LOW-COMPLEXITY SINUSOIDAL COMPONENT SELECTION USING LOUDNESS PATTERNS

Harish Krishnamoorthi, Visar Berisha, Andreas Spanias, Homin Kwon

Department of Electrical Engineering, SenSIP Center, Arizona State University, Tempe, AZ 85287-5706, USA
Email: [kharish, visar, spanias, homin]@asu.edu

ABSTRACT

Sinusoidal modeling of audio at low-bit rates involves selecting a limited number of parameters according to a quantitative or perceptual criterion. Most perceptual sinusoidal component selection strategies are computationally intensive and not suitable for real-time applications. In this paper, a computationally efficient sinusoidal selection algorithm based on a novel hybrid loudness estimation scheme is presented. The hybrid scheme first estimates efficiently the loudness of a multi-tone signal from the loudness patterns of its constituent sinusoidal components. Then it refines this estimate by performing a full evaluation of loudness but only in select critical bands. Experimental results show that the proposed technique maintains a low perceptual sinusoidal synthesis error at a much lower computational complexity.

Index Terms—sinusoidal synthesis, perceptual methods, loudness estimation, audio coding.

1. INTRODUCTION

Parametric speech and audio coding schemes [1, 7, 8, 11] have gained popularity owing to their ability to provide high quality speech and audio at lower bit-rates. For example, parametric techniques based on mixed basis representations [12] and on Sinusoids + Transients + Noise (STN) models [3, 4] have been successful in speech and audio synthesis. A number of important algorithms [2-5] have been proposed to estimate the amplitudes, frequencies and phases associated with the sinusoidal model. Some examples include peak picking in the short-time Fourier transform (STFT) domain [2], analysis-by-synthesis techniques [14], and matching pursuit decompositions [5].

In this paper, we focus on the problem of sinusoidal component selection based on perceptual criteria. In low bit-rate applications, only a limited number of sinusoidal parameters can be transmitted. In such situations, the goal is to select a subset of sinusoids deemed perceptually most relevant. In Fig. 1, the different stages involved in a typical perceptual sinusoidal component selection algorithm are shown. Due to the non-linear aspects of the perceptual model, an exhaustive search is often required to select the appropriate subset. As a result, the selection process can become computationally intensive. Several schemes have been proposed in the literature for selecting the perceptually salient sinusoids. For example, a strategy based on maximum signal to mask ratio (SMR) has been proposed for sinusoidal synthesis [1]. Additionally, excitation patterns [4] and loudness patterns [6] have also been employed for constrained sinusoidal representations. Peak-picking strategies based on maximum SMR or maximum SNR criteria focus on high-energy spectral regions, and therefore, can miss perceptually relevant sinusoids that are not located in these regions. Existing methods [4, 6, 16] focusing on excitation patterns and loudness are computationally expensive and hence inappropriate in delay-critical applications. Furthermore, their reliance on iterative, greedy component selection strategies can result in non-optimal sinusoidal subsets.

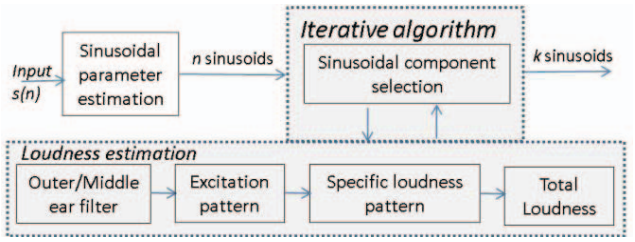

Figure 1: Block diagram of sinusoidal selection scheme.

In this paper, we propose a hybrid algorithm to estimate the loudness associated with a multi-tone signal. For tones separated by more than one critical band, the loudness of the multi-tone signal is obtained from the loudness patterns of individual sinusoidal components. This first step is done efficiently by taking the maxima among the specific loudness patterns; however, estimates need to be refined for multiple tones within a critical band. For this reason the loudness pattern is further evaluated precisely in select critical bands by evaluating the auditory filter shape. To show the validity of the model, we apply it in sinusoidal audio synthesis. The estimated loudness pattern is then used to select a constrained and perceptually optimal subset of sinusoidal components. This is done by using a greedy algorithm where each selected sinusoid is chosen such that it provides a maximal increment in loudness. The process of sinusoidal selection can be continued until a target bit rate is reached or target loudness in the synthesis signal is attained. The proposed method is computationally efficient and distinctly different than previously proposed sinusoidal synthesis approaches that employ the full Moore and Glasberg process [9] for sinusoidal loudness estimation. Instead the proposed method first approximates the loudness of distinct sinusoids and then refines these estimates by using more precise estimation only when multiple tones fall within a critical band. We show that CPU execution time is reduced by 90% thereby making loudness-based sinusoidal selection feasible.

The rest of the paper is organized as follows. In Section 2, an overview of the underlying loudness estimation algorithm is presented. Section 3 presents the proposed hybrid loudness estimation scheme for evaluating the loudness of multi-tone signals. In Section 4, the sinusoidal component selection based on the proposed hybrid loudness estimation scheme is described.

Section 5 provides simulation results and conclusions are presented in Section 6.

2. LOUDNESS ESTIMATION ALGORITHM

In this section, a brief description of the underlying loudness estimation algorithm employed is provided. A block diagram with the different stages of the algorithm is shown in Fig. 1.

2.1 Loudness estimation algorithm

The loudness patterns are computed according to the Moore & Glasberg loudness estimation procedure [9]. The steps involved are the following: i) First, the spectral components, $S(i)$, associated with an input audio segment, $s(n)$, undergo an outer and middle ear correction. Following this, the frequency scale is transformed into an auditory scale that is measured using an Equivalent Rectangular Bandwidth (ERB) number. The ERB number represents the number of equivalent rectangular bandwidth auditory filters that can be fitted below any frequency f (in Hz) and is calculated using

$$p = 21.4*\log(4.37f/1000+1) \text{ (in ERB units)} \quad (1)$$

where p represents the number of ERB units on the ERB scale. D detectors are placed uniformly at 0.1 ERB units; ii) Next, the excitation pattern (EP), $E(k)$, at any detector is calculated as the sum of the response from the different auditory filters. The auditory filters change shapes as a function of the center frequency and the total intensity within one ERB unit [13]; iii) The EP, $E(k)$, obtained is then transformed to a Specific Loudness Pattern (SLP), $SP(k)$, through the nonlinear transformation of the EP, i.e.,

$$SP(k) = c\cdot((E(k)+A(k))^{\alpha} - A(k)^{\alpha}) \quad (2)$$

where c = 0.047 and α=0.2 and $A(k)$ is a function of the peak excitation level at absolute threshold of hearing [9]. The specific loudness pattern represents the loudness density pattern, i.e., the loudness per ERB [9]; iv) The last stage involves calculating the area under the SLP to obtain the total instantaneous loudness (L).

3. PROPOSED HYBRID LOUDNESS ESTIMATION

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 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 in the following subsection.

3.1 Hybrid Loudness estimation

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. 2(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

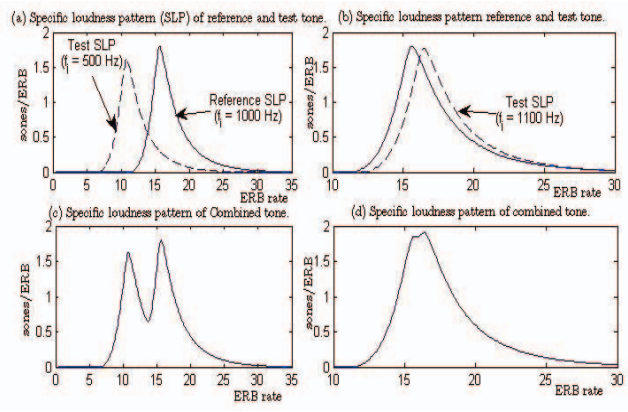

Figure 2: Plot of specific loudness patterns of reference, test and combined tones.

associated with the combined tone is plotted in Fig. 2(c) and (d). It can be observed that the envelope of the two specific loudness patterns in Fig. 2(a) and (b) closely resembles the exact specific loudness shown in Fig. 2(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 $L_T = \{d_k; |d_k - d_{k-1}|=0.1,\ k = 1,2,...D\}$ 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

$$\tilde{N}_{ij}(L_T) = \max(N_i(L_T), N_j(L_T)) \quad (3)$$

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

$$L_e \text{ (in sones)} = \int_0^{m \text{ ERB}} N_{ij}(z)dz - \int_0^{m \text{ ERB}} \tilde{N}_{ij}(z)dz \quad (4)$$

where N_{ij} represents the actual specific loudness pattern of the combined tone and m is the total number of ERB units. In Fig. 4, 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

$$d_{ij} \text{ (in ERB units)} = p_i - p_j \quad (5)$$

where p_i and p_j are computed using (1) and denote the ERB number associated with the reference and test tone respectively. It can be observed from Fig. 4 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 (3) when estimating the specific loudness pattern of the combined tone.

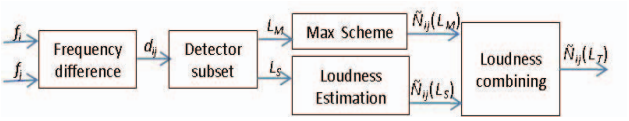

Figure 3: Block diagram of the proposed hybrid loudness estimation scheme.

To account for the change in filter shapes, we propose a novel approach that combines the "Max" scheme described in (3) 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 3. The steps are described below. First, the frequency separation (d_{ij}) between the test and reference tone is computed using (5). 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 (3).. Let p represent the ERB unit where the auditory filters change shapes. Let $L_S = \{d_k; |d_k - p| < m, k = 1,2...D\}$ 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 p^{th} ERB unit. For the subset L_S, all the steps associated with the loudness estimation procedure described in [9] 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 (3).

In Fig. 4, 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 [10] can be employed to further reduce the computational complexity.

4. SINUSOIDAL SELECTION BASED ON THE HYBRID LOUDNESS ESTIMATION SCHEME

In this section, the sinusoidal component selection algorithm based on the hybrid loudness estimation procedure is presented.

4.1 Sinusoidal component selection

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 [2] 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 j^{th} 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, $A = \{\varnothing\}$. 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

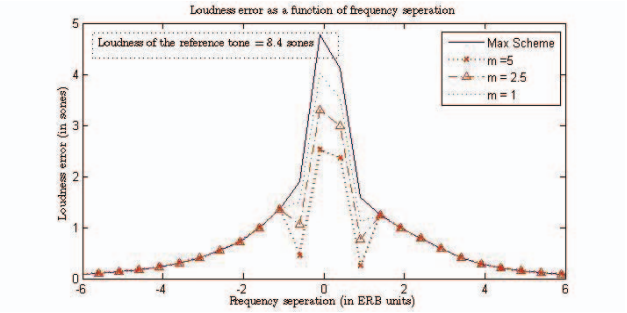

Figure 4: Plot of specific loudness patterns of reference, test and combined tones.

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 j^{th} iteration and the greedy nature of this algorithm requires that the loudness estimation algorithm be employed n-(j-1) times during the j^{th} iteration. Therefore, to select k sinusoids, the loudness estimation algorithm is executed n+ (n-1) + … + (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 2.1, 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 i^{th} sinusoid. Let N_i^{tr} represent the estimated specific loudness pattern during the i^{th} trial and N_j^S denote the estimated specific loudness pattern after j sinusoidal selections.

Algorithm: Computationally efficient sinusoidal selection.

$j = 1$; k = Total number of sinusoids to be selected.
m = number of ERB units.
while iteration $j<k$

- For $i = 1....|S|$
  - If f_i is within one ERB unit of any element of A
  - Define subset of detectors.
    - $L_S = \{d_k | |d_k - p_i| < m\}$
    - $L_M = L_T \setminus L_M$
  - Do selective evaluation of N_i^{tr} *at* L_S.
  - $N_i^{tr}(L_M) = max(N_i(L_M), N_{j-1}^S(L_M))$
  - L^i_{tr} (in sones) = Area under $N_i^{tr}(L_T)$
- End for

$q = argmax_i\, L^i_{tr}$
$f_S = f_a$; $N_i^S = N_a^{tr}$
$A = A \cup f_S$; $S = S \setminus f_S$;
end while

5. RESULTS

In this section we present simulation results. The performance of the algorithm was tested with different types of audio records

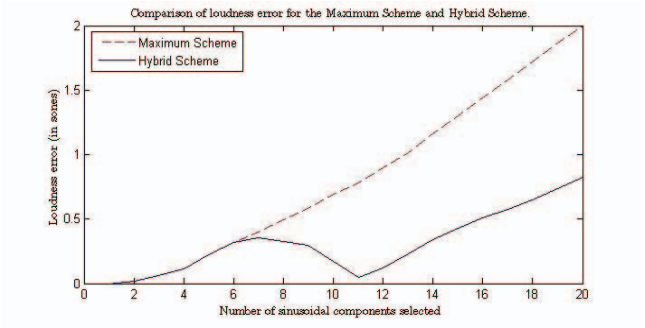

Figure 5: Loudness error estimates between the Maximum and Hybrid schemes.

obtained from the SQAM database [15]. 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.

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 I 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 I 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 II, 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. 5, 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. 5 that the hybrid scheme is associated with a lower average loudness error across all iterations.

Table I: 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 II: Sinusoidal component selection accuracy.

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. CONCLUSIONS

In this paper, we proposed a low complexity hybrid loudness estimation algorithm that estimates the loudness of a multi-tone signal from the specific loudness patterns of its constituent sinusoids. Our simulations show that the proposed low complexity algorithm in most cases selects a similar subset of sinusoids as the high complexity method that uses the full loudness estimation algorithm. We also show the low complexity algorithm is much more efficient than the full loudness estimation process and in fact in terms of CPU time the proposed algorithm requires 90% less time to execute.

7. REFERENCES

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

[2] 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, Aug 1986.

[3] 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.

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

[5] P. Vera-Candeas et. al, "A Sinusoidal Modeling Approach based on Perceptual Matching Pursuits for Parametric Audio Coding," *118th AES Convention,* May 2005.

[6] H. Purnhagen et.al., "Sinusoidal Coding Using Loudness Based Component Selection," *Proc. of IEEE ICASSP,* May 2002.

[7] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," *Proc. of IEEE,* vol. 88, no. 4, pp. 451-513, April 2000.

[8] A. Spanias, "Speech Coding: A tutorial review," *Proc. of IEEE,* vol. 82, no. 10, pp. 1441-1582, Oct 1994.

[9] B. C. J. Moore, B. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Aud. Eng. Soc.*, vol. 45, no. 4, pp. 224–240, Apr. 1997.

[10] H. Krishnamoorthi, V. Berisha and A. Spanias, "A Low-Complexity Loudness Estimation Algorithm," *Proc. of IEEE ICASSP 2008,* pp. 361-364, Las Vegas, April 2008.

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

[12] A. Spanias and P. Loizou, "A Mixed Fourier/Walsh Transform Scheme for Speech Coding at 4 KBPS," *Proc.IEE-PartI,* vol. 139, no. 5, pp. 473-481, Oct. 1992.

[13] B. C. J Moore, *An Introduction to the Psychology of Hearing,* Emerald Group Publishing, 2003.

[14] E. B. George and M. J. T. Smith, "Analysis-by-Synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Aud. Eng. Soc.*, pp. 497–516, Jun. 1992.

[15] "SQAM-sound quality assessment material: Recordings for subjective tests," EBU, Tech. Doc. 3253, 1988.

[16] 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.