A LOW-COMPLEXITY LOUDNESS ESTIMATION ALGORITHM

Harish Krishnamoorthi, Visar Berisha and Andreas Spanias

Dept of Electrical Engineering, Arizona State University, Tempe, AZ-85287
Email: [kharish, visar, spanias]@asu.edu

ABSTRACT

Audio processing applications such as rate determination, bandwidth extension, compression, and noise reduction make use of loudness metrics. Most loudness models are computationally expensive and often not suitable for real time applications. In this paper, we present a low-complexity loudness estimation algorithm for both steady and time-varying sounds. The model computes an estimate of the excitation pattern by simultaneously pruning the frequency components and detector locations. For time-varying sounds, intensity patterns across successive audio frames are further exploited to reduce complexity by partial evaluation of the excitation pattern in select critical bands. Comparative results indicate that the proposed algorithm performs consistently well for different types of audio signals at a reduced complexity.

Index Terms— audio coding, loudness, psychoacoustics.

1. INTRODUCTION

Several psychoacoustic models that make use of masking have been previously proposed in the literature [1, 2, 8, 9]. Although some of these models have been successful in audio coding applications, their use in other audio processing applications is not straightforward. A number of algorithms relying on loudness metrics have been introduced [8,11,14]. For example, loudness models have been used to develop objective and hybrid measures that predict subjective quality [8]. In addition, rate determination algorithms that use perceptual loudness have been proposed in [11]. Also hearing aid systems use loudness models to compensate for perceptual loss [14]. Recently, bandwidth extension algorithms employed loudness metrics to determine the perceptual importance of the different sub-bands [13] and reduce the side information hits. Sinusoidal analysis-synthesis algorithms based on loudness and excitation patterns have been proposed in [7] and [10]. Although the use of loudness patterns in all the aforementioned applications delivered very promising results, computational complexity for loudness estimation is very high.

A number of loudness models have been proposed in the literature. Some simple loudness estimation algorithms employ frequency weighting curves such as the A, B or C weighting [11] derived from the equal loudness curves to model the non uniform sensitivity of human hearing. These models do not account for masking and therefore perform poorly for transient and broadband sounds. Recent models rely on modeling the cochlear as a bank of auditory filters with bandwidths corresponding to critical bands [1,2,6,9]. Some of these models [2] account for both steady and time-varying sounds. The level-dependent auditory filters are either implemented in the frequency domain or in the time domain but in both cases real-time implementation is a challenge.

In this paper, we propose a new algorithm for estimating loudness starting from Glasberg’s model [2]. The proposed algorithm computes a fast estimate of the excitation pattern (EP) by selecting the most relevant frequency components and detectors in a non uniform manner. The high resolution EP is obtained by linearly interpolating the initial EP estimate. The specific loudness and total loudness are extracted from the EP estimate. A modified version of this algorithm for time-varying sounds is also proposed. This model exploits the intensity patterns across successive audio frames to further reduce the number of computation while maintaining acceptable performance. We compare the proposed model to the Moore and Glasberg process and show that the differences in the loudness estimates are minimal, while the complexity is reduced considerably.

The rest of the paper is organized as follows. Section II presents the Moore and Glasberg. This description is accompanied by analysis of performance and complexity. In section III, the proposed algorithm and its use for steady and time-varying sounds is presented. Section IV contains the experimental setup and sample results. Section V contains concluding remarks.

2. ANALYSIS OF MOORE & GLASBERG’S MODEL

An overview of the Moore and Glasberg’s model [1,2] is given in this section. The block diagram of the algorithm is shown in Fig. 1.

Fig. 1: Block diagram of the loudness model [2].

2.1 Model for steady sounds

i) The steady sounds model starts with the specification of the input spectral components $S(k)$; ii) the spectral components $S(k)$ undergo outer and middle ear correction and their combined response is modeled with an FIR filter; this stage has an $O(N)$ complexity. iii) The third stage involves the computation of an excitation pattern $E$ associated with the sound reaching the inner ear. Detectors are placed uniformly at 0.1 ERB (Equivalent Rectangular Bandwidth) units where the number of ERB units (also known as the ERB number) is defined as the number of equivalent rectangular bandwidth auditory filters that can be fitted below any frequency. Auditory filter shapes are estimated for every detector and frequency combination which is
an $O(N \times D)$ complexity. iv) Following this, the excitation pattern $E(i)$ at any detector $i$ is calculated as the sum of the response from the different auditory filters [5] which are quasi-linear at a given level but change shape with center frequency and level [3,4]. Hence this is an $O(N \times D)$ operation. v) The next stage involves transformation of the excitation pattern $E$ to a specific loudness pattern $SP$ according to the procedure described in [1]. Specific loudness represents the action of cochlea on the EP and can be thought of as a loudness density pattern, i.e. the loudness per ERB [5]. Therefore with $D$ detectors, this stage has an $O(D)$ complexity. vi) The final stage is the calculation of the area under the specific loudness pattern $SP$ in order to obtain the total instantaneous loudness $L$. This stage is associated $O(D)$ complexity.

2.2 Model for time-varying sounds
The steady sound loudness model [1] does not account for temporal masking effects. Real life signals in general are time-varying and exhibit temporal masking. In [2], a model for time-varying sounds is developed using attack and release time parameters to model temporal masking and obtain the short-term loudness of a signal. This model is similar to the steady state model. Finally, the short-term loudness $SL$ is obtained from the instantaneous loudness $L$ which is just a single operation per frame.

3. PROPOSED LOUDNESS ALGORITHM
From the previous section, we observe that the process associated with the highest complexity is the one used to evaluate auditory filter shapes and EP generation modules.

![Fig. 2: Plot showing cardinality of optimal detector set $L_o$ compared with reference set $L_r$ and estimated set $L_e$.](image)

In this section, we describe the proposed low-complexity loudness estimation algorithm for steady and time-varying sounds. A general high level overview of the proposed algorithm is shown in Fig 4.

3.1 Steady sound low-complexity algorithm
The number of frequency components ($N$) and the number of detector locations ($D$) are pruned in a manner consistent with human perception. It now remains to decide what frequency components $f_i$'s, $i \in \{1,2,...N\}$ and detector locations $d_k$'s $k \in \{1,2,...D\}$ to choose in order to estimate the model.

3.1.1 Detector Pruning
The loudness of a signal is directly related to the signal’s neural excitation pattern. The idea behind the proposed technique is to sample the excitation pattern at a sufficient number of points in order to capture its general shape. Most existing methods for generating excitation patterns place detectors uniformly along the basilar membrane. It is however sufficient to sample the EP at its maxima and minima to capture its shape; it is not necessary to sample uniformly. Let $L_r = [d_i | | d_i - d_{i-1}| = 0.1, i = 1,2,...D]$ denote the reference set of detector locations expressed in ERB units, such that they are uniformly spaced at 0.1 ERB units. Let $L_o = [d_k | \frac{\partial EP(k)}{\partial k} = 0, k = 1,2,...D]$ denote the “optimal” detector locations such that they correspond to the extrema of EP. For a linear interpolation scheme, sampling at the extrema is optimal. In our proposed algorithm, we estimate the EP at the detector locations specified by $L_r$ by linearly interpolating the EP obtained at the points specified by $L_o$, where $L_o$ is an estimate of the set $L_o$ because $L_o$ is unavailable to us. The following analysis shows that only a few detectors are sufficient for representing the EP. Firstly, an FFT 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 at least a ten fold reduction in the cardinality of set $L_r$. 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 set $L_o$ is of the order O (number of ERB units) which span the input audio spectrum, as shown in Fig. 2. In Fig. 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.

3.1.2 Frequency component pruning
It is known that multiple components falling inside the same critical band will have the same instantaneous loudness as any individual component with their combined sum of intensities [5]. This enables us to approximate the input audio spectrum inside each ERB unit with a single component of intensity equal to the combined sum of intensities within that ERB unit as shown in (1). In Fig. 3a, we show an example of a sample audio spectrum plotted on an ERB unit scale and the approximated spectrum obtained according to the equation below:

\[ \hat{S}_a(k) = \sum_{j \in \{k,k+1\}} S(j) \]  

(1)
where, $k$ is the ERB number, $j$ is the set of components in $k^{th}$ ERB. Note that $S(j)$ is the input spectral amplitude and $S_{\hat{k}}(k)$ is the approximated spectral amplitude in the $k^{th}$ ERB. However, the shape of the EP depends on the distribution of the frequency components inside each ERB unit. In order to minimize the error in the shape of the estimated EP, it is necessary to estimate the location of the approximated frequency components and detector locations inside each ERB unit.

### 3.1.3 Estimating pruned frequency locations and detector positions

Here, we describe a procedure that estimates the positions of the approximated spectral components that best structure the shape of the EP. This set of frequency components can then be directly mapped to a set of EP detectors such that they capture the extrema of the reference EP directly (without having to compute it). The specific form of the auditory filter shapes allows us to estimate the positions of maxima of the EP from the spectrum directly. The response at a particular detector is given by

$$EP(k) = \sum_i (1 + p_i * g_i) \exp(-p_i * g_i) * S_i \quad (3)$$

where $p_i$ is the slope of the auditory filter at a center frequency $f_i$, $g_i$ is the normalized deviation of the detector location $d_k$ from the frequency component location $f_i$, and $i$ represents the frequency index.

For any component, $S_i$, in the input spectrum, the maximum auditory filter response due to $S_i$ will occur at a detector location for which $|g_i| \leq 0$, as $\exp(-p_i * g_i) \geq 1$ in (3). As a result, we select the maximum $S_i$ inside each ERB unit and place a detector close to $S_i$ such that both $S_i$ and $\exp(-p_i * g_i)$ are maximized simultaneously in (3). In other words, the frequency component location corresponding to the maximum of the spectrum also corresponds to the maximum in the EP inside that ERB unit.

In order to preserve the shape of the estimated EP, in particular the positions of maxima in relation to the reference EP computed at the locations given by $L_k$, the approximated components are placed at the positions of maximal auditory filter response in each ERB. In Fig. 3b, we plot of reference excitation pattern and the estimated EP along with the positions of maximal auditory filter response.

### 3.2 Time-varying sound low complexity algorithm

In real life, one typically encounters time-varying sounds such as music or music. Therefore it is essential for any loudness estimation algorithm to be applicable to this general class of signals as well. The proposed algorithm for steady sounds requires the auditory filter shapes and EP to be computed for every audio frame and may not be suitable for real-time applications. Furthermore, using the steady-state model for time-varying signals leads to an underestimation of loudness since the temporal effects are discarded. In particular, post masking effects can extend up to 100 ms in duration after the masker has ceased to exist [6].

![Image of time-varying sound analysis](image_url)

**Fig. 3:** a) Plot of input and approximated spectrum. b) Plot of reference and estimated EP. c) Plot of reference and predicted EP.

Therefore in a typical audio frame of 30-60 ms duration considered for analysis, the loudness of the current frame is dependent largely on the post masking characteristics of the preceding frame(s). In this section, we describe the proposed low-complexity loudness estimation algorithm for time-varying sounds.

![Image of proposed loudness estimation algorithm](image_url)

**Fig. 4:** Block diagram of proposed loudness estimation algorithm for steady and time-varying sounds.

We develop the algorithm by exploiting the intensity pattern in two consecutive frames. We define the intensity pattern $I_k(i)$ as the total equivalent intensity inside each ERB. From (1) it can be seen that the intensities of the pruned components $S_{\hat{k}}(i)$ is also representative of the intensity pattern. A differential intensity pattern $D_{\hat{k}}(i)$ is computed as shown in (5) for every frame.

$$D_{\hat{k}}(i) = I_k(i) - I_{k-1}(i) \quad (5)$$

where $i$ represents the ERB number and $k$ is the frame index.

Since the auditory filters change their shapes with frequency and intensity level [3], they have to be re-computed in every frame according to the intensity pattern associated with the current frame. However, we exploit the differences in the intensity pattern across consecutive frames and re-evaluate the auditory filters shapes and the corresponding EP $\hat{E}_k(i)$ only in select ERBs where (6) is satisfied.

$$D_{\hat{k}}(i) \geq \tau_i \quad (6)$$

where $\tau_i$ is the threshold in dB in the $i^{th}$ critical band. Following this an excitation prediction step then estimates the final EP of the
current frame \( E_k(i) \) from the EP of preceding frame \( E_{k-1}(i) \) as shown in (7).

\[
E_k(i) = \begin{cases} 
   E_{k-1}(i) + DI_k(i) & , \text{ if } \Delta_i \\
   \hat{E}_k(i) & , \text{ otherwise}
\end{cases}
\]

(7)

where \( E_k(i) \) is obtained from the scaled EP in critical bands where the differential intensity pattern doesn’t exceed the threshold and is obtained from the re-evaluated EP in the other critical bands The subsequent stages in the model are similar to the steady sound algorithm as illustrated in Fig 4.

4. 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 an audio segment of 46 ms duration was 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 Average Error Energy (AEE) and Relative Error Energy (REE) in the EP as defined in (5) and (6). Also, the error in the total instantaneous loudness is obtained according to (7).

\[
AEE = 20 \cdot \log 10 \left( \frac{\sum_{i=1}^{y[D]} |\hat{E}(i) - E(i)|}{\sum_{i=1}^{y[D]} E(i)} \right)
\]

(5)

\[
REE = 20 \cdot \log 10 \left( \frac{\sum_{i=1}^{y[D]} |\hat{E}(i) - E(i)|}{E(i)} \right)
\]

(6)

\[
\text{Loudness error} = \sum_{i=\text{frames}} |\hat{L}_j - L_j|
\]

(7)

where \( \hat{E}(i), E(i) \) are the estimated and reference EP expressed in linear power units. \( \hat{L}_j, L_j \) are the estimated and reference instantaneous loudness. In Table 1, we show the average error in estimated EP and instantaneous loudness defined according to (5), (6) and (7) for different types of audio material. It can be seen that the proposed algorithm performs consistently for different types of audio signals within a tolerable error. Furthermore, we compare the complexity of the proposed algorithm with the standard approach followed in [1]-[4]. Due to the differing nature of operation in the each stage of the model, we compare the complexity in each stage separately as shown in Table 2.

Table 1: Normalized Error for 10 ms overlapping frames

<table>
<thead>
<tr>
<th></th>
<th>Rel Avg Error (dB)</th>
<th>Avg Rel Error (dB)</th>
<th>Loudness Error (sones)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Max</td>
<td></td>
</tr>
<tr>
<td>Single Instruments</td>
<td>-12.82</td>
<td>-14.84</td>
<td>0.72</td>
</tr>
<tr>
<td>Speech</td>
<td>-12.80</td>
<td>-14.73</td>
<td>0.29</td>
</tr>
<tr>
<td>Vocal</td>
<td>-12.03</td>
<td>-14.55</td>
<td>0.22</td>
</tr>
<tr>
<td>Solo Instruments</td>
<td>-12.42</td>
<td>-14.60</td>
<td>0.44</td>
</tr>
<tr>
<td>Vocal &amp; Orchestra</td>
<td>-13.4</td>
<td>-18.57</td>
<td>0.95</td>
</tr>
<tr>
<td>Orchestra</td>
<td>-11.52</td>
<td>-14.92</td>
<td>1.34</td>
</tr>
<tr>
<td>Pop Music</td>
<td>-12.58</td>
<td>-14.90</td>
<td>0.27</td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this paper, we proposed a low complexity loudness estimation algorithm for steady and time-varying sounds based on the Moore et. al [1, 2] model. For steady sounds, the proposed algorithm becomes more efficient because of pruning the less significant frequency components while preserving the energy per ERB and the number of detectors. This is done by retaining only those locations that provide the highest contribution. For time-varying sounds, a superior reduction in complexity is achieved through an excitation prediction step wherein the EP of the current frame is predicted from that of the preceding frame. The proposed algorithm was tested on a wide variety of input audio and is seen to perform consistently.

6. REFERENCES