SINUSOIDAL COMPONENT SELECTION BASED ON PARTIAL LOUDNESS CRITERIA

Harish Krishnamoorthi and Andreas Spanias

School of Electrical, Computer and Energy Engineering, SenSIP Center,
Arizona State University, Tempe, AZ, USA 85287-5706

ABSTRACT

Sinusoidal models are widely used in parametric speech and audio coding schemes. A common requirement in these applications is to select only a subset of components that provide the greatest perceptual benefit particularly at low bit-rates. Usually, perceptual sinusoidal component selection algorithms make use of greedy algorithms that are computationally expensive. In this paper, we present a new algorithm that selects sinusoidal components based on the partial loudness model proposed by Moore & Glasberg. We compare the performance of the proposed algorithm in terms of perceptual benefit and computational complexity to other existing sinusoidal selection algorithms.

Index Terms— loudness, sinusoidal models, parametric audio coding, audio coding, auditory patterns

1. INTRODUCTION

Parametric models are widely used in speech and audio coding algorithms owing to their ability to provide high quality audio at low bit rates compared to traditional transform-domain audio coders [1, 2, 3]. For example, the sinusoids + transients + noise (STN) model decomposes the signal into sinusoids, transients and noise components in order to obtain compact signal representations [4, 5, 6]. Similarly, the MPEG-4 audio standard consists of the HILN (Harmonics and Individual Lines plus Noise) audio coder and the HVXC (Harmonic Vector Excitation Coding) speech coder, both of which make use of parametric models and are widely used in internet streaming and broadcast applications [7]. However, in low bit-rate applications, only a limited number of parameters from individual parametric models can be encoded. Often times, it is desired that this limited set of sinusoidal parameters be selected such that the target bit-rate is scalable with perceptual quality, i.e., a gradual degradation in quality with decreasing bit-rates is desired. For example, the bandwidth extension algorithm proposed in [8] determined the importance of the higher sub-bands based on perceptual criteria.

In this paper, we focus on the sinusoidal parameter selection task where the objective is to select a limited number of perceptually salient sinusoids from a given set of candidate sinusoids. To this end, several perceptual techniques

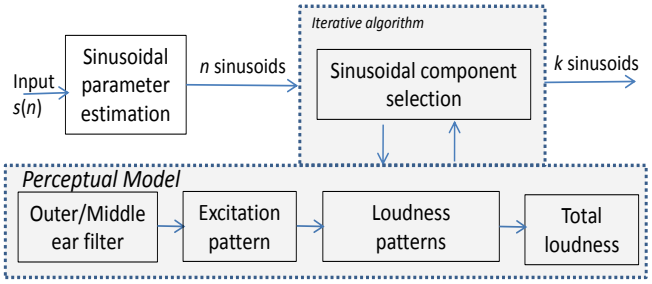

Fig. 1. General structure of a sinusoidal component selection task.

[6, 7, 9, 10] have been proposed in the literature. In [7], the signal-to-mask ratio (SMR) based criteria was suggested to select the components. Similarly, an *excitation pattern matching* algorithm [6] was proposed where the sinusoid whose excitation pattern resulted in the greatest matching (i.e., least error) to the reference signal pattern was selected. Later, a closely related technique where sinusoidal component selection was carried out based on *loudness pattern matching* was proposed [9].

Almost all of the above techniques employ iterative algorithms that are greedy in nature. That is, the perceptual model is employed repeatedly per candidate sinusoid and the sinusoid maximizing an underlying perceptual criteria is selected in that iteration. This process is repeated until the required number of sinusoids is selected. This process is computationally expensive and not adequate for practical applications. To this end, a few computationally efficient alternatives [10, 11, 12] were proposed. In [11], a pruning approach was described to evaluate the auditory model stages in a computationally efficient manner. In [10], a hybrid approach to loudness estimation for sinusoidal signals was proposed to speed up the sinusoidal component selection task. Specifically, they speed up the auditory model evaluations that are carried out repeatedly in every iteration.

In this paper, we propose a computationally efficient alternative that is based on the partial loudness model [13] proposed by Moore & Glasberg. The partial loudness model calculates the audibility (i.e., loudness) of a *signal of interest* in the presence of another *background signal*. At every

iteration, the proposed algorithm computes the partial loudness pattern associated with the set of candidate sinusoids (treated as the signal of interest) in presence of the sinusoidal components that are previously selected (treated as the background signal). The proposed algorithm then selects the next sinusoid based on the frequency region that shows the maximum partial loudness. While the other excitation/loudness pattern based approaches employ the auditory model repeatedly across all available candidate sinusoids before selecting a sinusoidal component, the proposed algorithm avoids this exhaustive search procedure. Therefore, the proposed technique is computationally efficient compared to the other component selection algorithms.

We compare the performance of the proposed algorithm to that of a loudness pattern matching algorithm [9] for component selection. Results indicate that the proposed algorithm selects $> 80\%$ of the same components as that selected by the loudness pattern matching algorithm. Furthermore, the proposed algorithm operates at 95% less time than the loudness pattern matching algorithm thereby achieving significant computational savings.

The paper is organized as follows. In section 2, the details of the perceptual model are provided. In section 3, the loudness pattern matching algorithm and the proposed algorithm are described. In section 4, the experimental setup and simulation results are presented. Finally, conclusions are given in section 5.

2. PARTIAL LOUDNESS MODEL

In this section, a brief overview of the steps associated with evaluating partial loudness patterns according to the Moore & Glasberg auditory model [13] is provided.

2.1. Auditory Model

Let $x(n)$ and $d(n)$ denote the signal of interest and the background signal ("noise") respectively. The combined signal is given by $y(n) = x(n) + d(n)$. The input signals are referenced to an assumed sound pressure level (SPL) of P dBSPL.

First, the input signals undergo an outer and middle ear correction so that the effective power spectrum reaching the inner ear is $P_x^c(\omega_j) = |M(\omega_j)|^2 P_x(\omega_j)$ and $P_d^c(\omega_j) = |M(\omega_j)|^2 P_d(\omega_j)$ where $|M(\omega_j)|$ denotes the frequency response of the outer/middle ear filter, $P_x(\omega_j)$ and $P_d(\omega_j)$ denote the power spectrum of $x(n)$and $d(n)$ respectively, $\omega_j = e^{\frac{i2\pi f_j}{f_s}}$ and f_s denotes the sampling frequency.

Let $\mathbf{A}$ denote a $D \times N$ matrix where each row of $\mathbf{A}$ represents an auditory filter's magnitudes at frequencies ω_k where $k \in \{1, ..., N\}$. Therefore, a set of D auditory filters are employed as indicated by the number of rows in $\mathbf{A}$. More details on computing the auditory filter magnitudes can be found in [14]. Also, the frequency scale is transformed into an auditory scale that is measured in equivalent rectangular bandwidth (ERB) units. The ERB scale is related to the frequency f according to the following relation:

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

A set of D detectors, $\{d_k\}_{k=1}^{D}$ are placed uniformly at 0.1 ERB units along the auditory scale (i.e., $|d_k - d_{k-1}| = 0.1$). Each detector d_k represents the centers of the D auditory filters employed in $\mathbf{A}$. Next, the excitation patterns E_{SIG} and E_{NOISE} associated with the input signals $x(n)$ and $d(n)$ are evaluated as the output of these D auditory filters to the effective spectrum reaching the inner ear and is given by:

$$\begin{aligned} E_{SIG} &= \mathbf{A}\mathbf{P}_{\mathbf{x}}^{\mathbf{c}} \\ E_{NOISE} &= \mathbf{A}\mathbf{P}_{\mathbf{d}}^{\mathbf{c}} \end{aligned} \quad (2)$$

where the vectors $\mathbf{P}_{\mathbf{x}}^{\mathbf{c}}$, $\mathbf{P}_{\mathbf{d}}^{\mathbf{c}}$ represents the effective power spectrum of the signal of interest and background signal respectively after outer/middle ear correction.

Finally, the partial loudness pattern of the signal $x(n)$ in the presence of the background signal $d(n)$ is evaluated. Let E_{THRN} denote the peak excitation of a sinusoidal signal when it is at its masked threshold (in the presence of the background signal) and E_{THRQ} denote the peak excitation when the sinusoid is at its absolute threshold. Let $D_1 = \{k|E_{SIG}(k) > E_{THRN}(k)\}$ and $D_2 = \{k|E_{SIG}(k) \leq E_{THRN}(k)\}$ denote the two sets of locations along the auditory scale where $E_{SIG} > E_{THRN}$ and $E_{SIG} \leq E_{THRN}$ respectively. Then, the partial loudness pattern at locations $D1$ is calculated according to [13]:

$$\begin{aligned} N'_{SIG} = C\{&[(E_{SIG} + E_{NOISE})G + A]^{\alpha} - A^{\alpha}\} \\ &- C\{[(E_{NOISE}(1+K) + E_{THRQ})G + A]^{\alpha} \\ &- (E_{THRQ}G + A)^{\alpha}\} \left(\frac{E_{THRN}}{E_{SIG}}\right)^{0.3} \end{aligned} \quad (3)$$

and the partial loudness pattern at locations $D2$ (i.e., when $E_{SIG} < E_{THRN}$) is calculated according to:

$$\begin{aligned} N'_{SIG} = C &\left(\frac{2E_{SIG}}{E_{SIG} + E_{THRN}}\right)^{1.5} ((E_{THRQ}G + A)^{\alpha} - A^{\alpha}) \\ &\left\{\frac{[(E_{SIG} + E_{NOISE})G + A]^{\alpha} - (E_{NOISE}G + A)^{\alpha}}{[(E_{NOISE}(1+K) + E_{THRQ})G + A]^{\alpha} - (E_{NOISE}G + A)^{\alpha}}\right\}. \end{aligned} \quad (4)$$

The indexing by D_1 and D_2 has been omitted in (3) and (4) for readability. The values of the constants in (3), (4) are $C = 0.047$, $\alpha = 0.2$, $G = 1$ and $A = 2E_{THRQ}$ and K is a frequency dependent constant. More details regarding the constants can be found in [13].

3. PROPOSED ALGORITHM

3.1. Auditory Pattern Matching

The auditory pattern matching algorithm [6, 9] makes use of excitation patterns or loudness patterns in order to select L

perceptually salient sinusoids out of N (where $L < N$) candidate sinusoids. The L sinusoids are selected such that the error between the auditory pattern associated with the modeled signal (containing any set of L sinusoids) and that of the reference signal (consisting of all N sinusoids) is minimized.

Due to nonlinear operations involved in evaluating the auditory patterns, the optimal solution is usually found through an exhaustive search procedure. This process is combinatorial in nature with $O\binom{N}{L}$ combinations and is therefore computationally intensive. Therefore, several iterative algorithms [6, 9, 10] have been proposed to carry out the sinusoidal component selection task in a computationally efficient manner.

The general idea behind these algorithms is as follows: In the first iteration, the auditory patterns (either excitation/loudness patterns) associated with each of the N sinusoids are individually evaluated and the one that results in the least error with respect to the reference signal's auditory pattern is selected. In the next iteration, each of the remaining $N-1$ sinusoids are individually combined with the selected sinusoid and the sinusoid that produced the least auditory pattern error is selected. More generally, in the i^{th} iteration, each of the remaining $N-(i-1)$ sinusoids are individually combined with the $i-1$ already selected sinusoids and the sinusoid that results in the least auditory pattern error is selected. This process is repeated until all L sinusoids have been selected. Therefore, the iterative schemes are therefore associated with $O(N+(N-1)+(N-2)+...+N-(L-1)) = O(NL+L(L-1)/2)$ computational complexity. In the next subsection, we describe the proposed algorithm for perceptual sinusoidal component selection task.

3.2. Partial loudness based sinusoidal selection (PLSS)

The proposed algorithm makes use of the partial loudness model described by Moore & Glasberg [13]. In particular, the proposed algorithm computes a partial loudness pattern that represents a frequency dependant audibility of a signal of interest in presence of another background signal. To that end, the proposed algorithm chooses the set of candidate sinusoids as the *signal of interest* and the set of selected sinusoids as the *background signal*. This ensures that we measure the loudness contributions from the remaining candidate sinusoids only. Next, we describe the steps involved in the proposed sinusoidal selection algorithm.

Let S^i, C^i denote the set containing the selected sinusoids and the candidate sinusoids respectively in the i^{th} iteration where $i \in \{1, ..., L\}$. Let $\mathbf{P_x}^i, \mathbf{P_d}^i$ represent the vector of power spectral components associated with the background signal, $d_i(n)$ and the signal of interest, $x_i(n)$, respectively. Also, let $\mathbf{P_d}^i = g(S^i)$ and $\mathbf{P_x}^i = g(C^i)$ denotes the fact that $\mathbf{P_d}^i$, $\mathbf{P_x}^i$ are frequency domain representation of a signal containing the frequencies in the sets S^i, C^i respectively. A pseudo-code describing the PLSS algorithm is listed in Algorithm 1.

Input: $C^1 = \{f_1, f_2, ..., f_N\}$; % signal of interest;
$S^1 = \emptyset$; %background signal;
$i = 1$ %iteration index;
while $i \leq L$ **do**
1. Compute $PL_i(k)$ = PartialLoudness($\mathbf{P_x}^i$,$\mathbf{P_d}^i$)
2. d_m =find($PL_i = \max(PL_i)$);
3. Define window $W_m = [d_m - 0.5, d_m + 0.5]$
4. $f_p = max(\mathbf{P_x}^i(W_m)$
5. $S^{i+1} = S^i \cup f_p$ and $C^{i+1} = C^i \setminus \{f_p\}$.
6. $\mathbf{P_x}^{i+1} = g(C^{i+1})$ and $\mathbf{P_d}^{i+1} = g(S^{i+1})$

end

Algorithm 1: Partial loudness based sinusoidal selection algorithm (PLSS).

Initialy, the set of sinusoids for the background signal is empty, i.e., $S^1 = \emptyset$. The set of candidate sinusoids for the signal of interest contains all the N candidate sinusoids, i.e., $C^1 = \{f_1, f_2, ..., f_N\}$. The partial loudness pattern, $PL_i(k)$, of $\mathbf{P_x}^i$ in the presence of $\mathbf{P_d}^i$ is evaluated according to the steps described in section 2. The detector location, d_m, at which $PL_i(k)$ attains a maximum value is evaluated. A narrow frequency region, W_m, in the vicinity of the detector location d_m that spans one ERB unit is defined. That is, $W_m = [d_m - 0.5, d_m + 0.5]$. The sinusoidal component with the maximum amplitude that falls within the frequency region defined by W_m is selected for the i^{th} iteration. Let f_p denote this selected sinusoid. The set of frequencies for the signal of interest and background signal are updated to $S^{i+1} = S^i \cup f_p$ and $C^{i+1} = C^i \setminus \{f_p\}$ respectively. The corresponding signals in the frequency domain therefore correspond to $\mathbf{P_x}^{i+1} = g(C^{i+1})$ and $\mathbf{P_d}^{i+1} = g(S^{i+1})$.

3.3. PLSS vs. Existing techniques

The PLSS algorithm is different from the existing auditory pattern matching algorithms in the following aspects:

- Firstly, the PLSS algorithm makes use of the *partial loudness patterns* instead of the *excitation/loudness* patterns used in the existing techniques [6, 9, 10].
- Secondly, the PLSS technique eliminates the repeated application of auditory model per candidate sinusoid in any given iteration. Instead, the partial loudness model is evaluated only once irrespective of the number of candidate solutions available. In particlar, the PLSS operates at a computational complexity of $O(L)$ instead of $O(NL+L(L-1)/2)$ followed by existing auditory pattern matching techniques [6, 9, 10].
- Most importantly, the frequency dependant partial loudness pattern computed at every iteration approximates the loudness error metric that is measured by existing auditory pattern matching algorithms [9, 10] to select sinusoids.

Table 1. Component Similarity Measure (CSM).

Audio Material	Number of selected components			
	L=5	L=10	L=15	L=20
Pop	83.4%	85.8%	88.8%	90%
Orchestra	83.8%	87.5%	91.6%	93%
Vocal+Orchestra	88.6%	90.5%	92.3%	92.7%
Solo Instruments	88.4%	92.2%	93.4%	92%

Table 2. CPU execution times for sinusoidal selection.

	CPU time (in seconds)		Percentage Reduction
	Reference	PLSS	
L=5	16.11	0.79	95.1%
L=10	37.13	1.58	95.74%
L=15	61.09	2.38	96.1%
L=20	87.59	3.24	96.3%

4. RESULTS AND DISCUSSION

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

4.1. Experimental Setup

The audio records from the SQAM database [15] were used to evaluate the performance of the proposed PLSS algorithm. The audio signals are sampled at 44100 Hz and split into small segments of size $N_f = 1024$ samples. Spectral analysis is carried out using an N_f-point fast Fourier transform (FFT). Furthermore, the spectral components are referenced to an assumed sound pressure level (SPL) of 90 dB SPL.

For every segment of audio, a set of sinusoids are extracted by following the simple peak-picking procedure. The peak picking procedure selects those components that exhibit a local maxima in the FFT spectrum, i.e., $(P_x(\omega_{k-1}) < P_x(\omega_k) \leq P_x(\omega_{k+1}))$. The local maxima's are picked so as to ensure that sinusoids are picked across the spectrum, not just based on their signal power. 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 loudness pattern based sinusoidal selection technique [9] is used as the reference algorithm to compare the performance of the proposed PLSS techniques. In particular, we evaluate the PLSS algorithm in terms of the component similarity measure (CSM), computational complexity and the residual loudness error (RLE).

The CSM metric measures how many sinusoidal components are in common between the two techniques being compared, i.e., it evaluates whether the PLSS algorithm selects the same sinusoids as the reference algorithm. In Table 1, we list the CSM for different types of audio material for $L = 5, 10, 15, 20$ components. It can be observed that the PLSS technique selects at least $> 80\%$ of the same components as that selected by the reference algorithm.

Next, we evaluate the computational complexity in terms of the CPU time associated with the PLSS algorithm and the reference algorithm. All simulations were performed using MATLAB on an Intel 2.4 GHz i5-core processor with 4 GB RAM. Table 2 lists the CPU execution times for the PLSS and the reference algorithm. Results indicate an average reduction of 95% in computational time between the PLSS and the reference algorithm. This is due to the fact that the PLSS algorithm operates at an $O(L)$ computational complexity whereas the reference algorithm operates at an $O(NL + L(L-1)/2)$ computational complexity.

Finally, the residual loudness error is measured between the modeled signal (with L sinusoids) and original signal (with N sinusoids) for both PLSS and reference algorithms. Figure 2 plots the residual loudness error in dB units. The PLSS algorithm comes very close to the reference algorithm in terms of the loudness of the synthesized signal.

5. CONCLUSION

In this paper, we proposed a sinusoidal component selection algorithm based on the partial loudness model developed by Moore & Glasberg. The proposed PLSS algorithm computes the partial loudness of candidate sinusoids in presence of previously selected sinusoids. We show that selecting sinusoids from frequency regions that exhibits maximum partial loudness in every iteration results in $> 80\%$ of similar sinusoidal selections as that obtained from a loudness pattern matching algorithm. Furthermore, we show that the PLSS algorithm operates at $0(L)$ computational complexity and requires 95% less time on average compared to the loudness pattern matching algorithm.

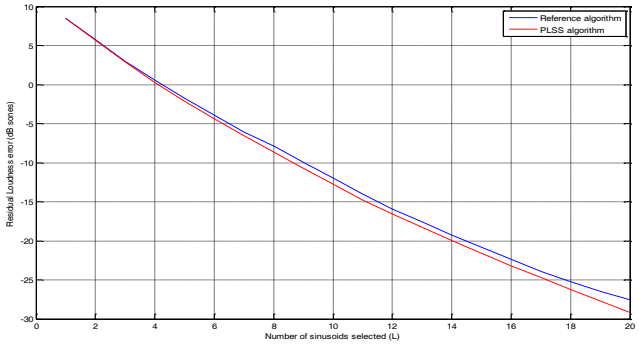

Fig. 2. Residual loudness error for reference and PLSS algorithms.

6. REFERENCES

[1] B. Edler and H. Purnhagen, "Parametric audio coding," in *Proc. International Conference on Signal Processing (ICSP2000)*, Aug 2000, vol. 1, pp. 21 –24.

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

[3] Jin Li and J.D. Johnston, "Perceptually layered scalable codec," in *Signals, Systems and Computers, Fortieth Asilomar Conference on*, 2006, pp. 2125–2129.

[4] 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.

[5] 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.

[6] 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, May 2005.

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

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

[9] H.Purnhagen, N. Meine, and B.Edler, "Sinusoidal coding using loudness based component selection," in *Proc. IEEE ICASSP*, May 2002.

[10] H.Krishnamoorthi, V.Berisha, A.Spanias, and H.Kwon, "Low-complexity sinusoidal component selection using loudness patterns," in *Proc. ICASSP*, April 2009.

[11] 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.

[12] R.J. Cassidy and J.O. 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.

[13] B. C. J. Moore, B. R. 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, April 1997.

[14] 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.

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