like bands. Thus, $Y(k)$ is a more suitable pattern than $S_x^c(i)$ to detect tonal components.

In such tonal bands, it can be assumed that the strongest spectral component will mask the neighboring weaker components. Hence, the pruned set of frequencies is obtained by selecting only the strongest spectral component and ignoring the other masked components.

Noise bands

In the other noise-like critical bands, estimating masked components is not straightforward. Therefore, frequency pruning is accomplished by further dividing each noise-like critical band into smaller sub-bands, $B_{1:Q}$, where Q denotes the number of smaller sub-bands. Here, each sub-band B_p is assumed to be approximately white. Each of these smaller sub-bands is now approximated with a single component, $\hat{S}_p$, with intensity equal to the combined sum of the intensities of all the components within that sub-band. Let M_p be the set containing the indices of the components in sub-band B_p. $\hat{S}_p$ is given by

$$\hat{S}_p = \sum_{j \epsilon M_p} S_x^c(j), \text{ for } 1 \leq p \leq Q. \tag{4.3}$$

We note that this process is consistent with the reference loudness estimation algorithm since it preserves the total intensity within any critical band thereby also preserving the auditory filter shapes. In Fig. 4.4, the input spectrum and the frequency pruned spectrum are shown.

Estimating pruned detector locations

We now describe the procedure to estimate the pruned set of detectors L_e. Due to the similar processes involved in the evaluation of EP and average intensity pattern, we make use of the average intensity pattern to estimate the set L_e. That is, the rounded exponential auditory filters transform the input spectrum to an EP defined along the ERB scale; similarly, the average intensity pattern can be thought of as a filtered version (with rectangular filter responses) of the intensity pattern. Hence, the maxima and minima associated with $Y(k)$, i.e., $L_e = \{d_k | \partial Y(k)/\partial k = 0, k = 1, \cdots, D\}$ can be used to determine the pruned set of detectors, i.e., L_e. We then estimate the EP at the detector locations specified by L_r by linearly interpolating the EP obtained at the points specified by L_e. In

Fig. 4.4, the reference EP (evaluated at L_r) and the estimated EP (evaluated at L_e) are plotted.

Properties of $Y(k)$

We note that the average intensity pattern $Y(k)$ is associated with a number of desirable properties: a) it provides a simple procedure to identify tonal bands in the input spectrum and thereby perform frequency pruning, b) further, due to the similar processes observed between excitation pattern evaluation and $Y(k)$ evaluation, it provides an elegant method for detector pruning, c) it can be obtained in a computationally efficient manner as described in (4.2) thereby keeping the overhead associated with the frequency/detector pruning approach minimal.

Results and Discussion

Experimental Setup

For the simulation, different types of audio provided in the Sound Quality Assessment Material (SQAM) database [69] were utilized. The audio signals are sampled at 44.1 KHz and audio segments of 23 ms durations were used for the simulations. Furthermore, each audio segment was referenced to an assumed Sound Pressure Level (SPL) between 30 and 90 dB randomly to evaluate the loudness estimation algorithm at all possible sound levels. Spectral analysis is done using a 1024 point FFT (i.e., $N = 513$). The reference set L_r of $D = 420$ detectors are uniformly spaced on the ERB scale. The experiments are performed on a 2 GHz Intel Core 2 duo processor with 2 GB RAM.

Frequency and Detector Pruning

Let N_r and D_r denote the average number of pruned frequency components and detectors respectively. The performance of the frequency and detector pruning approach is measured in terms of the percentage reduction in the number of frequency components and detectors, i.e., $(N-N_r)/N$ and $(D-D_r)/D$. The results

Table 4.1: Frequency and Detector Pruning Evaluation Results for $Q = 2$.

Type	Number of Components			Percent Reduction
	Maximum	Minimum	Average	
Frequency Pruning	66	56	$N_r = 63$	88%
Detector Pruning	102	81	$D_r = 87$	80%

are tabulated in Table 4.1. An average reduction of 88% and 80% is obtained for the frequency and detector pruning approaches respectively. This results in an average reduction of 97% ($= 1 - \frac{N r D_r}{ND}$) for the excitation and auditory filter evaluation stages, which have an $O(ND)$ complexity. In Table 4.2, a comparison of computational (CPU) time is shown, where the proposed approach achieves a 95% reduction in computational time for the excitation & auditory filter stages.

Loudness estimation

The absolute loudness error ($|L_r - L_e|$), and the relative loudness error ($|L_r - L_e|/L_r$) metrics are used to evaluate the performance of the proposed loudness estimation algorithm, where L_r, L_e represent the reference and estimated loudness (in sones) respectively. The results are tabulated in Table 4.3 for different types of audio signals[1]. It can be observed that the proposed frequency/detector pruning approach yields a very low average relative loudness error of about 5%.

Table 4.2: Computational Time: Comparison Results.

Stage	Computational Time (in seconds)		Reduction
	Reference	Proposed	
Auditory Filter & Excitation Pattern	0.407	0.01942	95%
Loudness Pattern	0.00128	0.00064	50%

4.5 Proposed Algorithm 2: Implementation and Results

An alternative approach to frequency and detector pruning is described which is computationally more efficient than the pruning approach described earlier.

[1] Synthetic signals were also tested and similar results as reported in Table 4.3 were obtained.

Table 4.3: Loudness estimation algorithm: Evaluation Results.

Type	Loudness Error $\|L_r - L_e\|$(in sones)			Relative Error
	Maximum	Minimum	Average	
Single Instruments	2.6	0.002	0.40	4.63%
Speech & Vocal	2.42	0.00312	0.41	3.80%
Orchestra	2.49	0.00662	0.42	5.18%
Pop Music	2.59	0.00063	0.45	4.25%
Band-limited Noise	4.4	0.09	1.02	7%

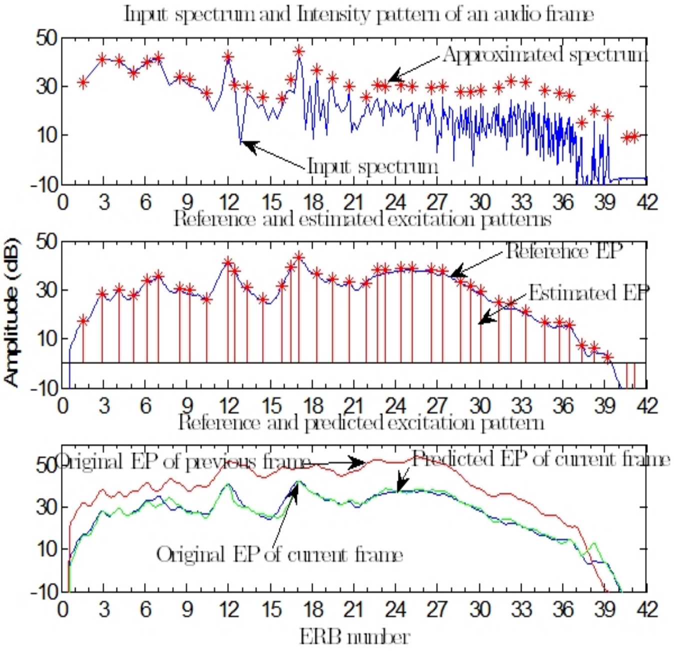

Figure 4.5: Top: Plot of input and approximated spectrum, Middle: Plot of reference and estimated EP, Bottom: Plot of reference and predicted EP.

However, the approximations involved in this approach make it less accurate in preserving the shape of the auditory patterns compared to that of the previous approach.

Frequency component pruning

It is known that multiple components with equal intensity falling inside the same critical band will have the same instantaneous loudness as any individual component with their combined sum of intensities [62]. This enables us to approximate the input audio spectrum inside each ERB unit (critical band) with a single component of intensity equal to the combined sum of intensities within that ERB unit as shown in (4.4).

$$S_a(m) = \sum_{i \in (m,m+1]} S_x^c(i) \tag{4.4}$$

where $S_x^c(i)$ is the input spectral amplitude after outer/middle ear correction, i represents the set of components in the m^{th} ERB unit and $S_a(m)$ is the approximated spectrum in the m^{th} ERB. In Fig. 4.5(a), an example of a sample audio spectrum and the approximated spectrum $S_a(m)$ are plotted on an ERB scale. Although, approximating the frequency spectrum preserves the final loudness estimates it does however distort the shape of the intermediate quantities (i.e., the excitation/loudness patterns) as these patterns depend on the intensity distribution of the spectral components inside each critical band (one ERB unit). In order to minimize the error in the shape of the estimated excitation/loudness pattern, a modification in the locations of the approximated spectral components $S_a(m)$ is proposed. In addition, it is necessary to estimate the locations of the detectors that capture the general shape of the excitation/loudness patterns (i.e., their maxima and minima positions).

Estimating pruned frequency and detector locations

Here, we describe a procedure that estimates the positions of the approximated spectral components $S_a(m)$ that best capture the structure of the excitation/loudness patterns. For any component, $S_x^c(i)$, the maximum response due to $S_x^c(i)$ will occur at a detector location for which $|g_{k,i}| \approx 0$, i.e., for which $\exp(-p_k.g_{k,i}) \approx 1$ in (5.2). That is, the auditory filter that is centered at d_k for which $g_{k,i} \approx 0$ will result in the maximum response from $S_x^c(i)$. However, in a generic spectrum, when multiple components are closely spaced, it is not straightforward to identify the detector with the maximum response as the relative magnitudes of the neighboring components can have an influence on which detector with result in the maximum response. In other words, the summation in (5.2) can show a maximum at any detector location d_k which is not necessarily close to a specific $S_x^c(i)$.

However, the following two properties aids in capturing at least the maxima of the excitation/loudness patterns: i) the specific form of the auditory filter shapes (with exponential fall off on both sides) ensures that the responses due to a frequency component is negligibly small in neighboring critical bands, ii) for the approximated spectrum $S_a(m)$, the frequency components are placed far apart (i.e., one in each critical band) and therefore each component's response attains locally maximum value within the critical band with negligible influence from components in neighboring critical bands.

Therefore, it is safe to conclude that the detectors that are close to the location of the approximated component $S_a(m)$ will attain a maximum response within that critical band. For the approximated spectrum $S_a(m)$, the response at a particular detector d_k is then given by

$$\hat{EP}(k) = \sum_{i=1}^{N} (1 + p_k g_{k,m}) exp(-p_k g_{k,m}) S_a(m) \tag{4.5}$$

Furthermore, among all the locations inside the critical band to place the approximated component $S_a(m)$, the most likely location to select is that of the maximum $S_x^c(i)$ component in the spectrum so that the peaks in the estimated excitation/loudness pattern are close to the actual peaks. Therefore, detector pruning is accomplished by directly mapping this set of frequency component locations (obtained from the maximum components inside each critical band) to a set of detectors such that they capture the general shape of reference excitation pattern directly (without having to compute it). In Fig. 4.5(b), we plot the reference excitation pattern and the estimated excitation pattern along with the positions of maximal auditory filter response.

Simulation Results

In this section, the experimental setup is described and evaluation results are provided. The performance of the proposed algorithm was tested with different types of audio provided in the Sound Quality Assessment Material (SQAM) database. The audio signals are sampled at 44.1 KHz and audio segments of 46 ms durations were used for the simulations. In real-life, sound levels can change abruptly across time. Therefore, each audio segment was referenced to an assumed Sound Pressure Level (SPL) between 30 and 90 dB randomly to account for these abrupt changes. We evaluate the performance of the proposed algorithm in terms of the Relative Error Energy (REE) and Average Error Energy (AEE) as defined in (4.6) for the excitation pattern which is indicative of the relative error at each detector location d_k and average error across all detector locations, i.e.,

$$REE = 20\log_{10}\left\{\sum_{k\in 1,2,\ldots D}\left|\frac{\hat{EP}(k) - EP(k)}{EP(k)}\right|\right\}. \tag{4.6}$$

Table 4.4: Computational requirements in various stages of the model for the standard and proposed algorithm.

Stages	Complexity Comparison		Complexity
	Original	Proposed	Reduction (S-P)/S
Auditory Filters: O(ND)	90962	1186	98%
Excitation pattern: O(ND)	90962	1186	98%
Specific Loudness-O(D)	415	43	89%
Total Loudness: O(D)	415	43	89%

The error in the estimated loudness is evaluated in terms of the Average Loudness Error (ALE) and the Maximum Loudness Error (MLE) which are defined in (4.7) and (4.8) respectively, i.e.,

$$ALE = \frac{1}{P}\sum_{j=1}^{P}|\hat{L}_j - L_j|, \tag{4.7}$$

$$MLE = max(|\hat{L}_j - L_j|),\ j \in 1, 2, ...P \tag{4.8}$$

where $\hat{EP}(k)$ and $EP(k)$ are the estimated and reference EP expressed in linear power units. $\hat{L}_j$, L_j are the estimated and reference instantaneous loudness. P represents the number of audio frames. In Table 1, we compute REE, AEE, ALE and MLE metrics for different types of audio material. The REE and AEE of the estimated excitation pattern are roughly about -12.5 dB and -15 dB respectively. The error on loudness measured using the ALE and MLE metrics are 0.6 sones and 2.6 sones on average across different audio signals. It can also be observed from Table 4.5 that the proposed algorithm performs consistently for different types of audio signals within a tolerable error.

Table 4.5: Loudness estimation algorithm: Evaluation Results.

Different types of audio	AEE (dB)	REE (dB)	MLE (sones)	ALE (sones)
Single Instruments	-12.82	-14.84	0.72	3.26
Speech	-12.80	-14.73	0.29	2.82
Vocal	-12.03	-14.55	0.22	2.60
Solo Instruments	-12.42	-14.60	0.44	2.25
Vocal & Orchestra	-13.4	-18.57	0.95	3.26
Orchestra	-11.52	-14.92	1.34	2.82
Pop Music	-12.58	-14.90	0.27	2.60
Average	-12.5	-15	0.6	2.6

Furthermore, we compare the computational complexity of the proposed algorithm with the standard approach followed in [3, 4, 52, 70]. We also highlight the complexity of each stage separately due to the differing nature of operations in each stage. From Table 4.4, it can be seen that the proposed algorithm achieves a significant reduction in complexity close to 96% on average.

4.6 Time-varying low-complexity algorithm

In real life, one typically encounters time-varying sounds such as speech or music. Therefore it is possible to exploit the time-varying nature in developing a computationally efficient loudness estimation algorithm. The auditory model requires the auditory filter shapes and the excitation/loudness pattern to be computed for every audio segment which may not be suitable for many real-time applications. In this section, we describe the proposed low-complexity algorithm for time-varying sounds. We begin by exploiting the intensity pattern in the current and preceding frames. We define the intensity pattern $I_p(m)$ as the total equivalent intensity in the m^{th} ERB as shown in (4.4). A differential intensity pattern $DI_p(m)$ is computed according to:

$$DI_p(m) = I_p(m) - I_{p-1}(m) \tag{4.9}$$

where m represents the ERB number and p is the frame index. Since the auditory filters change their shapes with frequency and intensity level [52], they have to be re-computed in every frame according to the intensity pattern $I_p(m)$ associated with the current frame. However, we exploit the differences in the intensity pattern in the current and previous frames and partially evaluate the auditory filters shapes and the corresponding EP $\hat{E}_p(k)$ in select ERBs where (4.10) is satisfied.

$$DI_p(m) > \tau_m \tag{4.10}$$

where τ_m is the threshold in dB in the m^{th} critical band. Following this an excitation prediction step estimates differential intensity $DI_p(m)$ at the detector locations d_k where the EP is computed, by linear interpolation. The final EP of the current frame $E_p(k)$ is predicted from the EP of preceding frame $E_{p-1}(k)$ according to:

$$E_p(k) = \begin{cases} E_{p-1}(k), & DI_p(m) < \tau_m \\ \hat{E}_p(k), & otherwise \end{cases} \tag{4.11}$$

wherein $E_p(k)$ is obtained either from the scaled EP in critical bands where the differential intensity pattern doesn't exceed the threshold and from partially evaluated EP in the other critical bands. In Fig. 4.5(c), we show a plot of the original EP of the current and preceding frame and the predicted EP of the current frame. It can be seen that the predicted EP closely follows the original EP of the current frame. The subsequent stages in the model are similar to the steady sound algorithm as illustrated in Fig. 4.1.

4.7 Summary

In this chapter, we described an efficient frequency pruning and a detector pruning algorithm to obtain estimates of excitation patterns, loudness patterns and the total loudness quantities. Our experiments indicate that the proposed frequency and detector pruning approach can achieve up to an 80% and 88% average reduction in the number of spectral components and detector locations respectively. The combined frequency/detector pruning performance results in 97% reduction in the computational complexity of the auditory filter evaluation and excitation pattern stages of an auditory model. Experimental results also indicate that the loudness estimates obtained with the proposed technique are associated with only a $4 - 7\%$ average relative loudness error.

For time-varying signals, we described a prediction algorithm that estimates the excitation pattern of the current frame from that of the preceding frame. The excitation pattern is partially evaluated in select critical bands where the auditory filter shapes exhibit significant changes. In the other critical bands, the excitation pattern of the previous frame are scaled appropriately to obtain the excitation pattern corresponding to the current frame. This resulted in additional computational savings for time-varying signals such as speech/music.

Chapter 5

Auditory Speech Enhancement

5.1 Introduction

Speech enhancement remains an open research problem for the past several decades. Several algorithms [71] have been proposed in the literature to address the two main issues faced by speech enhancement systems, i.e., improving the quality and intelligibility of degraded speech. Most of the algorithms employ techniques that roughly fall into one of the following frameworks: i) spectral subtractive type techniques [72], ii) statistical model based techniques [28], and iii) subspace based techniques [30]. In all of the above algorithms, the enhancement is usually carried out either using a time/frequency domain representation or in a suitable subspace. Such signal representations do not consider the mechanism utilized by the human auditory system.

Nevertheless, the perceptual effects of speech/noise have been studied and several strategies incorporating perceptual constraints have been proposed to partially account/model the properties of the human auditory system. For example, perceptual weighting filters, derived from the LP analysis of speech segments, are used to weight the residual noise so as to “hide" the noise in high energy spectral regions (i.e., formant peaks) and aggressively suppress noise near spectral valleys [29]. Similarly, masking thresholds have been employed to adapt the parameters in a spectral subtractive type algorithm [27]. Similar strategies for incorporating perceptual constraints have also been considered in statistical model based techniques [28] and subspace based techniques [30]. In another interesting work, a more elaborate auditory model has been used to detect the speech dynamics (in particular the non-stationary segments such as transients, plosive bursts, changing formants) and adapt the time-varying wiener filter [73]. Although most of these algorithms employ a tractable error criterion, the incorporation of perceptual constraints is usually done heuristically.

From a perceptual point of view, the enhancement task should be carried out by explicitly modeling the human auditory system. This involves obtaining auditory representations such as excitation patterns or loudness patterns corresponding to an acoustic signal. However, this approach has not been popular for the following reasons. Firstly, this requires reconstructing the acoustic signal back from its auditory representation, i.e, it involves an inverse mapping procedure. Therefore, auditory models have generally been used only in analysis frameworks and not in an analysis/synthesis framework. Alternatively, the perceptual objective function can be minimized by carrying out an exhaustive search over all possible candidate solutions in the time/frequency domain. However, this process is associated with a high computational complexity making them unsuitable for real-time applications. Recently, efficient techniques have been proposed [38] to reduce the computational complexity associated with auditory model implementations. Further, we note that algorithms that reconstruct an acoustic signal from its auditory representation have been proposed for certain auditory modeling frameworks [39]. The above considerations form a primary motivation to explore an auditory domain based speech enhancement system.

In this chapter, we describe a new approach for speech enhancement that employs auditory representations. We propose a speech enhancement technique that directly minimizes a perceptual error metric. In other words, the proposed technique finds an estimator that minimizes the error between the auditory representation associated with the enhanced speech and that associated with the desired speech. This approach is different from the existing approaches wherein the error criterion usually only involves some measure of the perceptual behavior (either in terms of thresholds or spectral weights) and does not the explicitly include the actual auditory perception. We describe a constrained optimization

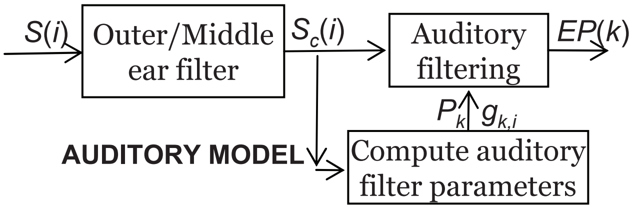

Figure 5.1: Block diagram of the auditory model.

framework to carry out the above task. Simulation results indicate that the proposed algorithm attains a lower relative loudness error compared to the Wiener or spectral subtraction technique and attains better performance at low signal-to-noise ratios.

Description of the Auditory model

In this section, a brief description of the steps involved with the Moore and Glasberg auditory model [3] is provided. A block diagram of the model is shown in Fig. 5.1.

The input signal $x(n)$ is referenced to an assumed sound pressure level (SPL) of P dB. Let $S_x(\omega_k)$ denote the power spectrum of $x(n)$ where $\omega_k = e^{\frac{j2\pi f_k}{f_s}}$ and f_s denotes the sampling frequency. Next, an outer and middle ear correction is applied so that the effective power spectrum reaching the inner ear is $S_x^c(\omega_k) = |M(\omega_k)|^2 S_x(\omega_k)$ where $|M(\omega_k)|$ denotes the frequency response of the outer/middle ear filter.

The excitation pattern associated with the sound reaching the inner ear is calculated next. The frequency scale is first transformed to an auditory scale that is measured using an equivalent rectangular bandwidth (ERB) number and is calculated using,

$$p \text{ (in ERB units)} = 21.4 \log_{10}(4.37f/1000 + 1) \tag{5.1}$$

where p represents the ERB number and f denotes frequency in Hz. A set of D detectors, $\{d_j\}_{j=1}^{D}$ are placed uniformly at 0.1 ERB units along the auditory scale. Each detector d_j represents the center of the auditory filters employed. Let $\{cf_j\}_{j=1}^{D}$ denote the center frequencies (in Hz) corresponding to the center $\{d_j\}_{j=1}^{D}$ of the auditory filters. The excitation pattern, $EP(j)$, is now evaluated as the output of these auditory filters to the effective spectrum reaching the inner ear, i.e., $S_x^c(\omega_k)$ and is given by,

$$EP(j) = \sum_{k=1}^{N} (1 + p_j g_{j,k}) \exp(-p_j g_{j,k}) S_x^c(\omega_k), \text{ for } 1 \le j \le D, \tag{5.2}$$

where p_j denotes the slope of the auditory filter centered at detector d_j and $g_{j,k} = |(f_k - cf_j)/cf_j|$, denotes the normalized deviation of f_k from cf_j. The slopes, $\{p_j\}_{j=1}^{D}$, have to be evaluated since they change as a function of the effective spectrum $S_x^c(\omega_k)$ reaching the inner ear. Further details of auditory filter shape evaluation can be found in [3].

For non-stationary signals such as speech, a short-time Fourier transform, $|X^l(\omega_k)|$, is obtained on a frame-by-frame basis and the power spectrum is approximated as $\hat{S}_x^l(\omega_k) = |X^l(\omega_k)|^2$, where l denotes the frame index and steps described above for the auditory model are carried out.

It should be noted that although (5.2) appears linear in $S_x^c(\omega_k)$, the values p_j change as a function of $S_x^c(\omega_k)$. The auditory filter shapes become shallower for higher intensity levels. In order to linearize the equation, we make the assumption that the auditory filters are not level-dependent, thereby removing the dependence of p_j's on $S_x^c(\omega_k)$.

5.2 Proposed speech enhancement algorithm

In this section, the idea behind the proposed speech enhancement algorithm based on auditory modeling is described.

System Model

Let $y(n)$ denote the noisy signal such that $y(n) = x(n) + d(n)$ where $x(n)$ is the desired signal and $d(n)$ represents uncorrelated additive noise. Due to the non-stationary nature of the speech signals, the signals are processed on a frame-by-frame basis. Therefore, a short-time fourier transform of the noisy speech is computed and the additive signal model can be equivalently expressed in the frequency domain as,

$$Y^l(\omega_k) = X^l(\omega_k) + D^l(\omega_k) \text{ for } k = 1, 2, \cdots, N. \tag{5.3}$$

Let $\{\mathbf{Y}, \mathbf{X}, \mathbf{D}\}$ represent $N \times 1$ frequency domain vectors containing the spectral components $\{Y^l(\omega_k), X^l(\omega_k), D^l(\omega_k)\}$ for $k = \{1, 2, \cdots, N\}$ associated with the noisy signal, clean signal and noise signal respectively.

Following the steps of the auditory model, we can evaluate the parameters $\{p_j, g_{j,k}\}$'s associated with auditory filters for $j = \{1, 2, \cdots, D\}$ and $k = \{1, 2, \cdots, N\}$. We can now equivalently express the operation in (5.2) in matrix notations as,

$$\mathbf{E_x} = \mathbf{A}(\mathbf{S^c_x})\mathbf{S^c_x} \tag{5.4}$$

where $\mathbf{A}$ is a $D \times N$ matrix and $\mathbf{A}(.)$ denotes that the matrix is a function of the parameter within the parenthesis. The elements of $\mathbf{A}$ are given by $a_{j,k} = (1 + p_j g_{j,k}) \exp(-p_j g_{j,k})$ and represent the auditory filter magnitudes. $\mathbf{S^c_x}$ represents the power spectrum of the incoming sound after outer/middle ear correction. If we remove the dependance of $\{p_j\}_{j=1}^{D}$'s on $\mathbf{S^c_x}$, then (5.4) can be expressed as,

$$\mathbf{E_x} = \mathbf{A}\mathbf{S^c_x} \tag{5.5}$$

Assuming uncorrelated additive noise, we can equivalently represent the additive signal model in terms of their auditory representation (i.e., excitation patterns) as $\mathbf{E_y} = \mathbf{E_x} + \mathbf{E_d}$. Here, $\mathbf{E_y}, \mathbf{E_x}, \mathbf{E_d}$ represent the $D \times 1$ excitation pattern vectors associated with noisy signal, clean signal and noise signal respectively.

Formulation of Perceptual Error Criterion

The objective is to obtain an estimate, $\hat{x}^l(n)$, of the clean speech signal, $x^l(n)$, such that the error between the excitation patterns of the estimated and clean speech is minimized, i.e., we wish to minimize the following error criterion:

$$C(\hat{\mathbf{E}}_{\mathbf{x}}, \mathbf{E}_{\mathbf{x}}) = ||\hat{\mathbf{E}}_{\mathbf{x}} - \mathbf{E}_{\mathbf{x}}||_2^2 \tag{5.6}$$

Assuming a linear estimator for $\hat{\mathbf{E}}_{\mathbf{x}}$, i.e., $\hat{\mathbf{E}}_{\mathbf{x}} = \mathbf{G}_{\mathbf{E}}\mathbf{E}_{\mathbf{y}}$, we can express the objective function in (5.6) as,

$$\begin{aligned} \mathbf{G}_{\mathbf{E}}^* &= \underset{\mathbf{G}_{\mathbf{E}}}{\operatorname{argmin}} \, ||\mathbf{G}_{\mathbf{E}}\mathbf{E}_{\mathbf{y}} - \mathbf{E}_{\mathbf{x}}||_2^2 \\ &= \underset{\mathbf{G}_{\mathbf{E}}}{\operatorname{argmin}} \, ||\mathbf{G}_{\mathbf{E}}\mathbf{A}\mathbf{S}_{\mathbf{y}}^{\mathbf{c}} - \mathbf{A}\mathbf{S}_{\mathbf{x}}^{\mathbf{c}}||_2^2 \end{aligned} \tag{5.7}$$

where $\mathbf{G}_{\mathbf{E}}$ is $D \times D$ matrix. Although, the estimator obtained in this manner minimizes the error between the auditory representations associated with the two signals $\hat{x}(n)$ and $x(n)$, it only results in an optimal estimator for $\hat{\mathbf{E}}_{\mathbf{x}}$ and not $\hat{\mathbf{S}}_{\mathbf{x}}^{\mathbf{c}}$, i.e., we still need to reconstruct the acoustic signal from its auditory representation before synthesizing the time-waveform. As mentioned in Section (5.1), this reconstruction is not straightforward due to the ill-conditioned nature and low rank of $\mathbf{A}$.

In order to simplify the procedure, we modify the formulation of the error criterion in (5.7) as,

$$\mathbf{G}_{\mathbf{S}}^* = \underset{\mathbf{G}_{\mathbf{S}}}{\operatorname{argmin}} \, ||\mathbf{A}\mathbf{G}_{\mathbf{S}}\mathbf{S}_{\mathbf{y}}^{\mathbf{c}} - \mathbf{A}\mathbf{S}_{\mathbf{x}}^{\mathbf{c}}||_2^2 \tag{5.8}$$

where $\mathbf{G}_{\mathbf{S}}$ is now a $N \times N$ matrix and represents a linear estimator for $\hat{\mathbf{S}}_{\mathbf{x}}^{\mathbf{c}}$, i.e., $\hat{\mathbf{S}}_{\mathbf{x}}^{\mathbf{c}} = \mathbf{G}_{\mathbf{S}}\mathbf{S}_{\mathbf{y}}^{\mathbf{c}}$. We note that (5.8) still minimizes the distortion between the excitation patterns of the associated signals similar to (5.7). The important difference being (5.8) results in an optimal estimator for $\hat{\mathbf{S}}_{\mathbf{x}}^{\mathbf{c}}$ whereas (5.7) results in an optimal estimator for $\hat{\mathbf{E}}_{\mathbf{x}}$.

For simplicity, we assume that $\mathbf{G}_{\mathbf{S}}$ is a diagonal matrix, i.e., the gain is applied individually to each frequency component in $\mathbf{S}_{\mathbf{y}}^{\mathbf{c}}$. Note that the entries of $\mathbf{E}_{\mathbf{x}}$ are positive since p_j and $S_x^c(\omega_k)$ are positive quantities (this can be seen from (5.2)). Therefore, in an attempt to prevent the estimated quantity $\hat{\mathbf{E}}_{\mathbf{x}} = \mathbf{A}\mathbf{G}_{\mathbf{S}}\mathbf{S}_{\mathbf{y}}^{\mathbf{c}}$

from becoming negative, we constrain the diagonal entries $g_s(\omega_k)$ of $\mathbf{G_S}$ to be positive. Furthermore, we minimize the error on a logarithmic scale than on a linear scale. This leads to the following constrained minimization problem:

$$\begin{aligned} \mathbf{G_S^*} &= \underset{\mathbf{G_S}}{\operatorname{argmin}} \,||\log(\mathbf{A}\mathbf{G_S}\mathbf{S_y^c}) - \log(\mathbf{A}\mathbf{S_x^c})||_2^2 \\ \text{subject to} &\quad 0 < g_s(\omega_k) < 1 \quad k = 1, 2, \cdots, N \end{aligned} \tag{5.9}$$

The enhanced signal, $\hat{x}_p^l(n)$ is obtained according to $\hat{x}_p^l(n) = F^{-1}\left[g_s(\omega_k)|Y^l(\omega_k)|e^{j\angle Y^l(\omega_k)}\right]$.

5.3 Implementation Details

In a practical scenario, both speech and noise power spectrums change with time and therefore it is necessary to obtain reasonable estimates of their respective power spectrums at regular intervals. Moreover, the computation of $\mathbf{G_S}$ in (5.9) depends largely on accurate estimation of the speech and noise power spectrums. In this section, we briefly describe the speech and noise power spectrum estimation techniques.

Estimation of Noise Power Spectrum

In this paper, we employ the minima-controlled recursive averaging algorithm proposed in [74] for noise spectrum estimation. The local noisy speech power spectrum $\hat{S}_y^l(\omega_k)$ is smoothed in time using a first-order recursive averaging procedure:

$$\hat{S}_y^l(\omega_k) = \alpha_s \hat{S}_y^{l-1}(\omega_k) + (1 - \alpha_s)|Y^l(\omega_k)|^2 \tag{5.10}$$

where α_s is the smoothing parameter. The noise power spectrum is obtained by tracking the minimum of $\hat{S}_y^l(\omega_k)$ over L frames. The noise power spectrum is then updated based on the signal presence probability $p^l(\omega_k)$ in the l^{th} frame as

$$\hat{S}_d^{l+1}(\omega_k) = \tilde{\alpha}_d \hat{S}_d^l(\omega_k) + [1 - \tilde{\alpha}_d]|Y^l(\omega_k)|^2 \tag{5.11}$$

where $\tilde{\alpha}_d = \alpha_d + (1 - \alpha_d)p^l(\omega_k)$ is a time-varying smoothing parameter which is updated according to the signal presence probability $p^l(\omega_k)$ and α_d is a fixed smoothing parameter.

Estimation of Speech Power Spectrum

From the noise power spectrum estimate, an estimate of the clean speech power spectrum is obtained based on the spectral over-subtraction technique proposed in [72]. Let $D^l(\omega_k) = |Y^l(\omega_k)|^2 - \alpha \hat{S}_d^l(\omega_k)$, then an estimate of the power spectrum is obtained by:

$$\hat{S}_x^l(\omega_k) = \begin{cases} D^l(\omega_k) & \text{if } D^l(\omega_k) > \beta \hat{S}_d^l(\omega_k), \\ \beta \hat{S}_d^l(\omega_k) & \text{otherwise} \end{cases} \tag{5.12}$$

where $0 < \alpha < 1$ is the over-subtraction factor which is adapted according to the posterior signal-to-noise ratio and $0 < \beta \ll 1$ is the noise floor parameter.

5.4 Experiments and Evaluation Results

In this section, we describe the experimental setup and compare the performance of i) the proposed algorithm with ii) the Wiener filter and iii) the spectral subtraction approach. All the three schemes are provided with the same speech and noise power spectrum estimates and their performance is evaluated.

Experimental Setup

The performance of the algorithms were evaluated using noisy speech excerpts available in the NOIZEUS corpus [75]. We considered white noise, airport noise and babble noise at four different signal-to-noise ratios (0dB, 5dB, 15dB and 20dB). The speech files were analyzed using short segments of 32-ms duration frames with 50% overlap between frames. The parameters used for the noise power spectrum estimation are as follows: $\alpha_s = 0.8$, $\alpha_d = 0.95$, $\delta = 5$, $\alpha_p = 0.2$ and search window $L = 1$ s. For speech power spectrum estimation, the noise floor parameter was set to $\beta = 0.002$.

The noisy speech files were referenced to an assumed sound pressure level (SPL) of 90 dB for calculating the auditory patterns and the final loudness estimates. The signal was reconstructed in the time-domain using an overlap-add synthesis procedure.

Wiener Filter and Spectral Subtraction

We compare the performance of the proposed algorithm with the Wiener filter. The Wiener filter minimizes $\mathrm{E}[(\hat{x}(n) - x(n))^2]$ where $\hat{x}(n)$ denotes the estimated signal and E denotes the expectation operator. Assuming a linear model for $\hat{X}(\omega_k)$, the Wiener solution can be equivalently represented in the frequency domain as,

$$H(\omega_k) = S_x(\omega_k)/(S_x(\omega_k) + S_d(\omega_k)) \tag{5.13}$$

where $S_x(\omega_k)$ and $S_d(\omega_k)$ denote the power spectral density of $x(n)$ and $d(n)$ respectively. However, due to the non-stationary nature of speech and noise signals, the true power spectrum in (5.13) is replaced by their estimated quantities $\hat{S}_x^l(\omega_k)$ and $\hat{S}_d^l(\omega_k)$. The signal is reconstructed in the time domain as $\hat{x}_w^l(n) = F^{-1}\left[H(\omega_k)|Y^l(\omega_k)|e^{j\angle Y^l(\omega_k)}\right]$

The proposed algorithm is also compared with a spectral subtraction type algorithm. The enhanced signal $\hat{x}_s^l(n)$ is obtained according to

$$\hat{x}_s^l(n) = F^{-1}\left[\sqrt{\hat{S}_x^l(\omega_k)}e^{j\angle Y^l(\omega_k)}\right] \tag{5.14}$$

Metrics

We compare the performance of the three schemes in terms of the signal-to-noise ratio (SNR), segmental SNR (SSNR), the absolute loudness error (ALE) and the relative loudness error (RLE). The absolute loudness error and the relative loudness error can be expressed as:

$$ALE = \frac{1}{K}\sum_{i=1}^{K} |L_{\hat{x}}(i) - L_x(i)| \tag{5.15}$$

$$RLE = \frac{1}{K}\sum_{i=1}^{K} \left|\frac{L_{\hat{x}}(i) - L_x(i)}{L_x(i)}\right| \tag{5.16}$$

where $L_{\hat{x}}(i)$ and $L_x(i)$ represent the loudness (in sones) of the i^{th} segment of the enhanced speech and the clean speech respectively. K denotes the total number of frames. We note that the relative loudness error is a more suitable metric for comparison than the absolute loudness error as it facilitates comparison over a wide dynamic range of sound intensities. The performance of the proposed estimator in terms of the extent of matching between the auditory patterns of the estimated signal and that of the clean signal can be judged based on their loudness differences. For this reason, we consider the ALE and RLE error metrics. In Tables 5.1, 5.2, 5.3, the comparative performance for babble noise, airport noise and white noise cases are shown for the three techniques being compared. Firstly, it can be observed that the proposed algorithm attains a minimum average value for the relative loudness error across all three noise types at different SNR condition. This behavior can be attributed to the fact that we minimize a squared error between the logarithm of the two auditory patterns which is equivalent to minimizing the squared error of the ratio of the auditory patterns.

Table 5.1: Comparison of Techniques for Babble Noise Case.

SNR	Metric	Wiener	Spec. Sub.	Proposed
0 dB	SNR/SSNR	2.48/-9.5	2.77/-9.2	3.32/-8.2
	ALE/RLE	9.64/1.32	9.4/1.3	8.08/1.09
5 dB	SNR/SSNR	6.34/-5.16	6.6/-4.84	6.97/-4.07
	ALE/RLE	6.36/0.93	6.2/0.91	5.39/0.76
10 dB	SNR/SSNR	11.32/-0.22	11.58/0.08	11.7/0.69
	ALE/RLE	3.72/0.67	3.63/0.6	3.23/0.49
15 dB	SNR/SSNR	14.91/3.51	15.1/3.74	15.08/4.18
	ALE/RLE	2.57/0.39	2.47/0.39	2.36/0.33

Secondly, comparison of ALE and RLE metrics for the Wiener and Spectral subtraction techniques reveal almost similar RLE measures for the two techniques indicating the fact that no explicit auditory modeling has been incorporated in them. This trend is seen across all types of noise considered at various input SNR levels. On the other hand, the proposed technique shows improvement in the ALE and RLE measures due to incorporation of an explicit auditory model. Moreover, the proposed technique also shows a corresponding performance improvement in the SNR and SSNR metrics. This effect can be attributed to the fact that the gain function $\mathbf{G_S}$ is applied to the power spectrum thereby also preserving the spectral characteristics of the estimated signal.

Finally, it can be observed that the proposed estimator is more effective at low input SNR conditions rather than at the high input SNR conditions thereby indicating that incorporation of auditory modeling might be more beneficial in low SNR conditions.

Table 5.2: Comparison of Techniques for White Noise Case.

SNR	Metric	Wiener	Spec. Sub.	Proposed
0 dB	SNR/SSNR	6.6/-5.9	6.84/-5.72	7.03/-4.3
	ALE/RLE	6.82/0.98	7.13/1.08	7.49/0.82
5 dB	SNR/SSNR	9.84/-2.56	10.0/-2.41	9.92/-1.53
	ALE/RLE	5.69/0.81	5.69/0.85	6.15/0.68
15 dB	SNR/SSNR	15.82/4.1	16.0/4.14	15.44/4.39
	ALE/RLE	3.35/0.44	3.31/0.44	3.9/0.39

5.5 Summary

In this chapter, we described the proposed auditory domain based speech enhancement algorithm that minimizes the error between the auditory representation associated with the estimated and the desired signal. We show that the proposed estimator attains a lower average relative loudness error compared to a Wiener or a spectral subtraction based technique with the same noise estimation algo-

Table 5.3: Comparison of Techniques for Airport Noise Case.

SNR	Metric	Wiener	Spec. Sub.	Proposed
0 dB	SNR/SSNR	2.91/-9.4	3.24/-9.0	3.86/-8.0
	ALE/RLE	8.84/1.08	8.57/1.25	7.33/1.03
5 dB	SNR/SSNR	7.44/-4.53	7.7/-4.22	8.02/-3.34
	ALE/RLE	5.58/0.84	5.46/0.83	4.68/0.67
10 dB	SNR/SSNR	10.75/0.02	10.96/0.27	11.11/0.83
	ALE/RLE	3.66/0.56	3.61/0.56	3.27/0.46
15 dB	SNR/SSNR	15.08/3.99	15.28/4.25	15.43/4.75
	ALE/RLE	2.21/0.36	2.14/0.35	1.9/0.29

rithm. Furthermore, the proposed algorithm models the mechanism of the human auditory system by including the auditory model characteristics explicitly in the error criterion rather than considering only a measure of perceptual behavior in a heuristic manner. We also note that the proposed technique avoids estimation of masked thresholds from the noisy input signals as is typically done in several perceptual speech enhancement algorithms.

Chapter 6

Perceptual Sinusoidal Component Selection

6.1 Introduction

Over the years, researchers have studied several mathematical representations of the human auditory system for the purpose of using them in audio compression algorithms. Perhaps, the most popular of these representations is the global masking threshold [2] which is used to shape the quantization noise (so that they are rendered inaudible) in standardized audio compression algorithms such as the ISO/IEC MPEG-1 layer 3 [19], the DTS [76], and the Dolby AC-3 [20] standards. Recent research [33] suggests that perceptual models that employ auditory patterns (AP) rather than masking thresholds maybe more beneficial in audio compression. This is because, auditory patterns not only take into account the masking phenomenon but also other perceptual aspects such as loudness, the nonuniform sensitivity of the human auditory system and the adaptive control of the cochlear gain to the intensity level of incoming audio [13,62]. In addition, the auditory pattern outputs correspond to physiological and neural responses at the different intermediate stages of the auditory system as shown in the bottom of Fig. 6.1. These advances in understanding the human auditory system have led to the development of several sophisticated auditory modeling techniques [3,8,10,11,13] that generate internal auditory representations (or auditory patterns).

In view of this, several techniques based on employing auditory patterns have been proposed for the purpose of speech/audio coding. For example, in [35], a bandwidth extension algorithm was proposed that makes use of auditory patterns to determine the perceptual importance of the different high-band sub-bands in order to reduce the amount of side-information bits transmitted. Similarly, in [36], a rate determination algorithm based on loudness criterion was proposed for use in variable bit-rate speech coders. Also, several objective metrics that predict

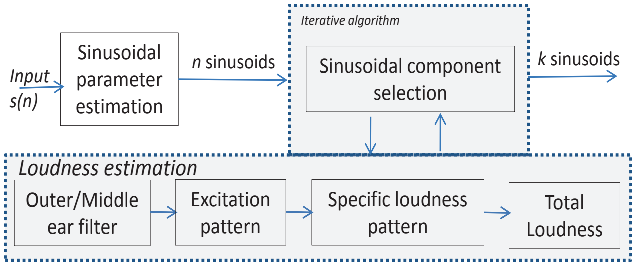

Figure 6.1: General structure of a sinusoidal component selection task.

subjective quality such as PERCEVAL [24], POM [23] or PESQ [22] make use of auditory patterns. However, in [41], speech coding was accomplished by encoding the auditory patterns rather than the frequency components and reconstruction was accomplished using an inverse auditory mapping. Finally, auditory patterns have been used to select perceptually salient sinusoids in several parametric coding techniques [33, 34, 44, 77] including the more recent MPEG-4 HILN (Harmonics plus individual lines and noise) audio coder [77].

In this chapter, we make use of the auditory model developed by Moore & Glasberg [3] to evaluate the auditory patterns. In particular, the model generates an excitation pattern and a loudness pattern during the intermediate stages of the model in addition to obtaining the final instantaneous loudness measure. The excitation pattern represents the magnitude of the basilar membrane vibrations whereas the loudness pattern represent the stimulation of the neural receptors present along the basilar membrane [62]. In this chapter, the excitation/loudness patterns are referred to as the “auditory patterns” or “auditory representations”.

6.2 Problem Statement

In this section, we focus on the problem of selecting perceptually salient sinusoids for use in parametric models of speech/audio coding. The parametric models make use of signal models or source models for compact signal representations, i.e., they exploit signal redundancy [77]. For example, the MPEG-4 audio standard consists of the HILN parametric audio coder. This coder makes use of a sinusoidal signal model, a transient signal model and a noise signal model to exploit signal redundancy. The sinusoidal model extracts frequencies, amplitudes and phases associated with individual frequency components in the underlying signal. The residual signal is treated as the noise component and its spectral envelope is modeled using linear prediction techniques.

These parametric techniques can achieve higher quality at much lower bit-rates compared to the traditional transform-domain audio coders [77]. Therefore, parametric methods have become popular in several Internet streaming and broadcasting applications. Due to the desire for low bit-rates, perceptual techniques are used to select parameters associated with a particular signal model. For example, loudness measures are used to select a limited number of sinusoidal components from the complete set of sinusoidal components. Often times it is also desired that these limited set of sinusoidal parameters be selected such that the target bit-rate is scalable with target perceptual quality. That is, a gradual degradation in quality with decreasing bit-rate is desired. Hence, source models are combined with perceptual models so that signal irrelevancy can also be exploited in addition to signal redundancy.

In this paper, we focus on a perceptual sinusoidal component selection task. First, a set of candidate sinusoids are first extracted using sinusoidal analysis techniques; here, we make use of the peak picking procedure described in [78]. Following this, a limited number of sinusoidal components are selected from this candidate set. To that end, perceptual models are employed since the final target is a human listener. Several techniques that make use of perceptual models have been proposed in the literature: For example, the MPEG-4 HILN audio coder makes use of the signal-to-mask ratio (SMR) criterion to identify perceptually salient sinusoidal components [77]. Recently, in [33], the authors describe an excitation pattern matching algorithm where the sinusoids whose excitation pattern results in the best matching (i.e., least error) to the original signal's excitation pattern are selected. This technique was later extended in [34] where the sinusoidal component selection was carried out based on loudness pattern matching. In both approaches, the authors show that, at low bit-rates, the set of sinusoids selected by minimizing either the excitation pattern or the loudness pattern differences are significantly different from those selected through a maximum-SNR (signal-to-noise ratio) or a maximum-SMR criteria. In Section , the difference between masking models and auditory pattern models are illustrated.

Although the existing techniques based on matching auditory patterns show improvement over masking-based approaches, these techniques are associated with a high computational complexity which makes them impractical for use in most audio coding applications. In the next section, we will highlight the computational complexity associated with different approaches and describe the proposed low-complexity approach. This is primarily due to the presence of multiple nonlinearities in the auditory model stages that presents difficulties in solving perceptual objective functions.

Computational Complexity Analysis

For the sinusoidal component selection task, a subset of L sinusoids need to be selected in a perceptually relevant manner out of N candidate sinusoids. The optimal solution is the one that results in the least error between the auditory patterns associated with the modeled signal (consisting of L sinusoids) and the original signal (consisting of N sinusoids) respectively. This optimal solution is usually found through an exhaustive search procedure, i.e., one has to evaluate the auditory patterns associated with each of the $\binom{N}{L}$ sinusoidal combinations in order to obtain the optimal selection. This process is combinatorial in nature and involves repeated application of the auditory model stages. It is associated with an $O\left(\binom{N}{L}\right)$ computational complexity which grows exponentially ($\approx O(N^L)$) with increasing L (for $L < N/2$) and hence is not suited for real-time systems.

In an alternate approach, suboptimal algorithms have been employed for sinusoidal component selection. These algorithms are greedy and iterative in nature, i.e., they select one sinusoid in every iteration until a required number of sinusoids are selected. In this paper, we consider the greedy excitation pattern matching (EP) algorithm proposed in [33] as a reference for performance comparisons. The details of the greedy algorithm are described in Section 6.3. The greedy approach is associated with an $O(N+(N-1)+\cdots+(N+(L-1))) = O(NL-(L-1)(L-2)/2)$ computational complexity. That is, in the first iteration, there are N available sinusoids to select from. In the second iteration, there are $N-1$ available sinusoids and so on. More generally, in the L^{th} iteration, there are $N-(L-1)$ candidate sinusoids available. Therefore, the computational complexity associated with the greedy approach is quadratic in L unlike the exponential growth associated with the exhaustive search procedure. Nevertheless, the greedy approach still requires repeated evaluation of the auditory model stages and therefore it is still associated with a high computational complexity.

Proposed approaches

In this paper, we propose a number of techniques that pose the problem of selecting perceptually salient sinusoids as a convex optimization problem. All of the proposed techniques attempt to maximize the matching (i.e., least error) between the excitation patterns associated with the modeled and the original signal respectively. The modeled signal is represented using the small subset of L sinusoidal components, whereas the original signal consists of all N candidate sinusoids. Moreover, recent advances in the field of convex optimization have led to the development of fast and efficient solvers for convex optimization problems.

We propose three techniques that pose the problem of perceptual sinusoidal selection as a convex optimization problem. The first technique minimizes the ℓ_1 error between the excitation patterns associated with the modeled and the original signal. This is referred to as Average Linear Error (ALE) minimization scheme. The second technique minimizes the maximum error between the excitation patterns associated with the modeled and original signal. This scheme is referred to as Maximum Linear Error (MLE) minimization scheme. The third technique minimizes a linear distance between the logarithms of the original and modeled signal's excitation patterns. This scheme is referred to as the Linear Logarithmic Error (LLE) minimization scheme.

The proposed techniques are different from the greedy EP matching algorithm in the following aspects: i) First, the proposed techniques linearizes the excitation pattern evaluation stage by removing the dependance of auditory filter shapes on the intensity level of spectral components. This approximation is required to formulate the ALE and MLE minimization schemes as a linear programming (LP) problem. ii) Secondly, in the proposed techniques, the L sinusoidal components are selected jointly rather than selecting the sinusoids one-by-one in each iteration as done by the greedy EP matching algorithm. That is, in the greedy approach, each sinusoid is optimal only in the particular iteration it is selected (as they are dependant on the sinusoids selected in earlier iterations). Therefore, the combined set of sinusoidal selections across iterations becomes sub-optimal, iii) Thirdly, in the proposed techniques, the auditory model stages need not be repeatedly employed as is carried out in the greedy approach. Therefore, the proposed techniques are computationally efficient than the greedy approach, iv) Finally, the LLE minimization scheme results in significantly lesser excitation pattern error compared to the greedy approach. This indicates that the LLE scheme results in more optimal sinusoidal selections than that obtained from the greedy EP matching approach. Since the results of the exhaustive search procedure are difficult to obtain, the excitation pattern error serves as a guideline in deciding on the optimality of the selected sinusoidal subset.

We evaluate the performance of the proposed technique with that of the greedy EP matching approach using the following metrics: i) residual loudness error (RLE), ii) excitation pattern error (EPE) and iii) the number of common sinusoids that are selected by both the greedy and the proposed schemes.

Simulation results indicate that both the ALE and MLE schemes result in $90 - 95\%$ similarity with the greedy approach in their selected sets of sinusoidal components. On the other hand, the LLE scheme results in only $60 - 70\%$ similarity with the greedy approach in their selected set of sinusoidal components. However, the proposed LLE scheme results in a lower residual loudness error and excitation pattern error compared to the ALE or the MLE schemes. This indicates that the sinusoidal selections obtained from the LLE minimization are closer to the optimal solution (obtained from the exhaustive search procedure) than that obtained from the ALE, MLE or the greedy approaches. Our results indicate that the proposed set of algorithms not only outperforms SMR-based sinusoid selection algorithms but also operates at a much lower computational complexity compared to existing excitation pattern matching algorithms and in some cases, also results in better sinusoidal selections than that obtained from the greedy scheme.

This paper is organized as follows. In Section 6.3 a brief overview of the sinusoidal model, the excitation pattern matching algorithm and the auditory model specifics are described. In section 6.4 we describe the proposed sinusoidal selection algorithms. In Section 6.4, we compare the performance of the proposed techniques to that of the greedy EP matching algorithm followed by concluding remarks in Section 6.6.

6.3 Background on Sinusoidal Model, Excitation Pattern Matching and Auditory Model

In this section, the underlying sinusoidal model and the excitation pattern matching algorithm are described briefly. The proposed fast implementation in [38] can also be employed to increase the computational efficiency of the sinusoidal selection process.

Sinusoidal Model

Let $x_j(n)$ and $\hat{x}_j(n)$ denote two length M discrete-time signals corresponding to the reference audio segment and a coded version of the same segment respectively, where j indicates the frame index. More specifically, $\hat{x}_j(n)$ represents an estimate of $x_j(n)$ using only L out of N possible sinusoids. Mathematically,

$$x_j(n) \approx \hat{x}_j(n) = \sum_{k=1}^{L} A_k(n) cos(\omega_k(n)n + \phi_k(n)), \tag{6.1}$$

where $A_k(n)$, $\omega_k(n)$ and $\phi_k(n)$ represents the time-varying amplitudes, frequencies and phases associated with each of the k sinusoidal components. For the purposes of illustrating the proposed idea, the amplitudes, frequencies and phases can be assumed to remain stationary within each audio segment; hence, we drop the argument n from them. Also, the frame index j is dropped for simplicity.

Excitation Pattern Matching

The excitation pattern matching algorithm was initially introduced in [33] and provides a framework to select a subset of perceptually salient sinusoids from a larger set of candidate sinusoids. That is, given a candidate set of N sinusoidal components, the excitation pattern matching algorithm selects a subset of L ($L << N$) sinusoids such that they provides a maximum perceptual benefit. The perceptual benefit is measured according to a perceptual objective function that includes the auditory model in its formulation.

The excitation pattern matching algorithm described in [33] is a greedy algorithm that selects one sinusoid in every iteration. For example, in the first iteration, the excitation pattern corresponding to each candidate sinusoid is computed individually and the one that provides the maximal increment in the excitation pattern is selected. Since the excitation pattern corresponds to a measure of basilar membrane vibrations, a maximal increment in excitation pattern also indicates a corresponding improvement in subjective performance. In other words, the following error is minimized:

$$\Delta_i = \sum_{k=1}^{D} E(k) - E_i(k) \tag{6.2}$$

where $E(k)$ denotes the reference excitation pattern with all the N sinusoids and $E_i(k)$ denotes the test excitation pattern with the i^{th} sinusoid included and D denotes the number of detector locations where the excitation pattern is evaluated.

Therefore, in the first iteration, one sinusoid that minimizes (6.2) is selected. In subsequent iterations, each of the remaining unselected sinusoids are combined individually with the previously selected sinusoids and the sinusoid that corresponded to the combination which resulted in a maximal increment in the excitation pattern is selected. More generally, in the p^{th} iteration, each of the remaining $n-(p-1)$ sinusoids are individually combined with the $p-1$ previously selected sinusoids and the sinusoid that resulted in a maximal increment in excitation pattern or minimum error according to (6.2) is selected. This procedure is repeated until a target number of sinusoids (corresponding to a desired bit-rate) are selected.

6.4 Proposed Sinusoidal Selection Algorithms

We now describe the proposed techniques for sinusoidal component selection by solving a set of constrained convex optimization problems. In order to formulate the problem, the excitation pattern evaluation stage is linearized and modeled using matrix notations. The matrix formulation is helpful in establishing the sinusoid component selection process as an optimization problem in the ensuing sections.

Excitation Pattern Modeling

We first express (5.2) using matrix notations. Let the spectrum $S_x^c(i)$ be denoted by the vector $\mathbf{x} \in \mathbf{R}^{N\times 1}$ and the resulting excitation pattern $E(k)$ by $\mathbf{E_x} \in \mathbf{R}^{D\times 1}$, we can then write:

$$\mathbf{E_x} = \mathbf{A}(\mathbf{x})\mathbf{x} \tag{6.3}$$

where $\mathbf{A} \in \mathbf{R}^{D\times N}$ and the elements of $\mathbf{A}$ are given by $a_{k,i} = (1+p_k g_{k,i})\exp(-p_k g_{k,i})$ which represent the auditory filter magnitudes. Moreover, $\mathbf{A}(.)$ denotes that the matrix is a function of the parameter within the parenthesis since the $\{p_k\}$'s are dependent on $S_x^c(i)$. If we remove the dependance of $\{p_k\}_{k=1}^{D}$'s on $\mathbf{x}$, then (6.3) can be expressed as,

$$\mathbf{E_x} = \mathbf{A}\mathbf{x} \tag{6.4}$$

This is equivalent to assuming that the auditory filters are symmetric and that the slopes $\{p_k\}$'s are no longer dependent on the intensity level $\mathbf{x}$, thereby linearizing (6.3).

The change in the filter shapes according to the intensity level of the incoming audio is consistent with the human auditory system's mechanism of controlling the gain of the cochlea. For example, at high intensity levels, the auditory system reduces the cochlear gain as a precautionary measure so as to prevent loudness levels from reaching thresholds of pain.

Optimized Selection of Sinusoids

We present three methods for the selection of sinusoids based on excitation pattern matching technique.

Let $\hat{\mathbf{x}}$ denote the magnitude spectrum associated with the reconstructed signal containing a L out of N sinusoidal components. This can be expressed as,

$$\hat{\mathbf{x}} = \mathbf{X}\mathbf{b}, \tag{6.5}$$

$\mathbf{X} = \mathrm{diag}(\mathbf{x})$, and $\mathbf{b} \in \{0, 1\}$ is a binary vector that selects a subset of frequency components from $\mathbf{x}$. In other words, $\hat{\mathbf{x}}$ contains zeros at all frequency locations not selected by the binary vector $\mathbf{b}$. The excitation pattern associated with the reconstructed signal $\hat{x}(n)$ is denoted by

$$\mathbf{E}_{\hat{\mathbf{x}}} = \mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\mathbf{X}\mathbf{b} \tag{6.6}$$

where $\mathbf{X} = diag\{\mathbf{x}\}$ is combined with the matrix $\mathbf{A}$ with auditory filter magnitudes.

Now, we would like to select the L perceptually most salient sinusoids that minimize the difference between the original and reconstructed excitation patterns. Expressing this problem mathematically,

$$\begin{aligned} \underset{\mathbf{b}}{\mathrm{argmin}} \quad & \mathcal{D}(\mathbf{E}_{\mathbf{x}}, \mathbf{A}\mathbf{X}\mathbf{b}) \\ \text{subject to} \quad & \mathbf{s}^{\mathrm{T}}\mathbf{b} = L, \\ & \mathbf{b} \in \{0, 1\}. \end{aligned} \tag{6.7}$$

where $\mathcal{D}(\mathbf{E}_{\mathbf{x}}, \mathbf{K}\mathbf{X}\mathbf{b})$ is a measure of the error between the original and reconstructed excitation patterns and $\mathbf{s} = [1, \ldots, 1]^{\mathrm{T}}$. The optimal solution, $\mathbf{b}^{\mathrm{opt}}$, contains L nonzero entries and the indices of these entries correspond to the perceptually salient sinusoids. In the ensuing sections, we discuss the solution to the problem in (6.7) for several distance metrics.

Minimizing the Linear Error

In this section, we formulate two novel methods of selecting a subset of perceptually salient sinusoids based on minimizing a function of the residual $\mathbf{e}$, where

$$\mathbf{e} = \mathbf{E}_{\mathbf{x}} - \mathbf{E}_{\hat{\mathbf{x}}} = \mathbf{E}_{\mathbf{x}} - \mathbf{A}\mathbf{X}\mathbf{b}. \tag{6.8}$$

The first method relies on minimizing $||\mathbf{e}||_1$, i.e., $\mathcal{D}(\mathbf{E_x}, \mathbf{AXb}) = ||\mathbf{E_x} - \mathbf{AXb}||_1$. Rather than minimizing the above formulation directly, we minimize a slightly different formulation by noting the following properties. Firstly, it is important to note that $\mathbf{b} \in \{0, 1\}$. This implies that $\mathbf{E_x} \geq \mathbf{E_{\hat{x}}}$ since $\mathbf{b} = 1$ corresponds to selecting all the components in the original vector $\mathbf{x}$. Secondly, since $\mathbf{e} \geq 0$, minimizing $||\mathbf{e}||_1$ is equivalent to minimizing the sum of the individual residual entries e_i. That is, the optimal selector vector $\mathbf{b}^{\mathrm{opt}}$ can be found by solving the following problem:

$$\begin{aligned} \underset{\mathbf{b}}{\text{argmin}} \quad & \sum_{i=0}^{N_{\text{det}}-1} e_i \\ \text{subject to} \quad & \mathbf{s}^{\mathrm{T}}\mathbf{b} \leq L, \\ & \mathbf{b} \leq 1, \\ & \mathbf{b} \geq 0. \end{aligned} \tag{6.9}$$

Notice, we replace the binary constraint in the formulation above with a linear constraint. The region of feasibility of the problem in (6.9) is a convex polytope resulting from the intersection of the three half-spaces describing the constraints. The vertices of this polytope include all possible combinations of L out of N ones in $\mathbf{b}$. For example, for selecting $L = 2$ out of $N = 3$ sinusoids the vertices of the polyhedron of feasibility are (0,0,0), (0,1,1), (1,0,1), and (1,1,0). It is a well-known fact that the solution of the linear programming problem will lie on the vertices of the region of feasibility, therefore the binary constraint in (6.7) is redundant. This greatly simplifies the problem and allows for the use of Linear Programming techniques for solving (6.9) in a computationally efficient manner.

An alternate formulation relies on minimizing the maximum of the residual error, i.e., minimizing $\max(\mathbf{e})$, instead of minimizing its L1-norm. We introduce a new scalar, t, to bound the largest value of the vector residual $\mathbf{e}$. That is, we minimize

$$\begin{aligned} \underset{\mathbf{b},t}{\operatorname{argmin}} \quad & t \\ \text{subject to} \quad & -t \leq \mathbf{e} \leq t, \\ & \mathbf{s}^{\mathrm{T}}\mathbf{b} \leq L, \\ & \mathbf{b} \leq 1, \\ & \mathbf{b} \geq 0. \end{aligned} \tag{6.10}$$

As in the first method, the binary constraint has been removed. Due to the added constraint that bounds the error, it is no longer apparent that the optimal solution is binary. From our experience, we see that the solution is almost binary; therefore we select the sinusoids corresponding to the closest binary solution to the minimizer of (6.10).

The solution to the linear programming (LP) problems in (6.9) and (6.10)) can be obtained through iterative algorithms. In this paper we make use of a variant of Mehrotra's predictor-corrector algorithm [79] optimized for solving LP problems with a large number of unknowns.

Minimizing the Log Error

Rather than minimizing a linear function of the error, we propose yet another formulation of the problem in (6.7). This is motivated by the 1 dB difference rule proposed by Zwicker. According to Zwicker's 1 dB model of difference detection [13], two signals $x(n)$ and $y(n)$ with excitation patterns $E_x(k)$ and $E_y(k)$ are perceptually indistinguishable if their excitation patterns differ by less than 1 dB at every frequency.

$$\begin{aligned} \underset{\mathbf{b}}{\text{argmin}} \quad & \sum_{k=0}^{D-1} \log(E_{\mathbf{x}}(k)) - \log(\mathbf{c}_i^{\mathrm{T}}\mathbf{b}) && (6.11) \\ \text{subject to} \quad & \mathbf{s}^{\mathrm{T}}\mathbf{b} \leq L, \\ & \mathbf{b} \leq 1, \\ & \mathbf{b} \geq 0. \end{aligned}$$

where $\mathbf{C}^{\mathrm{T}} = (\mathbf{AX}) = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_D]$

As was the case with the two previous algorithms, the binary constraint has been removed. Due to the added constraint that bounds the error, it is no longer apparent that the optimal solution is binary. Therefore, the resulting $\hat{\mathbf{b}}$ vector is rounded to the closest integer (either 0 or 1).

Results and Discussion

In this section, the performance of the proposed techniques are tested on different types of speech and music signals.

Experimental Setup

The audio signals are sampled at 44100 Hz and split into frames of size $N_f = 512$ samples. Spectral analysis is carried out using an N_f-point (=512) fast Fourier transform (FFT). Furthermore, the spectral components are referenced to an assumed playback level of 90 dB SPL (sound pressure level). For every frame of audio, a set of sinusoids are extracted by following the simple peak-picking procedure described in [78]. This set of estimated sinusoids constitute the candidate set of N sinusoids and the objective behind the proposed techniques is to select a subset L ($L << N$) of sinusoidal components in a perceptually relevant manner.

The greedy EP matching algorithm described in Section 6.3 is used as a benchmark to compare the performance of the proposed techniques. In particular, the following metrics are used to evaluate the performance: 1) the number of selected sinusoidal components that are in common between that selected by the proposed and the greedy approaches, i.e., their percentage similarity, 2) residual loudness error, 3) excitation pattern error.

Percentage Similarity

The proposed techniques as well as the greedy EP matching algorithm both attempt to maximize the matching between the modeled signal's excitation pattern to that of the original signal's excitation pattern. However, there are important differences between the proposed techniques and the greedy approach which result in different subsets of sinusoids being selected:

- Firstly, the greedy EP matching algorithm takes into account changes in auditory filter slopes according to changes in intensity level of incoming audio while evaluating the excitation patterns. The proposed techniques, on the other hand, assume that the auditory filters are fixed during the sinusoidal selection process. This difference can give rise to slightly different sinusoidal subsets being selected.

- Secondly, the greedy EP matching approach is an iterative algorithm where only one sinusoid is selected in each iteration. The iterations are continued until the required number of sinusoids are selected. On the other hand, the proposed techniques selects the subset of sinusoids jointly.

- Thirdly, since the greedy EP matching algorithm is iterative in nature, the sinusoids selected in future iterations are dependent on the selections made in each of the earlier iterations. This process of selecting sinusoids does not pick the optimal solution (defined as the one that results in the least EP error), i.e., the combined set of sinusoidal selections across all iterations are not jointly optimal even though the sinusoidal selection made in each iteration is individually optimal. This loss of optimality across iterations occurs due to the non-linear dependance of the auditory filter slopes on the intensity level of input frequency components. On the other hand, with the proposed techniques, the decisions on which sinusoids need to be selected are made jointly.

Therefore, it is important to evaluate the extent to which the proposed techniques select the same sinusoids as that of the greedy EP matching algorithm. In Figures 6.2 and 6.3, we plot the percentage of components that are in common between the greedy algorithm and each of the proposed techniques (i.e., ALE, MLE, LLE) for a speech and music signal. In addition, the number of sinusoidal components selected by the SMR approach that are in common with the greedy approach is also shown for comparison.

In the case of both speech and music signals, the ALE Scheme as well as the MLE Scheme selects between 90% − 95% of the same sinusoids as that of the greedy approach. Similarly, the LLE Scheme results in 60% − 70% similar subset of sinusoids as that of the greedy approach. On the other hand, the SMR Scheme results in the least similarity with only 15% − 20% common sinusoids with that of the greedy approach. This shows that the SMR metric of selecting sinusoids is vastly different from the EP matching technique followed by the ALE, MLE, LLE and greedy schemes.

Residual Loudness Error

The residual loudness error measures the difference in loudness between the reference signal (represented with N sinusoids) and the modeled signal (represented with $L << N$ sinusoids). More specifically, the residual error is measured according to:

$$L_e(\text{in sones}) = \frac{1}{P}\sum_{k=1}^{P} L_r(k) - L_m(k) \tag{6.12}$$

where L_e denotes the average loudness difference in sones; L_r and L_m denote the loudness associated with the reference signal and the modeled signal respectively. Here, P denotes the number of frames and k is the frame index.

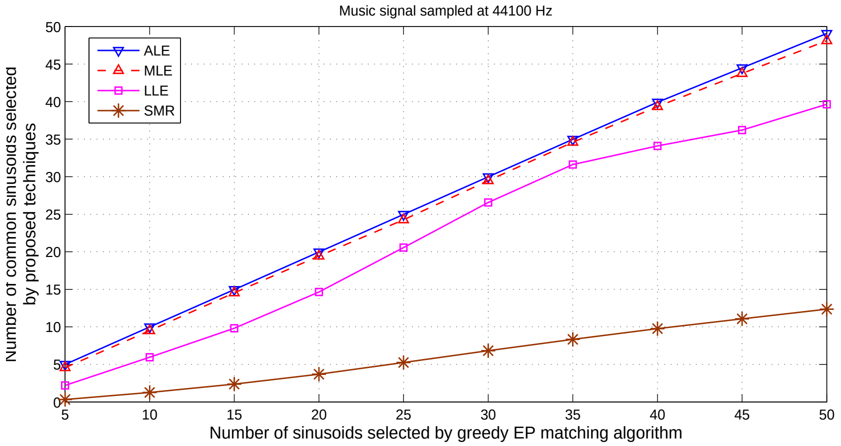

Figure 6.2: Percentage of sinusoidal components in common between the proposed and greedy scheme.

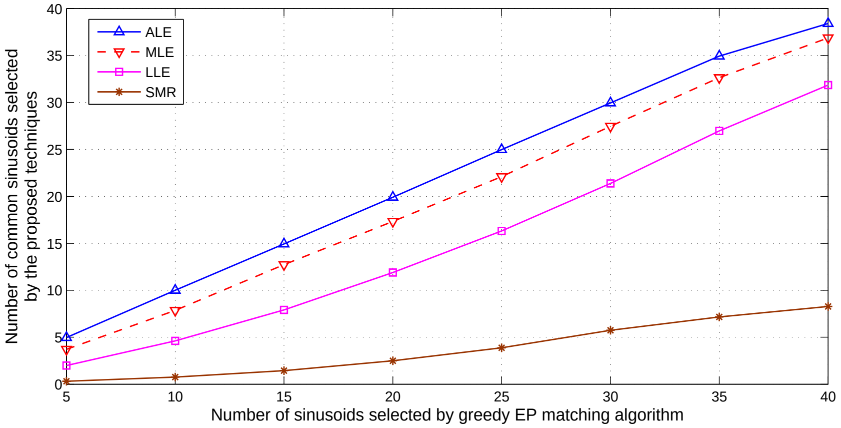

Figure 6.3: Percentage of sinusoidal components in common between the proposed and greedy scheme.

The residual loudness error is chosen as a metric for the following reasons:

1. First, the residual loudness error is more closer to the human perception of sounds as it measures the error in terms of loudness units.

2. The proposed as well as the greedy algorithm optimize to maximize the EP matching, i.e., the modeled signal's excitation pattern should come close to the original signal's excitation pattern. Since, the loudness pattern is related to the excitation pattern through a compressive non-linearity, the better the

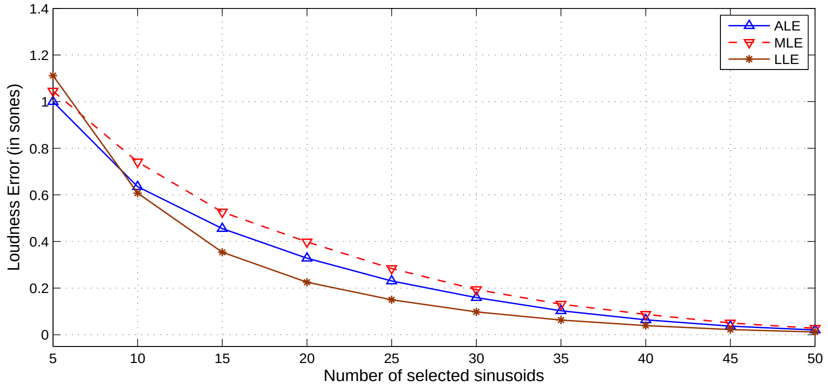

Figure 6.4: Residual Loudness error associated with the different schemes for a speech signal.

EP matching, the closer the modeled signal's loudness pattern to that of the original signal thereby resulting in a smaller residual loudness error. However, the different techniques carry out EP matching by selecting different sets of sinusoids; therefore, the resulting residual loudness error could be vastly different particularly when selecting a smaller subset of sinusoids.

3. Another motivation behind employing the residual loudness error is to assess the optimality of the different techniques as compared to the exhaustive search procedure. It is evident that the exhaustive search procedure selects the optimal subset and the best EP matching and therefore also results in the lowest possible residual loudness error. Therefore, the lower the residual loudness error associated with a particular approach, the closer it comes to the results obtained from the exhaustive search procedure. This can be used to rank the optimality of the different approaches.

In Fig. 6.5 and 6.4, we plot the residual loudness error associated with the different schemes (i.e., ALE, MLE, LLE and Greedy schemes) for a speech and music signal respectively. In both cases, the greedy EP, the ALE and the MLE schemes perform similarly. The MLE scheme has a slightly higher residual loudness error than the ALE or the greedy approach. The LLE scheme, on the

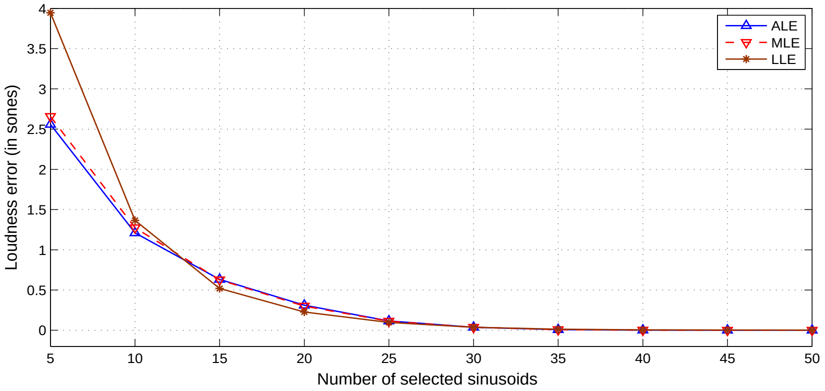

Figure 6.5: Residual Loudness error associated with the different schemes for a music signal.

other hand, exhibits a cross-over point between selecting 10 and 15 sinusoids. For subsets greater than 10, the LLE approach has a lower residual loudness error. This indicates better EP matching capability of the LLE approach compared to the other techniques.

Both the performance metrics (the residual loudness error and components selected) together indicate which of the proposed techniques comes closer to the optimal solution (that found using an exhaustive search procedure).

Motivation for using Excitation Pattern Error Metric
Masked Thresholds vs. Auditory Patterns

There are significant differences between employing a masking threshold based approach versus an auditory pattern evaluation based approach. Firstly, the masking threshold is measured along the frequency axis whereas the auditory patterns (e.g., excitation pattern or loudness pattern) are measured along the length of the basilar membrane (“Auditory domain"). Therefore, masking threshold represents an indirect way of judging the inner ear responses.

Secondly, the global masking thresholds represents the cumulative effect of the masking thresholds associated with individual frequency components. This methodology is followed in several state-of-the-art perceptual audio coders in-

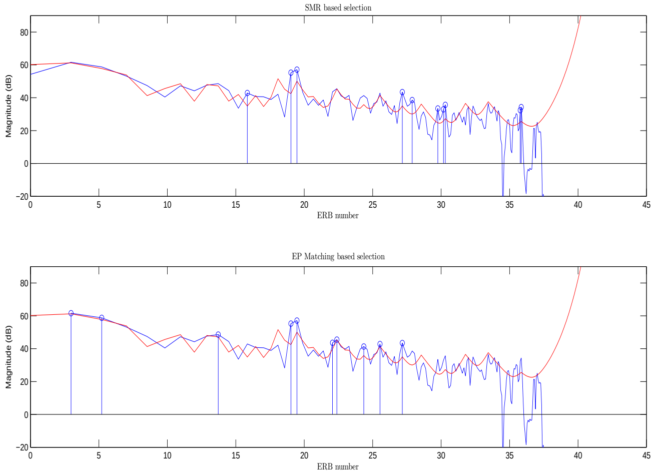

Figure 6.6: A set of 10 sinusoids are selected based on SMR selection criterion (Top) and EP Matching criterion (bottom).

cluding MPEG-1 psychoacoustic model, advanced audio coding (AAC), AC3 etc. Here, the individual masking thresholds are characterized from psychoacoustic experiments that are based on simple combinations of masker and maskee signal types (i.e., either noise or tone like signals). This is because, there does not exist general rules that predict masking behavior for an arbitrary spectrum. In other words, the masked thresholds do not make use of explicit auditory modeling techniques that attempt to model the functioning associated with different parts of the human auditory system. For example, the frequency selectivity of the auditory system is modeled by a bank of bandpass auditory filters whose bandwidths change according to the filter's center frequency. Similarly, the human auditory system adapts the cochlear gain according to the intensity level of the incoming audio in order to prevent hearing damage when sound levels reach the thresholds of pain. This behavior is in turn modeled through a corresponding change in filter slope according to the level of incoming audio. Also, the behavior of the hair cells in converting the mechanical basilar membrane vibrations to electrical nerve impulses is modeled through a nonlinear compressive operation.

Thirdly, with the masking based approach, the masking effect at a particular frequency location is created by all the frequency components in the spectrum. This is very useful in cases where it is required to ”hide" the quantization noise under the masked thresholds so that they become inaudible. For example, in an MP3 coder, the masked thresholds are used for the purpose of quantization, i.e., to decide on the number of bits, such that the resulting quantization noise falls below the masked thresholds. Therefore, when the coded signal is synthesized, both the signal plus quantization noise appear at the decoder. However, the masked threshold that is associated with the original signal effectively masks the quantization noise introduced during the coding process. This is not the case in a sinusoidal component selection task. Here, the objective is also to select a subset of candidate sinusoids in some optimal manner that maximizes auditory perception. Making use of masked thresholds give meaningful results only when the resultant signal (modeled signal in this case) contains all the frequency components that was present in obtaining the masked threshold. However, in a sinusoidal component selection task, the modeled signal (represented by a subset of sinusoidal components) is different from the original signal (represented with a full set of sinusoids). Therefore, making use of masked thresholds (that are obtained from the original signal components) to decide on a suitable subset of sinusoidal components (for representing the modeled signal) is not tractable since the masking effect at a particular frequency location is different for different combinations of candidate sinusoids used for representing the modeled signal.

From an auditory modeling perspective, this can be explained as follows. Every frequency component creates a response along the entire length of the basilar membrane. However, the masked threshold represents the effect of all the frequency components at a particular frequency location. Therefore, it is easier to measure the relative contribution of individual frequency components towards

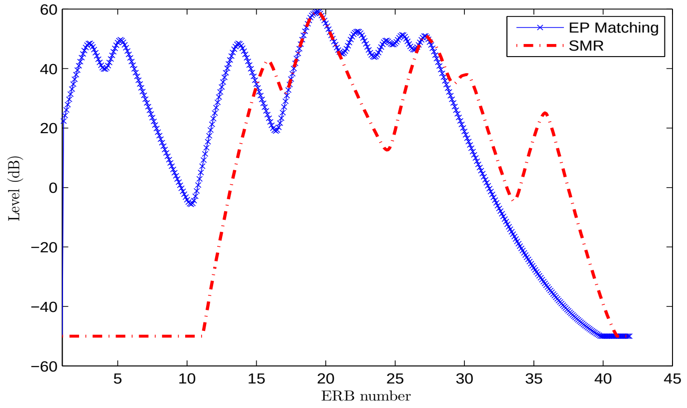

Figure 6.7: Excitation Pattern of sinusoids selected using the SMR based and EP Matching selection criterion.

auditory perception with the auditory pattern evaluation approach rather than with the masked thresholds approach.

For example, in Fig. 6.6, we plot the set of sinusoids that are selected using two different approaches. In the first approach, the sinusoids that correspond to the highest signal-to-mask ratio's are selected as shown in the top of Fig. 6.6. The masking thresholds are also shown for comparison. The second approach is based on the EP matching technique proposed in [2] where the sinusoids that result in a maximal matching between the reference and the modeled excitation patterns are selected (i.e., the least linear error between them). This is shown in the bottom plot of Fig. 6.6. It can be observed that the set of sinusoids selected using the two approaches are vastly different. In particular, it is interesting to note that the SMR approach fails to select the two low-frequency sinusoids that the EP matching approach selects despite them being the strongest components.

With the EP matching approach, this is avoided as each sinusoidal component is selected only if it's individual contribution to the overall auditory pattern is higher than that of the other sinusoidal components. This is best illustrated in Fig. 6.7 where the excitation pattern associated with the original audio segment and that corresponding to the two reconstructed versions of the same audio

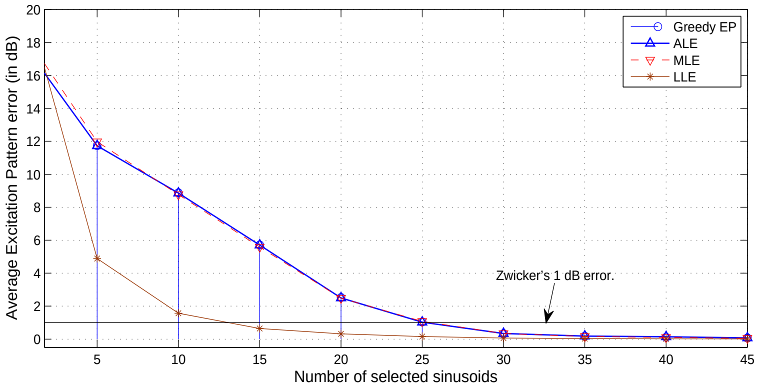

Figure 6.8: Average Excitation pattern error associated with the different schemes for a music signal.

segment are shown. It can be seen that, for the same number of sinusoids, the excitation pattern corresponding to the EP matching approach comes close to the reference EP compared to that resulting from the SMR approach.

Results: Excitation Pattern error

The Excitation pattern error metric is motivated based on the 1 dB detection criterion proposed by Zwicker in [13]. According to the criterion, two signals whose excitation patterns differ by < 1 dB at all center frequencies are perceptually indistinguishable from each other. For the sinusoidal selection task, the proposed techniques are evaluated using the excitation pattern error criterion in order to assess which techniques come close to meeting the < 1 dB criterion faster.

In Figures 6.8 and 6.9, the average excitation pattern error for a music and speech signal are plotted. It can be observed from the figures that the LLE minimization schemes attains a < 1 dB excitation pattern error faster compared to the other proposed techniques or the greedy approach.

6.5 A Hybrid Loudness Estimation Scheme

In this section, the proposed hybrid loudness estimation scheme for sinusoidal signals is described. The idea behind the proposed technique is to estimate the loudness associated with a multi-tone signal from the specific loudness pattern of

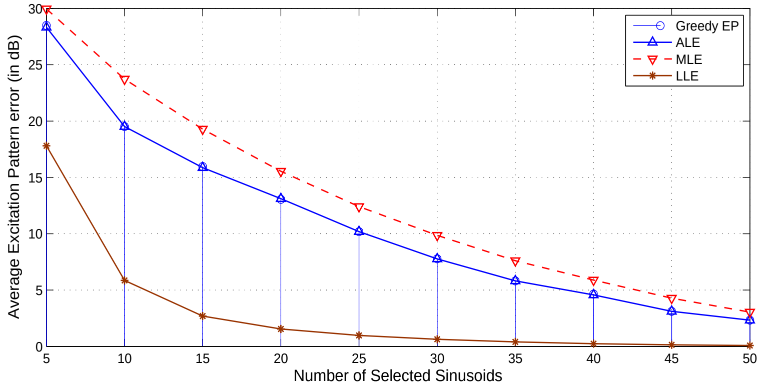

Figure 6.9: Average Excitation pattern error associated with the different schemes for a speech signal.

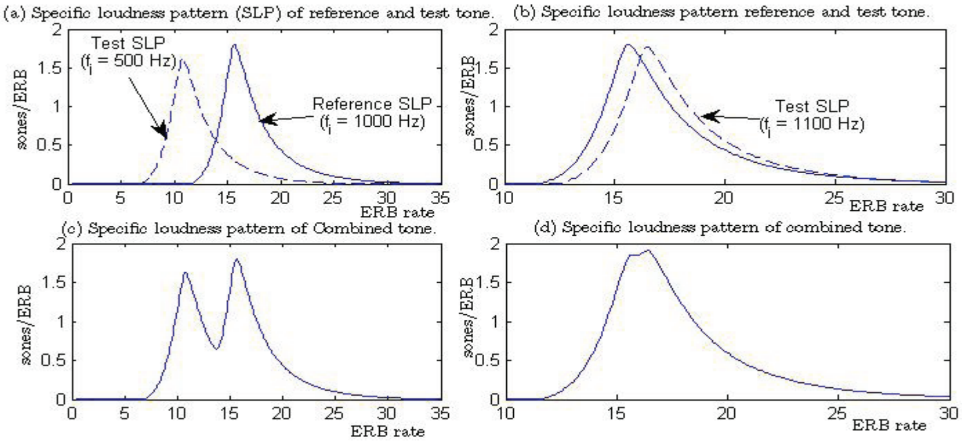

Figure 6.10: Plot of specific loudness patterns of reference, test and combined tones.

its constituent sinusoids. It will then be required to compute the specific loudness patterns of candidate sinusoids only once. An experiment to study the shape of the specific loudness pattern of the combined tone with respect to the specific loudness pattern of the individual sinusoids is described next.

A reference tone of frequency f_i is combined individually with a test tone of frequency f_j to form the combined tone $f_{i,j}$. The specific loudness pattern associated with the reference, test and combined tone is computed. The frequency of the test tone f_j is now varied and the experiment is repeated keeping the frequency of the reference tone fixed. In Fig. 6.10(a) and (b), we plot the specific loudness patterns associated with two different test tone frequencies along with that of the reference tone. The corresponding specific loudness pattern associated with the combined tone is plotted in Fig. 6.10(c) and (d). It can be observed that the envelope of the two specific loudness patterns in Fig. 6.10(a) and (b) closely resembles the exact specific loudness shown in Fig. 6.10(c) and (d). The above experiment was repeated with different choices for the frequency of the reference tone. Based on the experimental observations, we propose a scheme that enables us to estimate the specific loudness pattern of the combined tone from the specific loudness patterns of the constituent sinusoids by retaining the point wise maximum among them. Let $L_T = \{d_k||d_k - d_{k-1}| = 0.1, k = 1, 2, \cdots, D\}$ denote the set of detector locations placed along the ERB scale. If the specific loudness patterns are evaluated on the detector locations described by L_T, then mathematically, this process can be expressed as:

$$\tilde{N}_{ij}(L_T) = max(N_i(L_T), N_j(L_T)) \tag{6.13}$$

where N_i and N_j represent the specific loudness patterns associated with reference and test tones respectively. $\tilde{N}_{ij}$ represents the estimated specific loudness pattern associated with the combined tone $f_{i,j}$. We will refer to this scheme as the “Max" approach. We evaluate the performance of the “Max" scheme in terms of the loudness error, L_e, as

$$L_e \text{ (in sones)} = \int_0^m N_{ij}(z)dz - \int_0^m \tilde{N}_{ij}(z)dz \tag{6.14}$$

where N_{ij} represents the actual specific loudness pattern of the combined tone and m is the total number of ERB units. In Fig. 6.12, we plot the loudness error (L_e) as a function of the frequency separation (in ERB units) between the test and reference tones. The frequency separation (d_{ij}) is obtained using

$$d_{ij} \text{ (in ERB units)} = p_i - p_j \tag{6.15}$$

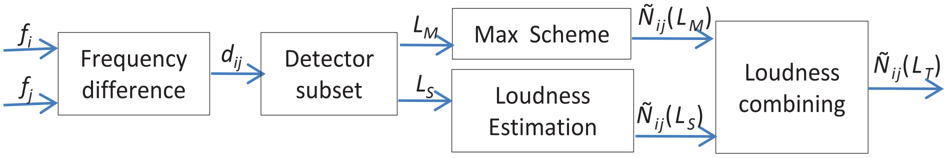

Figure 6.11: Block diagram of the proposed hybrid loudness estimation scheme.

where p_i and p_j are computed using (5.1) and denote the ERB number associated with the reference and test tone respectively. It can be observed from Fig. 6.12 that the error in loudness increases as the frequency separation (d_{ij}) decreases. This can be partly attributed to the fact that when the test and reference tones fall within one ERB unit, the total intensity level within that ERB unit changes causing the auditory filters to change their shapes. This causes a corresponding change in the shape of the specific loudness pattern of the combined tone. However, this change in the auditory filter shape is not accounted in (6.13) when estimating the specific loudness pattern of the combined tone.

To account for the change in filter shapes, we propose a novel approach that combines the “Max" scheme described in (6.13) with an evaluation of the specific loudness pattern in select ERBs. The block diagram of the proposed hybrid loudness estimation process is shown in Fig. 6.11. The steps are described below. First, the frequency separation (d_{ij}) between the test and reference tone is computed using (6.15). If the test and reference tones fall within the same ERB unit, i.e., if their frequency separation, $d_{ij} < 1$ (in ERB units), then an evaluation of specific loudness pattern in select ERBs is employed. A subset of detectors, which we represent by the set L_S, are chosen at locations where there is a significant deviation in the shape of the specific loudness pattern relative to that obtained from (6.13). Let p represent the ERB unit where the auditory filters change shapes. Let $L_S = \{d_k || d_k - p| < m, k = 1, 2, \cdots, D\}$ denote the subset of detectors where the specific loudness patterns are evaluated. Here, m

represents the number of ERB units on either side of the pth ERB unit. For the subset L_S, all the steps associated with the loudness estimation procedure described in [3] are followed. These include the auditory filter shape evaluation, excitation pattern and specific loudness pattern calculation stages. Next, a subset of detector locations L_M is chosen such that $L_T = L_M \cup L_S$ and the specific loudness pattern of the combined tone at detector locations L_M is now estimated according to (6.13). In Fig. 6.12, we plot the loudness error for the proposed hybrid scheme for different values of m. We observe that the hybrid approach is associated with a lower error in loudness and that the loudness error decreases as the detector subset L_S grows. However, the computational complexity increases as the cardinality of the set L_S increases. A detector pruning scheme described as part of a low-complexity loudness estimation procedure in [38,43] can be employed to further reduce the computational complexity.

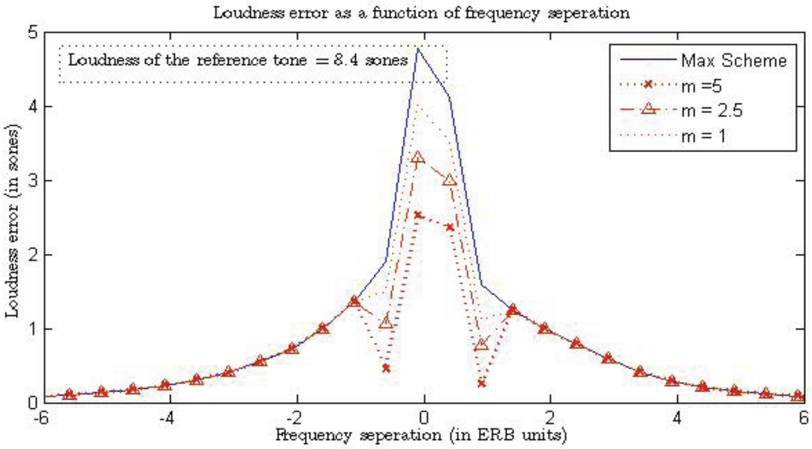

Figure 6.12: Plot of loudness error as a function of frequency separation.

Sinusoidal Selection based on Hybrid Algorithm

In this section, the sinusoidal component selection algorithm based on the hybrid loudness estimation procedure is presented. An input audio segment $s(t)$ is subject to a sinusoidal parameter estimation process. Here, a complete set of n sinusoids is estimated by peak picking [?] in the STFT domain. Let S denote the set of all candidate sinusoids available and $|S|$ denote the cardinality of S. The objective now is to select a subset of k out n sinusoids that provide a maximal increment in the total loudness. An iterative maximization algorithm is employed where the objective in the jth iteration is to select a sinusoid that provides the largest increment in loudness given the previous $j-1$ sinusoidal selections. Let A denote the set containing the selected sinusoids. Initially, $A = \{\}$. During the first iteration, the loudness associated with each sinusoid in S is computed. The sinusoid that provides the largest increment in loudness is selected and added to the set A. During the second iteration, each of the remaining sinusoids in S is individually added to the selected sinusoids in A to form a set of $n-1$ trial signals. The loudness associated with each of the trial signals is evaluated and the sinusoid that contributes towards a largest increment in loudness is selected during the second iteration. This procedure is repeated until all k sinusoids are selected. A total of $n-(j-1)$ trials are associated with the jth iteration and the greedy nature of this algorithm requires that the loudness estimation algorithm be employed $n-(j-1)$ times during the jth iteration. Therefore, to select k sinusoids, the loudness estimation algorithm is executed $n + (n-1) + \text{Ě} + (n - (k-1)) = nk + (k-1)(k-2)/2$ times. This repeated application of the loudness estimation algorithm is computationally demanding and not suitable for real-time applications. We describe below a computationally efficient sinusoidal selection scheme based on the proposed hybrid loudness estimation procedure. A step-by-step description is shown in the algorithm below. Here, instead of evaluating the loudness in each trial by employing all the steps described in Section 3.5, the loudness is estimated from the specific loudness patterns of individual sinusoids using the hybrid scheme. Let i index the set of sinusoids in S. Let p_i and N_i represent the ERB number and specific loudness pattern associated with the ith sinusoid. Let N_i^{tr} represent the estimated specific loudness pattern during the

ith trial and N_j^S denote the estimated specific loudness pattern after j sinusoidal selections.

Results

In this section we present simulation results. The performance of the algorithm was tested with different types of audio records obtained from the SQAM database [69]. The audio signals are sampled at 44.1 kHz and audio segments of 20 ms duration referenced to an assumed Sound Pressure Level (SPL) of 90 dB were used in our simulations. A set of $n = 40$ sinusoids are extracted from each audio segment.

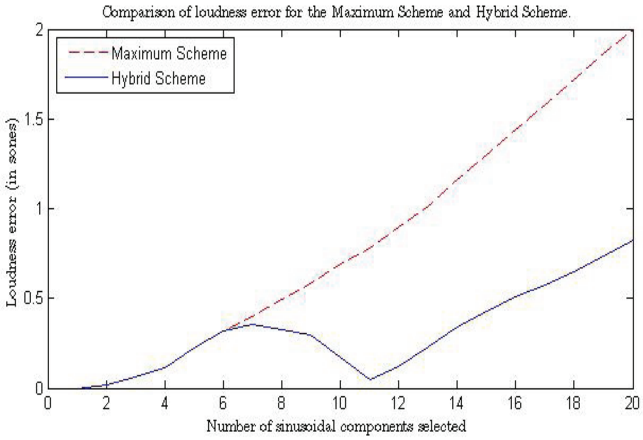

Figure 6.13: Plot of Loudness error for maximum and hybrid scheme for different number of components.

The accuracy of the sinusoidal component selection using the proposed estimation scheme is measured relative to those selected when a complete loudness estimation procedure is employed. That is, we evaluate whether the proposed method selects the same sinusoids as the full estimation method. To that end, Table 6.1 lists the percentage of sinusoids that are in common with the two methods. In essence, this is a metric of how good this approximation is. We tabulate results for different types of audio segments corresponding to four different scenarios. It can be seen from Table 6.1I that in most cases the proposed low complexity algorithm selects a set of sinusoids that is 90 % similar on the average to the set obtained from the full estimation (high complexity) algorithm. In Table 6.2, we present the CPU execution times for sinusoidal selection based on the proposed low complexity hybrid loudness estimation scheme when compared relative to the reference (high complexity) loudness estimation procedure. All simulations were performed using MATLAB (v7.5) on an Intel 2 GHz dual-core processor with 2 GB RAM. Results indicate that the proposed algorithm achieves a significant reduction in execution time. In Fig. 6.13, we compare the error in the loudness estimates between the “Max" scheme and the Hybrid scheme after each sinusoid is selected. It can be observed from Fig. 6.13 that the hybrid scheme is associated with a lower average loudness error across all iterations.

Table 6.1: Sinusoidal Component Selection Accuracy

	k=5	k=10	k=15	k=20
Pop	97 %	95 %	90 %	88 %
Solo Instruments	97 %	93 %	86.5 %	84.5 %
Orchestra	96.5 %	94.5 %	91.5 %	89.2 %
Speech	94.2 %	86.8 %	83.2 %	82.67 %

Table 6.2: Computational time comparison

k	CPU execution time (in seconds)	
	Reference Scheme	Hybrid Scheme
5	8.3	0.15
10	17.9	1.1
15	27.35	2.8
20	36.25	4.9

6.6 Summary

In this chapter, we proposed a sinusoidal selection algorithm based on two different approaches. The first approach formulates the problem as a convex optimization problem. The second approach describes the proposed hybrid loudness estimation scheme for use in sinusoidal component selection. It should be noted that the solution obtained by the greedy algorithm is acceptable as far as perceptual saliency is concerned. The only issue with the greedy algorithm is the high computational complexity.

In this chapter, we described the proposed perceptual sinusoidal selection algorithm by formulating it using convex optimization techniques. In particular, the following three techniques are described: i) the ALE scheme, ii) the MLE scheme, iii) the LLE scheme. The ALE scheme selects $> 90\%$ similar sinusoids as that of the greedy approach while operating at a much lower computationally complexity. The LLE scheme attains lower residual loudness error compared to the ALE, MLE or the greedy techniques. This indicates that the resulting set of sinusoidal selections from the LLE scheme are more optimal than that obtained from the other techniques. We note that the proposed algorithms can further benefit by incorporating the detector pruning algorithm proposed in [38] for evaluating excitation and loudness patterns.

Moreover, in contrast to existing perceptual coding techniques that focus on signal masking, we proposed a technique that uses auditory excitation level matching for audio coding. More specifically, in the context of peak-picking of sinusoidal transform coding, we propose an optimized selection criteria that minimizes the error in the excitation pattern between the original and the reconstructed signal. Our results indicate that the proposed algorithm outperforms existing maximum SMR sinusoid selection algorithms, while operating at a much lower complexity than existing excitation-pattern matching algorithms. Future work in this area will focus on embedding excitation pattern criteria in coding applications, speech enhancement, and audio classification. Further, the existing algorithms can be tailored to include real-time convex optimization solvers.

Chapter 7

Conclusions

In this dissertation, we addressed the problems associated with directly embedding an auditory model in an objective function. The main idea behind embedding an auditory model in an objective function is to process signals according to the properties of human perception. In particular, two different paradigms were investigated to solve perceptual distortion functions in a computationally efficient manner. The first paradigm involved repeatedly employing an auditory model over the entire search space of time/frequency domain candidate solutions. The second paradigm involved transforming the signals into their equivalent auditory patterns (either excitation or loudness patterns) and solve for an optimal solution. This required inverse auditory mapping techniques to map the auditory pattern estimate to its time or frequency domain representation. This dissertation described the development of efficient techniques to embed perceptual models following either of the two paradigms.

The first set of proposed algorithms reduce the computational complexity associated with the auditory model evaluation stages. To that end, a frequency and detector pruning approach [38] and a hybrid algorithm specifically for use with sinusoidal signals were proposed. The main idea behind the frequency and detector pruning approach is to reduce the number of frequency components and detector locations in a manner consistent with human perception. Experimental results indicate that the pruning approach achieves up to an 80 % and 88% reduction in the number of frequency components and detector locations respectively. It also results in up to 97 % reduction in the computational complexity of the auditory filter shape and excitation pattern evaluation stages while resulting in only $4-7\%$ average relative loudness error.

It should be noted that the performance of the loudness estimation algorithm is dependant on the accuracy of the spectral estimation process. Therefore, in applications that involve re-synthesis of signals, it is desired to maintain a good tradeoff between time and frequency resolution. For example, adaptive windows or multi-resolution windows can be used to improve the spectral estimation accuracy so that the resulting loudness estimates are accurate.

The main idea behind the hybrid algorithm is to make use of the auditory model stages only when masking phenomenon is suspected. If masking is not expected then a lookup table approach is employed for evaluating auditory patterns. The hybrid algorithm proposes an auditory pattern combining technique that combine the results from the two different auditory pattern evaluation approaches. It exploits the frequency separation between individual sinusoids and employs either a full loudness estimation process or the table lookup process depending on whether the individual sinusoids fall within the same critical band or not. The proposed hybrid algorithm was further incorporated in a perceptual sinusoidal component selection task where the objective was to select a small subset of sinusoidal components from an available set of candidate sinusoids in a perceptually relevant manner. Simulation results indicate that the hybrid algorithm resulted in 90% reduction in the computational complexity while maintaining a sinusoidal selection accuracy of $80 - 90\%$.

To solve a perceptual objective function following the second paradigm, a constrained mapping scheme was proposed that minimizes a perceptual objective function while simultaneously obtaining a time or frequency domain solution. The main idea behind the proposed technique is to overcome the inverse mapping of the auditory patterns to its corresponding time/frequency domain vector. The constrained mapping scheme is incorporated in an auditory domain based speech enhancement algorithm and a perceptual component selection task.

In the speech enhancement task, the proposed technique avoids the estimation of masked thresholds from noisy inputs unlike other perceptual speech enhancement schemes. Therefore, it is particularly more effective at the low signal-to-noise ratios. Furthermore, it attains a lower average relative loudness error compared to Wiener and spectral subtraction based technique thereby highlighting the merits of including perceptual models.

The constrained mapping scheme was incorporated in a sinusoidal selection scheme where the objective is to select a limited number of perceptually relevant sinusoids from a candidate set of sinusoids by maximizing the matching between the modeled signal's auditory pattern and the original signal's auditory pattern. Three different perceptual objective functions were tested with the proposed mapping scheme and compared to the greedy excitation pattern matching algorithm. Results indicate that the LLE minimization technique attains a lower average residual loudness error compared to the greedy approach indicating that the resulting sinusoidal selections are more optimal than that obtained from the greedy approach. The ALE and MLE minimization schemes represent computationally efficient alternatives to the greedy approach and result in $> 90\%$ similar sinusoidal selections as that of the greedy approach.

Finally, Simulink implementations of the different stages in the Moore & Glasberg auditory model are developed. These include the sound pressure normalization, outer and middle ear filtering, excitation pattern evaluation, loudness pattern evaluation, instantaneous loudness, short-term and long-term loudness evaluation blocks. These building blocks were subsequently used in the development of a number of Simulink demos. The first demo mimics the human auditory system and estimates the loudness perception associated with any incoming audio stimuli. It also obtains estimates of auditory patterns such as excitation and loudness patterns. The second demo highlights the difference between employing an energy based measure versus a loudness based measure. This is demonstrated by considering two signal with identical energies but have different loudness measures. In the third demo, a loudness control application is developed that controls the output loudness according to a preset loudness that is desired at the output.

7.1 Future Directions

Current speech enhancement algorithms include human perceptual characteristics in a heuristic manner. More direct methods of including perceptual properties in speech enhancement algorithms can be investigated. In particular, the following aspects can be investigated:

a) Study the performance of the proposed mapping technique in other speech enhancement algorithms,

b) Development of an inverse mapping technique to carry out speech enhancement in the auditory domain.

c) Use of partial loudness measure as a metric to compare the performance of different speech enhancement and noise estimation algorithms.

The significance of embedding perceptual models directly in speech enhancement algorithms can be highlighted by their performance at low signal-to-noise ratios.

Sinusoidal modeling techniques have become popular in parametric audio coding techniques. In this dissertation, sinusoidal selection has been carried out based on loudness measures. Alternatively, sinusoidal selection can be carried out based on partial loudness patterns. The partial loudness pattern predicts the loudness associated with one signal in presence of a background signal. This metric can also be used in parallel selection of sinusoids, i.e., selecting multiple sinusoids in a single iteration.

In coding applications, instead of quantizing a time or frequency domain signal, the corresponding auditory patterns can be quantized and transmitted.

This requires development of robust mapping techniques from the auditory representation to its time or frequency domain representation.

REFERENCES

[1] H. Fletcher, “Auditory patterns,” Rev. Mod. Phys., vol. 12, no. 1, pp. 47–65, Jan 1940.

[2] T. Painter and A. Spanias, “Perceptual coding of digital audio,” Proceedings of the IEEE, vol. 88, no. 4, pp. 451–513, April 2000.

[3] B. C. J. Moore, B. R. Glasberg, and T. Baer, “A model for the prediction of thresholds, loudness, and partial loudness,” Journal of Audio Engineering Society, vol. 45, no. 4, pp. 224–240, April 1997.

[4] B. R. Glasberg and B. C. J. Moore, “A model of loudness applicable to time-varying sounds,” Journal of Audio Engineering Society, vol. 50, no. 5, pp. 331–342, May 2002.

[5] H. Fletcher and W. A. Munson, “Loudness, its definition, measurement and calculation,” The Journal of the Acoustical Society of America, vol. 5, no. 2, pp. 82–108, Oct 1933. [Online]. Available: http://link.aip.org/link/?JAS/5/82/1

[6] ——, “Relation between loudness and masking,” The Journal of the Acoustical Society of America, vol. 9, no. 1, pp. 1–10, July 1937. [Online]. Available: http://link.aip.org/link/?JAS/9/1/1

[7] S. S. Stevens, “Procedure for calculating loudness: Mark vi,” The Journal of the Acoustical Society of America, vol. 33, no. 11, pp. 1577–1585, Nov 1961. [Online]. Available: http://link.aip.org/link/?JAS/33/1577/1

[8] T. Dau, D. Püschel, and A. Kohlrausch, “A quantitative model of the “effective” signal processing in the auditory system. i. model structure,” The Journal of the Acoustical Society of America, vol. 99, no. 6, pp. 3615–3622, 1996. [Online]. Available: http://link.aip.org/link/?JAS/99/3615/1

[9] ——, “A quantitative model of the “effective” signal processing in the auditory system. ii. simulations and measurements,” The Journal of the Acoustical Society of America, vol. 99, no. 6, pp. 3623–3631, 1996. [Online]. Available: http://link.aip.org/link/?JAS/99/3623/1

[10] T. Irino and R. D. Patterson, “A compressive gammachirp auditory filter for both physiological and psychophysical data,” The Journal of the Acoustical Society of America, vol. 109, no. 5, pp. 2008–2022, 2001. [Online]. Available: http://link.aip.org/link/?JAS/109/2008/1

[11] E. A. Lopez-Poveda and R. Meddis, "A human nonlinear cochlear filterbank," The Journal of the Acoustical Society of America, vol. 110, no. 6, pp. 3107–3118, 2001.

[12] E. Skovenborg and S. H. Nielson, "Evaluation of different loudness models with music and speech material," in In Proceedings of 117th A.E.S Convention, Oct 2004.

[13] H. Fastl and E. Zwicker, Psychoacoustics: Facts and Models, 3rd ed. Springer, 2006.

[14] B. R. Glasberg and B. C. J. Moore, "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, vol. 53, no. 10, pp. 906–918, October 2005.

[15] R. Patterson and et.al., "Complex sounds and auditory imaging," in Auditory Physiology and Perception, Y. Cazals, K. Horner, and N. Demany, Eds., 1992, pp. 429–443.

[16] R. D. Patterson and B. C. J. Moore, Auditory filters and excitation patterns as representations of frequency resolution. Academic Press, 1986, ch. 3, pp. 123–177.

[17] E. Zwicker, "Procedure for calculating loudness of temporally variable sounds," The Journal of the Acoustical Society of America, vol. 62, no. 3, pp. 675–682, Sep 1977. [Online]. Available: http://link.aip.org/link/?JAS/62/675/1

[18] A. Spanias, T. Painter, and V. Atti, Audio Signal Processing and Coding. Wiley-Interscience, Feb 2007.

[19] J. MPEG, Information technologyŮCoding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/sŮPart 3: Audio, ISO/IEC Std., 1992.

[20] L. D. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, and S. Vernon, "AC-2 and AC-3: Low-complexity transform-based audio coding," in Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction, 5 1996. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7132

[21] ISO/IEC, Information technologyŮGeneric coding of moving pictures and associated audioŮPart 7: Advanced audio coding, ISO/IEC Std., 1997.

[22] A. Rix, M. Hollier, A. Hekstra, and J. Beerends, “Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment,” Journal of Audio Engineering Society, 2002.

[23] C. Colomes, M. Lever, Y.-F. Dehery, and G. Faucon, “A perceptual objective measurement system (POM) for the quality assessment of perceptual codecs,” in Audio Engineering Society Convention 96, 2 1994.

[24] B. Paillard, P. Mabilleau, S. Morissette, and J. Soumagne, “PERCEVAL: perceptual evaluation of the quality of audio signals,” vol. 40, no. 1-2, pp. 21–31, 1992.

[25] W. Yang, M. Benbouchta, and R. Yantorno, “Performance of the modified bark spectral distortion as an objective speech quality measure,” in Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, vol. 1, may 1998, pp. 541 –544 vol.1.

[26] Y. Hu and P. Loizou, “Incorporating a psychoacoustical model in frequency domain speech enhancement,” IEEE Signal Processing Letters, vol. 11, no. 2, pp. 270–273, Feb. 2004.

[27] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” IEEE Trans. on Speech and Audio Processing, vol. 7, no. 2, pp. 126–137, 1999.

[28] P. Loizou, “Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum,” IEEE Trans. on Speech and Audio Processing, vol. 13, no. 5, pp. 857–869, 2005.

[29] Y. Hu and P. Loizou, “A perceptually motivated approach for speech enhancement,” IEEE Transactions on Speech and Audio Processing, vol. 11, no. 5, pp. 457–465, Sept. 2003.

[30] ——, “A generalized subspace approach for enhancing speech corrupted by colored noise,” IEEE Transactions on Speech and Audio Processing, vol. 11, no. 4, pp. 334–341, July 2003.

[31] T. Verma and T. Meng, “Sinusoidal modeling using frame-based perceptually weighted matching pursuits,” in Acoustics, Speech, and Signal Process-

ing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on, vol. 2, Mar 1999, pp. 981–984.

[32] R. Heusdens, R. Vafin, and W. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," IEEE Signal Processing Letters, vol. 9, no. 8, pp. 262–265, Aug 2002.

[33] T. Painter and A. Spanias, "Perceptual segmentation and component selection for sinusoidal representations of audio," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 2, pp. 149–162, March 2005.

[34] H. Purnhagen, N. Meine, and B. Edler, "Sinusoidal coding using loudness-based component selection," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02), vol. 2, 2002, pp. 1817–1820.

[35] V. Berisha and A. Spanias, "Wideband speech recovery using psychoacoustic criteria," EURASIP Journal on Audio, Speech and Music Processing, vol. 2007, p. 18, 2007.

[36] V. Atti and A. Spanias, "Perceptually-motivated all-pole modeling," IEEE Signal Processing Letters, vol. 16, no. 8, pp. 695–698, Aug. 2009.

[37] H. Krishnamoorthi, V. Berisha, and A. Spanias, Method and system for determining an auditory pattern of an audio segment. Patent Pending, filed by ASU, May 2010.

[38] H. Krishnamoorthi, A. Spanias, and V. Berisha, "A frequency/detector pruning approach for loudness estimation," IEEE Signal Processing Letters, vol. 16, no. 11, pp. 997–1000, Nov. 2009.

[39] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," IEEE Trans. on Information Theory, vol. 38, no. 2, pp. 824–839, Mar 1992.

[40] G. Kubin and W. B. Kleijn, "Multiple-description coding (MDC) of speech with an invertible auditory model," in Proc. IEEE Workshop on Speech Coding, 20–23 June 1999, pp. 81–83.

[41] G. Kubin and W. Bastiaan Kleijn, "On speech coding in a perceptual domain," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '99, vol. 1, 15–19 March 1999, pp. 205–208.

[42] J. H. Plasberg and W. B. Kleijn, “The sensitivity matrix: Using advanced auditory models in speech and audio processing,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 1, pp. 310 –319, Jan. 2007.

[43] H. Krishnamoorthi, V. Berisha, and A. Spanias, “A low-complexity loudness estimation algorithm,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008, March 31 2008–April 4 2008, pp. 361–364.

[44] H. Krishnamoorthi, V. Berisha, A. Spanias, and H. Kwon, “Low-complexity sinusoidal component selection using loudness patterns,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2009, 19–24 April 2009, pp. 301–304.

[45] H. Krishnamoorthi, A. Spanias, V. Berisha, H. Kwon, and H. Thornburg, “An auditory domain based speech enhacement algorithm,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP ’10, 2010.

[46] H. Krishnamoorthi, V. Berisha, and A. Spanias, “Sinusoidal component selection based on excitation level matching,” to be submitted to IEEE Trans. Speech Audio and Language Processing, 2011.

[47] H. L. F. Helmholtz, On the Sensations of Tone as a Physiological Basis for the Theory of Music, 4th ed. Longmans, Green, and Co., 1912.

[48] G. V. Bekesy, Experiments in Hearing. McGrawŰHill, 1960.

[49] E. Terhardt, “Calculating virtual pitch.” Hear. Res., vol. 1, no. 2, pp. 155–182, March 1979.

[50] E. Zwicker and H. Fastl, “On the development of the critical band,” The Journal of the Acoustical Society of America, vol. 52, no. 2B, pp. 699–702, 1972. [Online]. Available: http://link.aip.org/link/?JAS/52/699/1

[51] E. Zwicker and E. Terhardt, “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency,” The Journal of the Acoustical Society of America, vol. 68, no. 5, pp. 1523–1525, 1980. [Online]. Available: http://link.aip.org/link/?JAS/68/1523/1

[52] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” Hear. Res., vol. 47, pp. 103–138, Aug 1990.

[53] R. D. Patterson, “Auditory filter shapes derived with noise stimuli,” The Journal of the Acoustical Society of America, vol. 59, no. 3, pp. 640–654, 1976. [Online]. Available: http://link.aip.org/link/?JAS/59/640/1

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

[55] B. C. J. Moore, “Masking in the human auditory system,” in Collected Papers on Digital Audio Bit-Rate Reduction, Audio Engineering Society, 1996.

[56] ——, “Characterization of simultaneous, forward and backward masking,” in AES 12th International Conference, May 1993, pp. 22–33.

[57] R. L. Wegel and C. E. Lane, “The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear,” Phys. Rev., vol. 23, no. 2, pp. 266–285, Feb 1924.

[58] J. Herre and J. D. Johnston, “Enhancing the performance of perceptual audio coders by using temporal noise shaping,” in 101st Audio Engineering Society Convention, 1996.

[59] J. Herre, “Temporal noise shaping, quantization and coding methods in perceptual audio coding: A tutorial introduction,” in Audio Engineering Society 17th International Conference, 1999, pp. 312–325.

[60] W. Jesteadt, S. P. Bacon, and J. R. Lehman, “Forward masking as a function of frequency, masker level, and signal delay,” The Journal of the Acoustical Society of America, vol. 71, no. 4, pp. 950–962, 1982. [Online]. Available: http://link.aip.org/link/?JAS/71/950/1

[61] E. Zwicker and B. Scharf, “A model of loudness summation,” Psychological Review, vol. 72, pp. 3–26, 1965.

[62] B. C. J. Moore, An Introduction to the Psychology of Hearing, 5th ed. Academic Press, 2003.

[63] S. S. Stevens, “A scale for the measurement of a psychological magnitude: loudness,” Psychological Review, vol. 43, no. 5, pp. 405–416, Sep. 1936.

[64] B. C. J. Moore and B. R. Glasberg, “A revision of zwicker’s loudness model,” Acustica - Acta Acustica, vol. 82, pp. 335–345, 1996.

[65] A. Harma and K. Palomaki. (1999, Oct.) Hutear - matlab toolbox for auditory modeling. [Online]. Available: http://www.acoustics.hut.fi/software/HUTear/

[66] R. D. Patterson, M. H. Allerhand, and C. Giguère, “Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform,” The Journal of the Acoustical Society of America, vol. 98, no. 4, pp. 1890–1894, 1995. [Online]. Available: http://link.aip.org/link/?JAS/98/1890/1

[67] R. P. Hellman, “Rationale for a new loudness standard,” The Journal of the Acoustical Society of America, vol. 119, no. 5, pp. 3291–3291, May 2006. [Online]. Available: http://link.aip.org/link/?JAS/119/3291/3

[68] ISO389-7, Acoustics-Reference zero for the calibration of audiometric equipment. Part 7: Reference threshold for hearing under free-field and diffuse-field conditions, International Organization for Standardization Std., 1996.

[69] “SQAM-sound quality assessment material: Recordings for subjective tests,” European Broadcasting Union, Tech. Doc. 3253, 1988.

[70] B. C. J. Moore and B. R. Glasberg, “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” The Journal of the Acoustical Society of America, vol. 74, no. 3, pp. 750–753, Sep 1983. [Online]. Available: http://link.aip.org/link/?JAS/74/750/1

[71] P. C. Loizou, Speech Enhancement Theory and Practice. CRC Press, 2007.

[72] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, Apr. 1979, pp. 208–211.

[73] T. Quatieri and R. Dunn, “Speech enhancement based on auditory spectral change,” in IEEE ICASSP, 2002, pp. 257–260.

[74] I. Cohen and B. Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” IEEE Signal Processing Letters, vol. 9, no. 1, pp. 12–15, Jan 2002.

[75] Y. Hu and P. Loizou, “Subjective evaluation and comparison of speech enhancement algorithms,” Speech Communication, vol. 49, pp. 588–601, 2007.

[76] “The digital theater systems (dts),” web-page: www.dtsonline.com.

[77] F. Pereira and T. Ebrahimi, The MPEG-4 Book. IMSC Press Multimedia Series, Prentice Hall PTR, 2002.

[78] R. McAulay and T. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 34, no. 4, pp. 744–754, Aug 1986.

[79] S. Mehrotra, “On the implementation of a primal-dual interior point method,” SIAM Journal on Optimization, vol. 2, pp. 575–601, 1992.

[80] J. Markel and A. Gray, Linear prediction of speech. Springer-Verlag, 1976.

[81] A. Biswas and A. den Brinker, “Laguerre-based linear prediction using perceptual biasing,” in Signals, Systems and Computers, 2006. ACSSC ’06. Fortieth Asilomar Conference on, 2006.

[82] A. den Brinker, V. Voitishchuk, and S. van Eijndhoven, “IIR-based pure linear prediction,” IEEE Transactions on Speech and Audio Processing, vol. 12, no. 1, pp. 68 – 75, Jan. 2004.

[83] V. Atti and A. Spanias, “Rate determination based on perceptual loudness,” in IEEE Proceedings of ISCAS, vol. 2, May 2005, pp. 848–851.

[84] ——, “Speech analysis by estimating perceptually relevant pole locations,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’05), vol. 1, March 18–23, 2005, pp. 217–220.

[85] P. V. Candes and et. al., “A sinusoidal modeling approach based on perceptual matching pursuits for parametric audio coding,” in 118th AES Convention, May 2005.

[86] R. Cassidy and J. Smith, "Efficient time-varying loudness estimation via the hopping goertzel dft," in 50th Midwest Symposium on Circuits and Systems (MWSCAS), Aug. 2007, pp. 421–422.

[87] Y. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," in Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on, 1991, pp. 961–964.

[88] E. B. George and M. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," Journal of Audio Engineering Society, pp. 497–516, June 1992.

[89] R. Heusdens and J. Jensen, "Jointly optimal time segmentation, component selection and quantization for sinusoidal coding of audio and speech," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05), vol. 3, 18–23 March 2005, pp. 193–196.

[90] S. Levine and J. Smith, "A sines+ transients+ noise audio representation for data compression and time/pitch scale modifications," in Proc. 105th Conv. Aud. Eng. Soc., Sep. 1998.

[91] J. W. Shin and N. S. Kim, "Perceptual reinforcement of speech signal based on partial specific loudness," Signal Processing Letters, IEEE, vol. 14, no. 11, pp. 887–890, Nov. 2007.

[92] A. S. Spanias, "Speech coding: a tutorial review," Proceedings of the IEEE, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.

[93] A. S. Spanias and P. C. Loizou, "Mixed fourier/walsh transform scheme for speech coding at 4.0 kbit/s," IEE Proceedings I Communications, Speech and Vision, vol. 139, no. 5, pp. 473–481, Oct. 1992.

[94] P. J. Wolfe and S. J. Godsill, "Perceptually motivated approaches to music restoration," Journal of New Music Research, vol. 30, no. 1, pp. 83–92, 2001.

[95] [Online]. Available: http://www.britannica.com/EBchecked/topic/175622/ear

[96] R. Lyon, A. Katsiamis, and E. Drakakis, "History and future of auditory filter models," in Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), May 30-Jun. 2 2010, pp. 3809 –3812.

[97] R. F. Lyon, M. Rehn, S. Bengio, T. C. Walters, and G. Chechik, “Sound retrieval and ranking using sparse auditory representations,” Neural Computation, vol. 22, no. 9, pp. 2390–2416, 2010.

[98] S.-W. Huang, T.-H. Tsai, and L.-G. Chen, “A low complexity design of psycho-acoustic model for mpeg-2/4 advanced audio coding,” IEEE Transactions on Consumer Electronics, vol. 50, no. 4, pp. 1209 – 1217, Nov. 2004.

[99] T.-H. Tsai, J.-H. Luo, S.-W. Huang, and S.-C. Li, “Low complexity architecture design of mdct-based psychoacoustic model for mpeg 2/4 aac encoder,” in Proc. of IEEE International Symposium on Circuits and Systems, 2006, pp. 141–144.

[100] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, “Practical gammatone-like filters for auditory processing,” EURASIP J. Audio Speech Music Process., vol. 3, pp. 1–15, October 2007. [Online]. Available: http://dx.doi.org.ezproxy1.lib.asu.edu/10.1155/2007/63685

[101] A. Katsiamis, E. Drakakis, and R. Lyon, “A biomimetic, 4.5 mu w, 120+ db, log-domain cochlea channel with agc,” IEEE Journal on Solid-State Circuits, vol. 44, no. 3, pp. 1006 –1022, March 2009.

APPENDIX A

MOORE & GLASBERG AUDITORY MODEL TOOLBOX

A.1 Introduction

This appendix provides a software description of the various functions required to implement the Moore & Glasberg auditory model. Both Matlab and Simulink implementations of the auditory model stages have been developed. In this appendix, the Simulink implementations are described while references to the corresponding Matlab functions will be highlighted. There are a number of advantages with implementing the functions in Simulink:

1. It allows one to interact with the simulations at run-time.
2. It provides a more intuitive interface to visualize blocks and the signal flow between them. In addition, hierarchical blocks can be designed to hide the details of implementation at the top level.
3. It is easier to generate embedded C-code for a number of targets (e.g., DSP processors).
4. It is relatively simple to generate fixed-point and floating-point models from a base model.

The Simulink implementations are usually developed from built-in blocks present in the Simulink Library. These basic blocks can be configured to operate at different word lengths. Also, it is easier to generate C-implementations of these Simulink models than it is from their Matlab implementations. With these advantages in mind, the Moore and Glasberg auditory model stages were developed in Simulink. The Simulink models can later be modified to account for different word lengths or different target processors. This saves development time as it is easier to test the performance of these models in Simulink and later port them to a fixed/floating-point DSP processor. On the other hand, the Matlab implementations require rewriting existing code repeatedly whenever word length or the target processor changes.

The Simulink models were developed based on a strong need to demonstrate the perceptual aspects of including auditory models in speech/audio applications. For example, a loudness control application is developed that controls the output level of an audio signal such that it has a fixed average output loudness. In a separate demo, the difference between energy and loudness metrics are highlighted by subjective listening experiments.

A.2 Simulink Models

Sound Pressure Level Normalization :

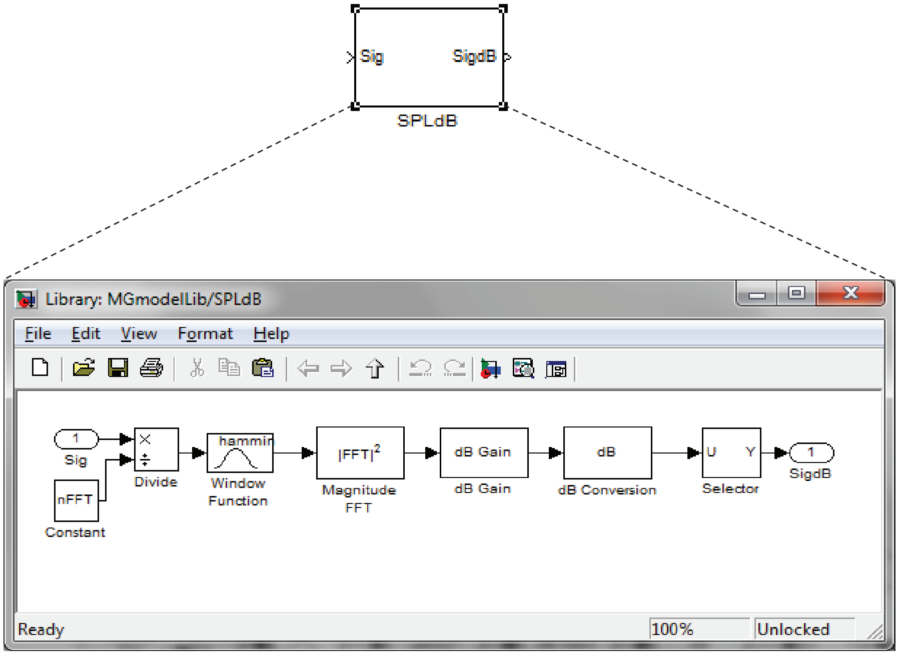

Figure A.1: Sound pressure level normalization.

This block operates on short segments of input audio on a running basis. Each segment is normalized by the FFT length as illustrated by the Divide block in Figure A.1. The Window function applies any chosen window (e.g., hamming window in this case) to the input audio segment. The Magnitude FFT block calculates a squared magnitude of the fast Fourier transform of the windowed segment. All auditory models require the sounds to be referenced to an assumed sound pressure level. In this case, the dB Gain block assumes a playback level of 90 dB as this reference SPL. The dB Conversion block converts these spectral magnitudes to dB units. Since the FFT spectrum is symmetric for real-valued signals, the Selector block is employed to select only the first half of spectral magnitudes. It should be noted that the DC component is omitted from this first half as the auditory models are invariant to DC component. The output of this block consists of spectral magnitudes expressed in dB SPL units. The corresponding Matlab function that carries out the same function is JNDSpectralAnalysis.m.

Outer and Middle Ear Filtering :

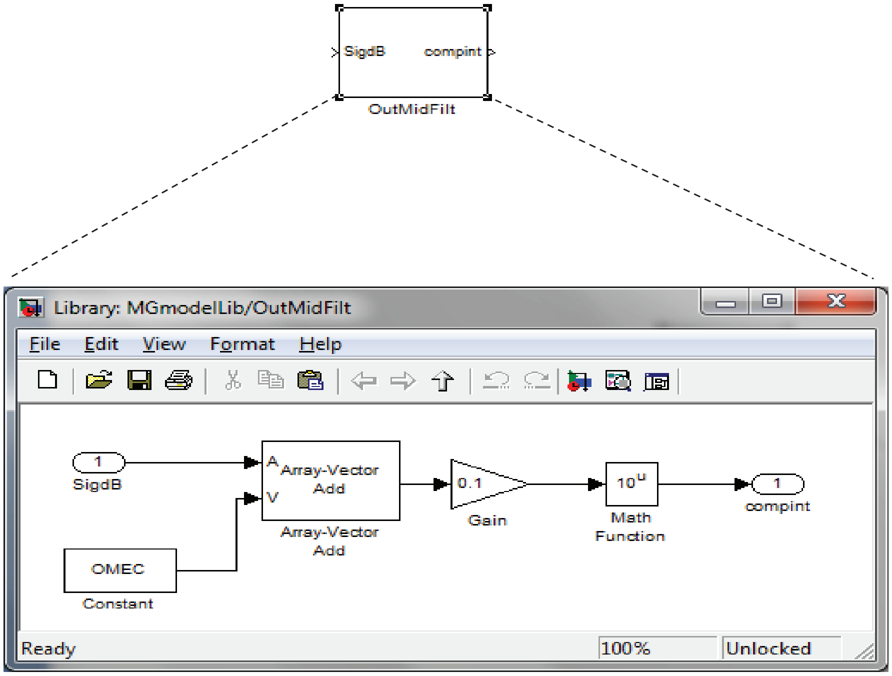

Figure A.2: Outer and Middle ear filtering.

This block performs the combined function of outer and middle ear filtering. The outer and middle ear frequency response is pre-computed and stored in the OMEC variable as shown in Figure A.2. The filtering operation is carried out by the Array-Vector add block that performs vector addition of the two inputs. The vectors are added because they are both represented in dB units. The Gain and the Math function together convert the resulting filtered dB signal magnitudes to linear power units. The outmidfilter.mat file contains the combined outer and middle ear response.

Excitation Pattern Evaluation :

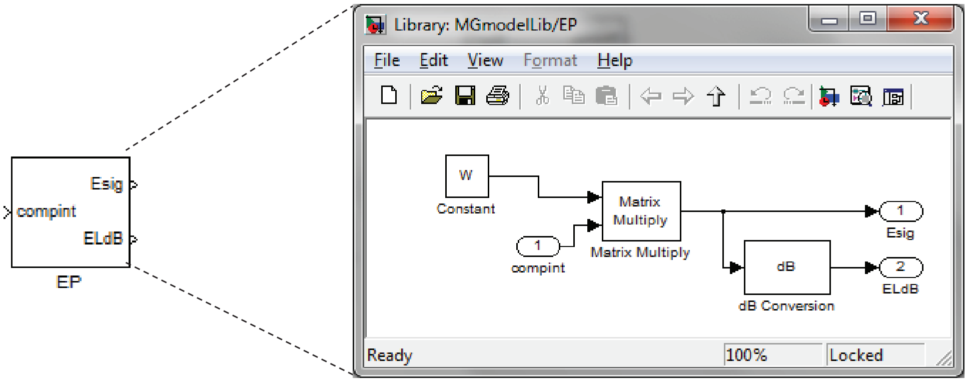

Figure A.3: Excitation Pattern evaluation

The excitation pattern evaluation block is implemented as a matrix multiplication operation as shown in Figure A.3. The output signal from the outer and middle ear filtering stage is multiplied with W. Here W is a $D \times N$ matrix where D represents the number of detector locations and N denotes the number of input spectral components. The W matrix is pre-computed and contains auditory filter magnitudes. The resulting quantity is called the excitation pattern and is also converted to dB units as shown in Figure A.3. The excitcal.m and erbintensty.m Matlab functions together evaluate the auditory filter magnitudes and the excitation patterns.

Loudness Pattern Evaluation :

The loudness pattern block implements (3.11) using Simulink blocks. The constants c, $alpha$ and A are pre-defined in the initialize.m Matlab function. The quantity $A = 2E_{THRQ}$ is a frequency dependent threshold that is pre-determined,

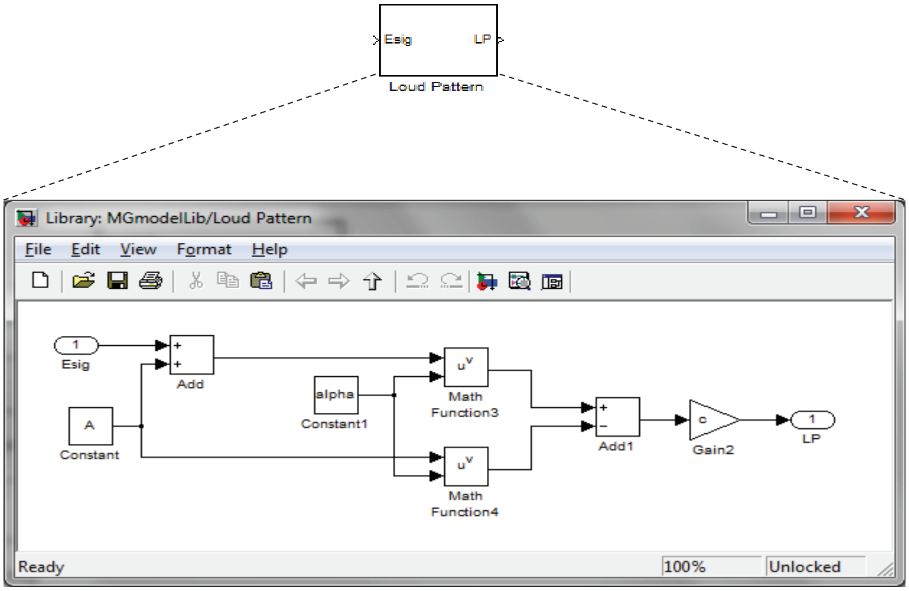

Figure A.4: Loudness pattern evaluation

whereas $c = 0.047$ and $alpha = 0.2$ are numerical constants. The output of this block represents the specific loudness pattern, i.e., the loudness density per ERB.

Instantaneous Loudness Calculation :

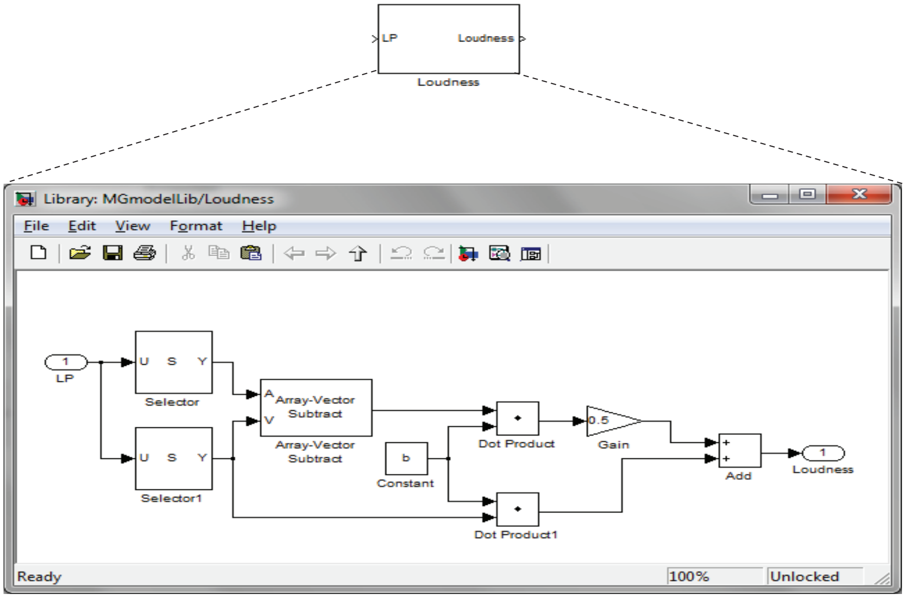

Figure A.5: Instantaneous loudness evaluation.

The instantaneous loudness block calculates the area under the specific loudness pattern. To accommodate arbitrary detector resolutions, the total loudness is calculated by approximating the area under the loudness pattern with rectangular and triangular regions. More specifically, the area is calculated by evaluating the following mathematical operation:

$$L = \sum_{k=1}^{D-1} (SP(k)(d_{k+1} - d_k) + 0.5(SP(k+1) - SP(k))(d_{k+1} - d_k). \tag{A.1}$$

The first term in the summation calculates the area under the rectangle between successive detector locations d_{k+1} and d_k. The second term in the summation improves the approximation by calculating the area of a triangle that is fitted above the rectangular region but below the actual loudness pattern. The area of the fitted triangle is added or subtracted depending on whether it is a falling or a rising edge. For example, if $SP(k+1) - SP(k) \geq 0$ then, it is a rising edge and the area of the triangle is added to that of the rectangle and vice versa for $SP(k+1) - SP(k) < 0$. The corresponding Matlab code that implements this functionality can be found in sploudarea.m function.

Short – term and Long – term Loudness Calculation :

This block calculates the short-term loudness associated with time-varying audio signals. The short-term loudness is calculated according to (3.15) and takes into account temporal masking properties. It is modeled either as an attack or a release effect depending on whether the instantaneous loudness is greater or lesser than the short-term loudness obtained at the previous instance. This is modeled using an If block that compares the magnitudes of the instantaneous loudness and short-term loudness as shown in Figure A.6. The instantaneous loudness is obtained from the input port whereas the short-term loudness is obtained through a feedback path with a delay element z^{-1} from the output port. The If Action Subsystems implement one of the cases in (3.15) depending on whether it is an attack or a release. The constants sa and sr in the action subsystems denote the attack and release parameters.

The long-term loudness is obtained in the same manner and therefore the same Simulink blocks are used. The only difference is a different attack parameter la and release parameter lr which model long-term memory effect.

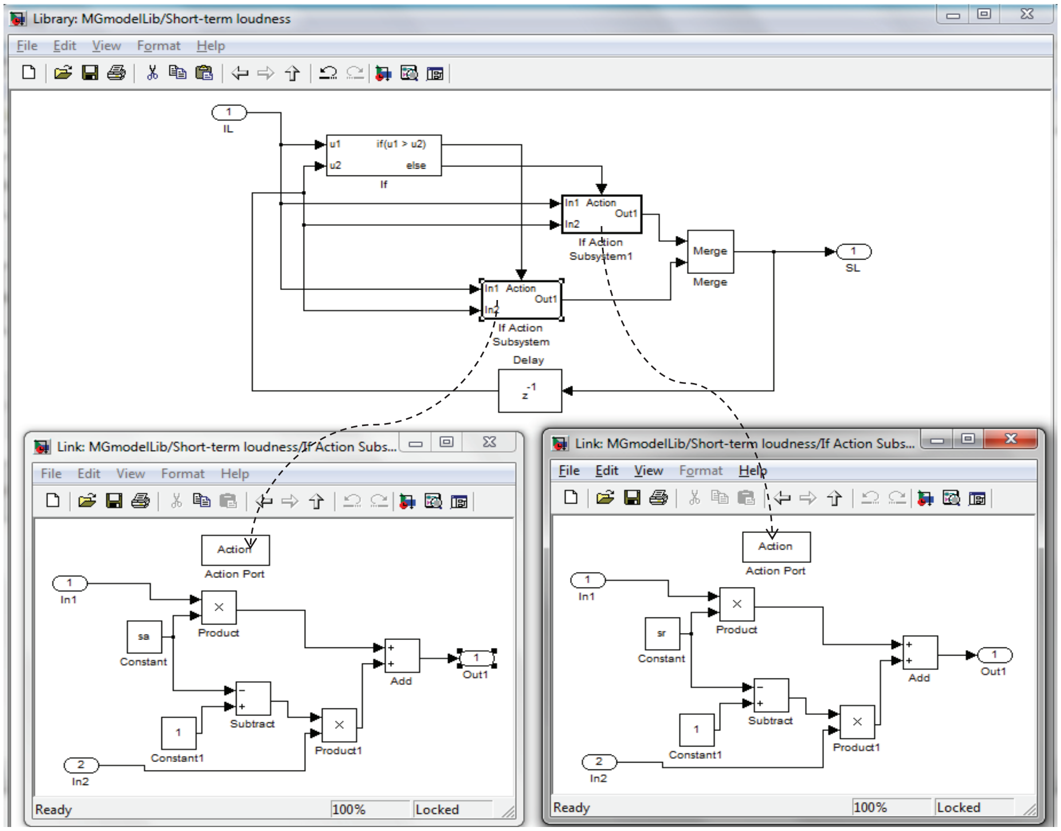

Figure A.6: Short-term loudness evaluation

ERB Bandwidth Calculation :

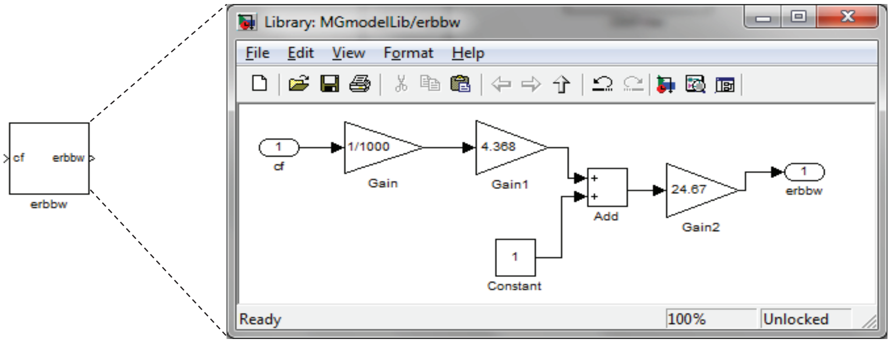

Figure A.7: ERB Bandwidth calculation.

This block calculates the critical bandwidth at any center frequency cf (in

Hz). The critical bandwidth is measured in terms of an equivalent rectangular bandwidth instead of the earlier Bark bandwidths. The following mathematical relationship is implemented using Simulink blocks as shown in Figure A.7:

$$CB(f) = 24.67(4.368\frac{cf}{1000} + 1). \tag{A.2}$$

The corresponding Matlab code that implements this functionality is erbbandwidth.m

ERB to Frequency Mapping :

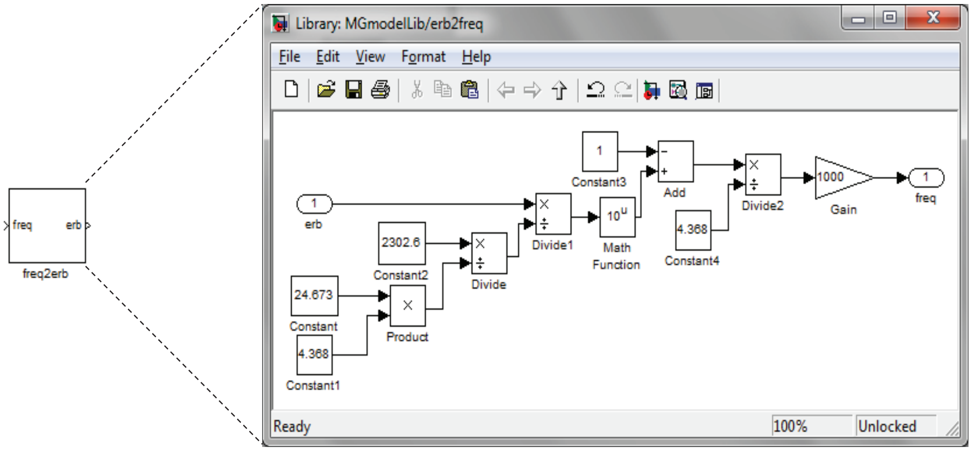

Figure A.8: ERB to frequency mapping.

This block converts ERB number *erb* to its corresponding frequency *freq* in Hz. In particular, it implements the following equation using Simulink blocks.

$$erb \text{ (in ERB units)} = 21.4\log_{10}(4.37freq/1000 + 1) \tag{A.3}$$

The corresponding Matlab code that implements this functionality is erbtofreq.m.

Frequency to ERB Mapping :

This block converts frequency *freq* in Hz to its equivalent ERB number *erb*. In particular, it implements the following equation using Simulink blocks.

$$erb \text{ (in ERB units)} = 21.4\log_{10}(4.37freq/1000 + 1) \tag{A.4}$$

The corresponding Matlab code that implements this functionality is freqtoerb.m.

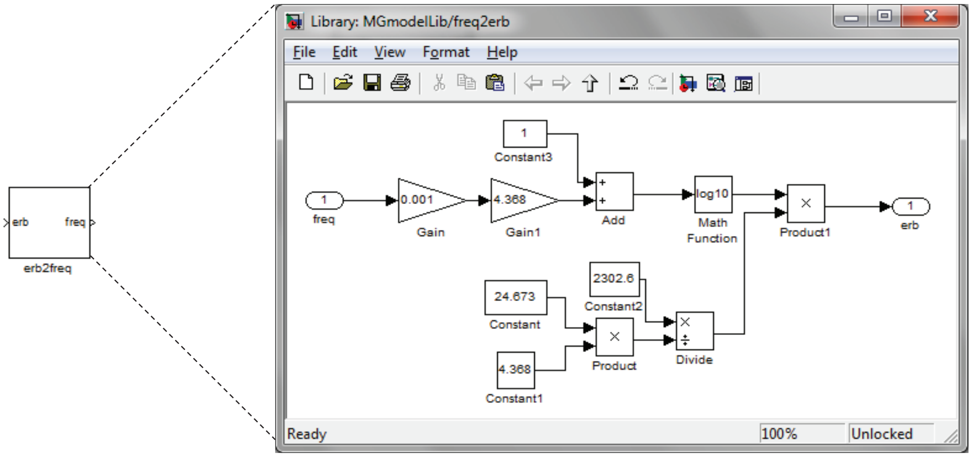

Figure A.9: Frequency to ERB mapping.

APPENDIX B

SIMULINK DEMOS

In this appendix, a number of demos are built using the Simulink blocks developed in Appendix A. These demos make use of the Moore & Glasberg auditory model stages to incorporate perceptual characteristics.

B.1 Auditory pattern evaluation and Loudness Estimation

This demo evaluates the excitation pattern, the loudness pattern, the instantaneous loudness, the short-term and long-term loudness quantities associated with an audio signal. Furthermore, it allows one to configure the model parameters to account for different frame lengths, sampling frequencies, FFT lengths and detector resolutions. These parameters can be changed in the initialize.m file that executes before evaluating the auditory model stages.

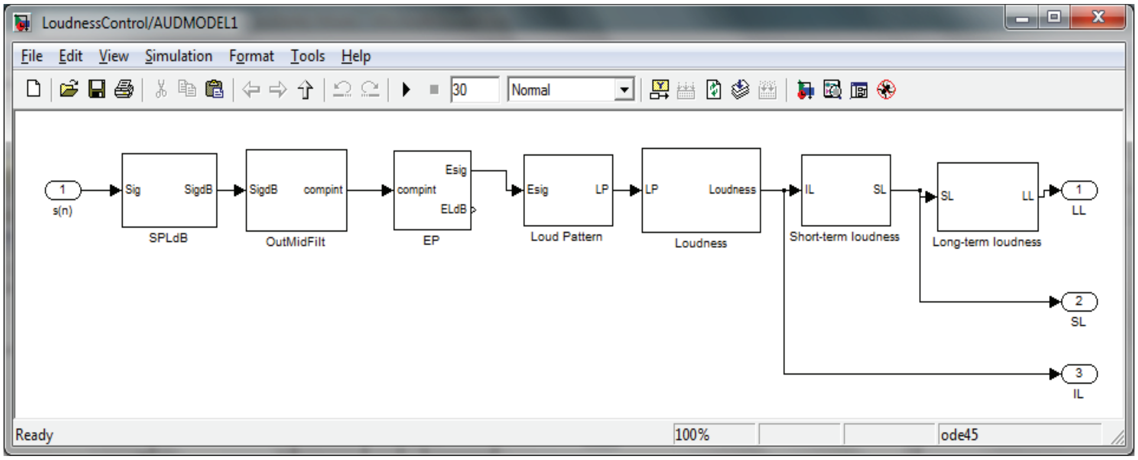

Figure B.1: Simulink implementation of Moore & Glasberg auditory model stages.

The Simulink model shown in Figure B.1 obtains the various auditory pattern outputs and loudness measures. It consists of the following sub-blocks:

- SPL dB normalization (implemented by the SPLdB sub-block),
- Outer and Middle ear filtering (implemented by the OutMidFilt sub-block)
- Excitation pattern evaluation (implemented by the EP sub-block)
- Loudness pattern evaluation (implemented by the Loud Pattern sub-block)

- Instantaneous loudness evaluation (implemented by the Loudness sub-block)
- Short-term and long-term loudness evaluation (implemented by the short-term loudness and long-term loudness sub-blocks)

B.2 Energy vs. Loudness

The objective behind this demo is to highlight the fact that two signals with identical energies can have different loudness measures. In other words, there is a difference between processing signals according to its energy content than according to its loudness measure. To illustrate this, a subjective listening experiment is developed wherein subjects are presented with two different signals with identical energies through headphones. In addition, their loudness measures according to the Moore & Glasberg auditory model are evaluated.

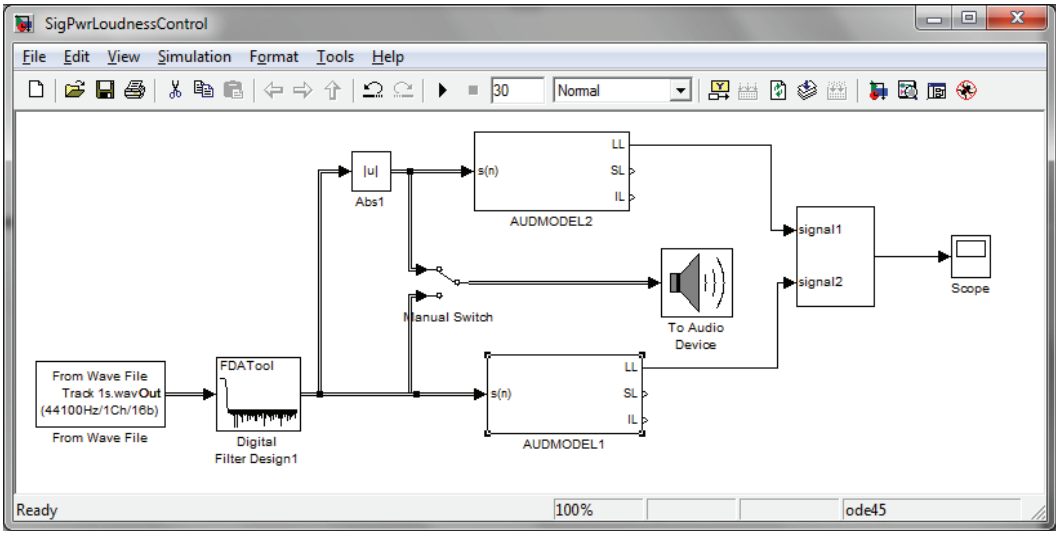

Figure B.2: Demo that illustrates the difference between energy and loudness.

This is accomplished in the following manner. An input audio signal is filtered with a low-pass filter so that the resulting signal is band-limited to a narrow frequency range. The Digital Filter Design block is used to design an appropriate filter as shown in Figure B.2. The loudness associated with the filtered signal is obtained using the auditory model block shown in Figure B.1. Moreover, the absolute value of the filtered signal is obtained using Simulink's Absolute value block. This performs the action of a rectifier and results in a signal with only positive samples. The loudness associated with the rectified signal is then evaluated in the top branch as shown in Figure B.2. It should be noted that both the rectified and non-rectified signals have identical energies. The manual switch between the two branches allows us to listen to one of the signals and judge its loudness subjectively. In addition, the auditory model outputs provide a numerical loudness measure.

In both cases, it was observed that the loudness of the rectified signal was higher than that of the non-rectified signal. This is due to the fact that the rectification operation introduces harmonics and spreads the bandwidth of the filtered signal while maintaining the same energy. The spread in bandwidth of the rectified signal is responsible for the increase in loudness as additional nerve cells along the basilar membrane are excited.

B.3 Loudness Control

In this demo, the output playback level of an audio signal is controlled so as to attain a fixed target loudness. In practice, this is accomplished through the automatic gain control circuits that measure and subsequently modify the signal's energy content in order to attain the desired target loudness. As shown in the demo in Figure B.2, there is no one-to-one correspondence between energy content and its loudness measure. This makes it difficult to control the output loudness by modifying the energy content of the signal. Therefore, there is no simple way to estimate the output loudness without resorting to subjective experiments or employing auditory models.

In this demo, we make use of the Moore & Glasberg auditory model to predict the loudness of a signal being presented to a human listener. The AUDMODEL1 block measures the loudness associated with the original signal whereas

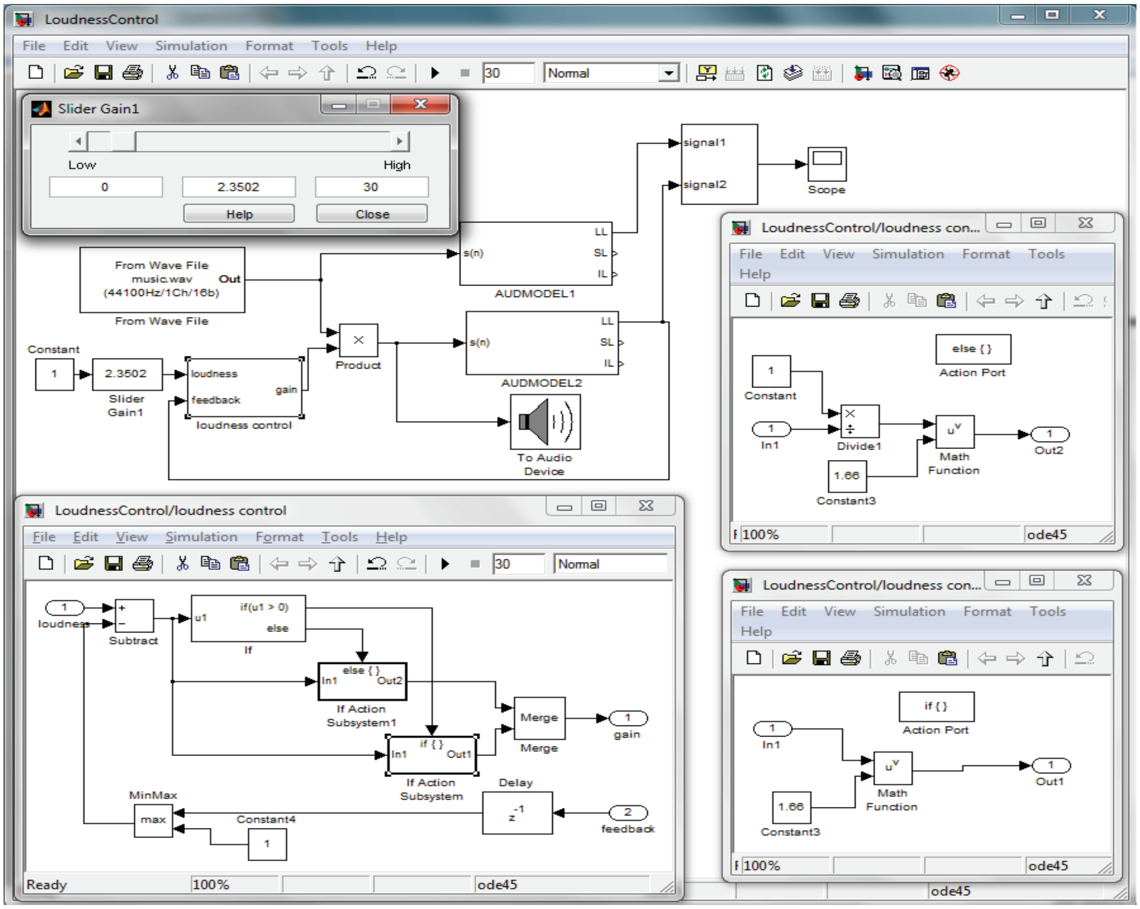

Figure B.3: Simulink model for loudness control

AUDMODEL2 block measures the loudness associated with the output signal that is adjusted for a target loudness. The desired target level for output loudness is specified by means of the slider gain block as shown in the top left corner of Figure B.3.

The loudness control block calculates an appropriate gain for each frame and applies it to the time-domain signal. The product block shown in Figure B.3 implements this operation. Although, changing the energy content of a signal has a corresponding effect on the output loudness, there is no simple way to predict the output loudness without resorting to subjective experiments or employing auditory models. In this demo, an auditory model is employed to predict the output loudness and feed back to the loudness control block.

In particular, the gain is calculated as follows: First, the difference between the desired target loudness and the actual loudness of the output signal is calculated. Here, the actual loudness corresponds to the output of the AUDMODEL2 block which is then fed back to the loudness control block. The difference in loudness should be mapped into a corresponding difference in signal intensity. Ideally, an inverse auditory mapping should be carried out. However, due to the lack of reliable inverse mapping techniques, an alternative mapping procedure is developed. In this demo, an inverse non-linear mapping based on the relationship between loudness level to loudness is employed. In particular, the following mapping is employed:

$$G = (L1 - L2)^{1/(2\log_{10}(2))} \text{ if } L1 > L2 \tag{B.1}$$

The Simulink model for loudness control is shown in Figure B.3.