A LOW-COMPLEXITY LOUDNESS ESTIMATION ALGORITHM

Harish Krishnamoorthi, Visar Berisha and Andreas Spanias

Department of Electrical Engineering, Arizona State University, Tempe, AZ-85287, USA
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 estimation algorithms are computationally expensive and often not suitable for real time applications. In this paper, we present a low-complexity loudness estimation algorithm applicable to both steady and time-varying sounds. The model computes an estimate of the excitation pattern by simultaneously pruning the frequency components and detector locations. 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 [6, 10, 11]. For example, loudness models have been used to develop objective and hybrid measures that predict subjective quality [7]. In addition, rate determination algorithms that use perceptual loudness have been proposed in [11]. Also hearing aid systems use loudness models to compensate for perception loss [5]. Recently, bandwidth extension algorithms employed loudness metrics to determine the perceptual importance of the different sub-bands [6] and reduce the side information bits. 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 estimation algorithms have been proposed in the literature. Some simple loudness estimation algorithms employ frequency weighting curves such as the *A, B or C weighting* [9] 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 cochlea as a bank of auditory filters with bandwidths corresponding to critical bands [1,2,5,9]. 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.

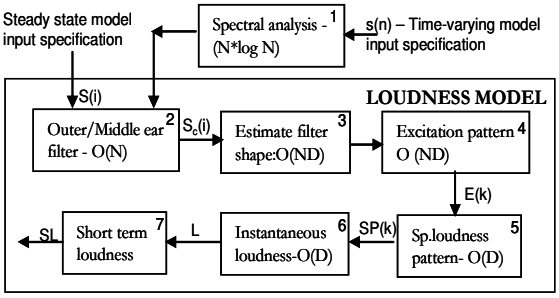

Fig. 1: Block diagram of the loudness model [2].

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 component locations 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. We compare the proposed algorithm 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 model. This description is accompanied by an analysis of performance and complexity. The proposed algorithm applicable to steady and time-varying sounds is presented in section III. 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.

2.1 Model for steady sounds

i) The specifications of the input spectral components *S(i)* are provided to the steady sounds model. ii) Following this, the spectral components *S(i)* undergo an outer and middle ear correction. This stage has an *O(N)* complexity, where N represents the number of frequency components. iii) The third stage computes an excitation pattern *E(k)* associated with the sound reaching the inner ear. Detectors are placed uniformly at 0.1 ERB (Equivalent Rectangular Bandwidth) units. ERB units are measured using an ERB number which represents the number of equivalent rectangular bandwidth auditory filters that can be fitted below any frequency. Auditory filter shapes are estimated for all detector location and frequency component combination which is an *O(ND)* complexity, where D is the number of detectors. iv) Following this, the excitation pattern *E(k)* at any detector k is calculated as the sum of the response from the different auditory filters [4] which change shape with center frequency and level [3]. This stage has an *O(ND)* complexity. v) The EP *E(k)* obtained is then transformed to a specific loudness pattern *SP(k)* according to the procedure described in [1]. Specific loudness represents the action of cochlea on the EP and represents loudness density pattern i.e. the loudness per ERB [5]. Therefore with *D* detectors this stage has an *O(ND)* 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 with an *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 and long-term loudness.

In [2], spectral analysis is performed with six parallel FFTs on hanning windowed segments of 2, 4, 8, 16, 32, 64 ms duration. Different frequency components are extracted from the appropriate segments to obtain appropriate time and frequency resolution. The spectrum obtained is updated at 1 ms intervals. The subsequent stages in the model are similar to the steady state model. Finally, the short-term loudness *SL* is obtained from the instantaneous loudness $\underline{L}$ which is 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. In this section, we describe the proposed low-complexity loudness estimation algorithm for steady and time-varying sounds.

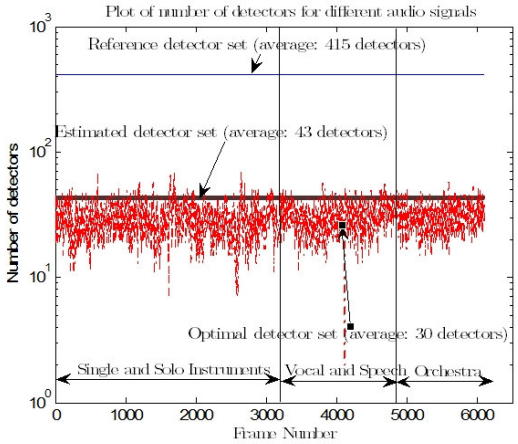

Fig. 2: Plot showing cardinality of optimal detector set L_o compared with reference set L_r and estimated set L_e .

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 $f_i's$, where $i \in \{1,2...N\}$ and detector locations $d_k's$ where $k \in \{1,2...D\}$ 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 $L_r = \{d_k; |d_k - d_{k-1}|=0.1, k=1,2,...D\}$ denote the reference set of detector locations expressed in ERB units, such that they are uniformly spaced at 0.1 ERB units. Let $L_o=\{d_k|\partial EP(k)/\partial k=0, k=1,2,...D\}$ denote the "optimal" set of 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_e, where L_e is an estimate of the set L_o because L_o is unavailable to us. The following two 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. 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).

$$I(m) = \sum_{i \in (m, m+1]} S(i) \qquad (1)$$

where $S(i)$ is the input spectral amplitude and $I(m)$ is the intensity pattern in the m^{th} ERB and i represents the set of components in the m^{th} ERB unit. In Fig. 3a, we show an example of a sample audio spectrum and intensity pattern $I(m)$ plotted on an ERB unit scale.

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 and detector locations

Here, we describe a procedure that estimates the positions of the approximated spectral components that best capture the structure 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 d_k is given by

$$EP(k) = \sum_i (1 + p_i.g_i) \cdot \exp(-p_i.g_i).S(i) \qquad (2)$$

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 $|g_i| \approx 0$, as $exp(-p_i.g_i) \approx 1$ in (2). As a result, we select the maximum $S(i)$ inside each ERB unit and place a detector close to it such that both $S(i)$ and $exp(-p_i.g_i)$ are maximized simultaneously in (2). 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.

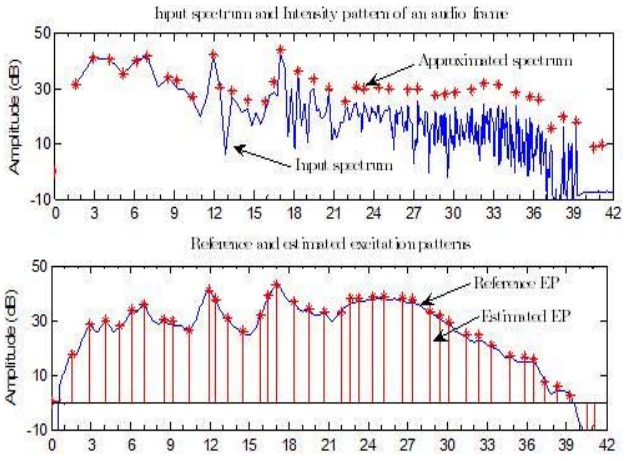

Fig. 3: a) Plot of input and approximated spectrum. b) Plot of reference and estimated EP.

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_r, the approximated components are placed at positions of maximal auditory filter response in each ERB. In Fig. 3b, we plot the reference excitation pattern and the estimated EP along with the positions of maximal auditory filter response.

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 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 (3) and (4) for the EP which is indicative of the relative error at each detector location d_k and average error across all detector locations.

$$\text{REE} = 20 \cdot \log_{10} \left\{ \sum_{k \in \{1,2..D\}} \left| \frac{\hat{E}(k) - E(k)}{E(k)} \right| \right\} \qquad (3)$$

$$\text{AEE} = 20 * \log_{10} \left\{ \frac{\sum_{k \in \{1,2,...D\}} \left| \hat{E}(k) - E(k) \right|}{\sum_{k \in \{1,2,...D\}} E(k)} \right\} \qquad (4)$$

Table 1: REE, AEE, ALE and MLE metric evaluation of the proposed algorithm for different audio material

Different types of audio	***AEE (dB)***	***REE (dB)***	***Loudness Error (sones)***	
			ALE	*MLE*
Single Instruments	-12.82	-14.84	0.72	3.2647
Speech	-12.80	-14.73	0.29	2.8186
Vocal	-12.03	-14.55	0.22	2.6029
Solo Instruments	-12.42	-14.60	0.44	2.2574
Vocal & Orchestra	-13.4	-18.57	0.95	3.2647
Orchestra	-11.52	-14.92	1.34	2.8186
Pop Music	-12.58	-14.90	0.27	2.6029
Average	**-12.5**	**-15**	**0.6**	**2.6**

The error in the estimated loudness is evaluated in terms of the Average Loudness Error (ALE) and the Mean Loudness Error (MLE) which are defined in (5) and (6) respectively.

$$\text{ALE} = \frac{1}{P}\sum_{j=1}^{P}\left|L'_j - L_j\right| \quad (5)$$

$$\text{MLE} = max(|\,L'_j - \,L_j\,|),\ j\epsilon\{1,2\ldots P\} \quad (6)$$

where $\hat{E}(k)$, $E(k)$ are the estimated and reference EP expressed in linear power units. 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 1 that the proposed algorithm performs consistently for different types of audio signals within a tolerable error.

Furthermore, we compare the computational complexity of the proposed algorithm with the standard approach followed in [1]-[4]. We also highlight the complexity of each stage separately due to the differing nature of operations in each stage. From Table 2, it can be seen that the proposed algorithm achieves a significant reduction in complexity close to 96% on average.

Table 2: Number of operations required in various stages of the model for the standard and proposed algorithm.

Stages	*Complexity comparison*		*Complexity Reduction (S-P)/S*
	D =415 N =512 standard(S)	*D = 43 N = 41 proposed(P)*	
Auditory filters: O(ND)	90692	1186	98%
EP: O(ND)	90692	1186	98%
Specific loudness-O(D)	415	43	89%
Total loudness: O(D)	415	43	89%
Overhead	5120	4460	12%
Total Complexity	187334	6650	96%

5. CONCLUSION

In this paper, we proposed a low-complexity loudness estimation algorithm applicable for steady and time-varying sounds based on the Moore et. al [1, 2] model. The proposed algorithm becomes more efficient by pruning the less significant frequency components. The algorithm also prunes the number of detectors by retaining only those detector locations that receive the highest response inside each ERB unit. The proposed algorithm is also seen to perform consistently for a wide variety of input audio material.

6. REFERENCES

[1] 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(4), pp. 224-240, April 1997.

[2] 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.

[3] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, p.103, 1990.

[4] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The JASA*, vol. 74, no. 3, pp. 750-753, 1983.

[5] E. Zwicker and H. Fastl, *Psychoacoustics: facts and models.* ,Berlin, Germany: Springer-Verlag, 1990.

[6] V. Berisha and A. Spanias, "Wideband Speech Recovery Using Psychoacoustic Criteria," *EURASIP Journal on Audio, Speech, and Music Processing*, Article ID 16816, 18 pages, 2007.

[7] T. Painter and A. Spanias, "Perceptual coding of Digital Audio," *Proceedings of the IEEE*, 88(4), pp. 451-513, April 2000.

[8] 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," *J.A.E.S*, vol. 53, No.10,Oct 2005.

[9] E. Skovenborg and S.H. Neilson, "Evaluation of Different Loudness Models with Music and Speech material," in *Proc. 117th conv. Aud. Eng. Soc.*, Oct.2004.

[10] T. Painter and A. Spanias, "Perceptual segmentation and component selection for sinusoidal representations of Audio," *IEEE Trans. on SAP*, vol. 13, no. 2, pp. 149-162, 2005.

[11] V. Atti and A. Spanias, "Rate Determination based on perceptual loudness," *IEEE ISCAS 2005*, pp. 848-851, May 2005.

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