Perceptual Matching Pursuit for Audio Coding

Hossein Najaf-Zadeh, Ramin Pichevar, Hassan Lahdili, and Louis Thibault

Advanced Audio Systems, Communications Research Centre Canada

3701 Carling Ave., Ottawa, Ontario, Canada K2H 8S2

Correspondence should be addressed to Hossein Najaf-Zadeh (hossein.najafzadeh@crc.ca)

ABSTRACT

This paper introduces a Perceptual Matching Pursuit (PMP) algorithm for audio coding. A masking model has been developed and integrated into the matching pursuit algorithm to account for the characteristics of the hearing system. By doing so, only an audible kernel is extracted at each iteration. Moreover, contrary to the matching pursuit algorithm, PMP will stop decomposing an audio signal once there is no audible part left in the residual. We have used ITU-R PEAQ to compare audio materials decomposed by PMP and by matching pursuit. Objective scores for PMP increase by up to 1 unit. A semi-formal listening test has verified the objective scores and shown the perceptual superiority of PMP over the matching pursuit algorithm.

1. INTRODUCTION

Matching pursuit is an iterative technique in which a signal is map onto a set of kernels [1]. It has been used in audio source coding for years [2]. However, there have been some problems that make this approach less practical. Heavy computation is the main obstacle, which has been largely overcome by the huge advancement in processors. Another issue with matching pursuit algorithm is no stop criterion has been reported to limit the number of extracted kernels. This problem results in selecting a large number of kernels to assure high audio quality. The parameters associated with each extracted kernel need to be coded and transmitted (or stored). A number of the extracted kernels (depending on the total number of the extracted kernels) are not perceptually relevant and can be ignored without any loss in audio quality. This report introduces a Perceptual Matching Pursuit (PMP) algorithm, in which only perceptually-relevant kernels are extracted and moreover signal decomposition stops when there is no audible part left in the audio signal. PMP performs better than the matching pursuit algorithm as it reduces the computational load by limiting the number of the extracted kernels. Moreover, PMP delivers better audio quality (for a given number
of the extracted kernels) by extracting only audible kernels at each iteration.

2. PERCEPTUAL MATCHING PURSUIT

One approach to incorporate auditory masking into a matching-pursuit-based audio encoder is to extract many kernels followed by passing the extracted kernels through a masking module to eliminate the inaudible kernels [3]. In that approach lots of kernels (including inaudible ones) are extracted that cause unnecessary computational load. Another approach is to give perception-based weights to different kernels based on their perceptual importance [4][5]. This method will result in better audio quality. However, it does not suggest any stop criterion once there is no audible part left in the audio signal. Our approach is to progressively create a Time-Frequency (TF) masking pattern to determine a masking threshold at all time indexes and frequencies. Once no kernel with its magnitude above the masking threshold can be extracted, the decomposition stops and the audio signal can be reconstructed from the extracted kernels without loss in quality. The accuracy of the masking model plays a key role in the quality of the reconstructed audio signal.

2.1. TIME-FREQUENCY MASKING MODEL

A masking model has been developed for the matching pursuit-based audio encoder reported in [3]. However, it can be used in any matching-pursuit-based audio encoder with minor modifications. In the encoder reported in [3], gammachirp kernels (i.e., gammatone functions with an adaptive chirp factor) are used in the matching pursuit algorithm. Since there is not much difference in the spectra of gammatone and gammachirp functions, gammatone functions are employed to develop the masking model. In each critical band, the temporal forward and backward masking threshold produced by a gammatone kernel is calculated. For forward and backward masking, we assume a linear relation between the masking threshold (in dB) and the logarithm of the time delay between the masker and the maskee in msec [6]. Since the effective duration of forward masking depends on the masker duration [7], we define an effective duration for forward masking in critical band \( k \) as follows

\[
F_d_k = 100 \arctan(d_k),
\]

\( d_k \) denotes the effective duration of the \( k \)th kernel, defined as the time interval between the points on the temporal envelope of the \( k \)th kernel where the amplitude drops by 90%.

The forward masking threshold is given by

\[
FM_i(n) = (SL(i) - c_{kn}) \frac{\log_{10}\left(\frac{n}{n_i + L_k + FL_k}\right)}{\log_{10}\left(\frac{n_i + L_k + 1}{n_i + L_k + FL_k}\right)}
\]

where \( SL_i(i) \) (in dB) is the sensation level of the \( i \)th kernel in critical band \( k \), \( n_i \) is the start time index of the \( i \)th kernel, \( L_k = d_k f_s \) is the effective length of the gammatone function in critical band \( k \), and

\[
FL_k = \text{round}(F_d_k f_s),
\]

\( f_s \) denotes the sampling frequency,

\[
n_i + L_k + 1 \leq n \leq n_i + L_k + FL_k,
\]

and \( c_{kn} \) (in dB) is an offset value in critical band \( k \) and time index \( n \), subtracted from the sensation level to determine the masking threshold. Our experiments have shown that for a strongly tonal part of the spectrum, this offset is around 20 dB. However, for noise-like parts of the spectrum, this offset can be reduced to elevate the masking threshold. We have empirically defined the following offset value as a function of the tonality level in each critical band for the frames of 1024 audio sample

\[
c_{kn} = 4 \tau_{kn} + 16,
\]

where \( \tau_{kn} \) is the tonality index for critical band \( k \) at time index \( n \). The tonality index is between zero (for noise-type signal) and 1 (for a pure sinusoid). Appendix A describes a method to determine the tonality index.

For the backward masking, we assume 3 msec as the effective duration of masking for all critical bands regardless of the effective duration of gammatone functions. Hence, the backward masking threshold is given by
The sensation level is given by
\[
SL_k(i) = 10 \log_{10} \left( \frac{A^2_k(i)G_k^2}{QT_k} \right)
\]
where \( A_k(i) \) is the magnitude of the \( i \)th kernel extracted in critical band \( k \), \( G_k \) is the peak value of the Fourier transform of the normalized kernel in critical band \( k \), and \( QT_k \) is the elevated threshold in quiet for the same critical band.

Since the effective duration of gammatone kernels are less than 200 msec, the absolute threshold of hearing is elevated by 10 dB/decade [11]. The elevated threshold in quiet in critical band \( k \) is given by
\[
QT_k = AT_k + 10(\log_{10}(200) - \log_{10}(d_k))
\]
where \( AT_k \) is the absolute threshold of hearing in critical band \( k \).

The masking threshold in a critical band at any time instance is determined by taking the maximum of the masking threshold caused by the extracted kernels in the same critical band and two adjacent critical bands. We set the initial level for the masking pattern in critical band \( k \) to \( QT_k \) and consider three situations for the masking pattern caused by a kernel. When a maskee starts within the effective duration of the masker, the masking threshold is given by
\[
M_k(n_i : n_i + L_k) = \max(M(n_i : n_i + L_k), SL_k(i) - c_{kn}),
\]
where \( M_k(n) \) is the masking pattern (in dB) in critical band \( k \).

Other situations are when a maskee starts after the effective duration of the masker (i.e., forward masking), and when a maskee starts before a masker (i.e., backward masking).

This forward masking contributes to the global masking pattern in critical band \( k \) as follows
\[
M_k(n_i + L_k + 1 : n_i + L_k + FL_k) = \max(M_k(n_i + L_k + 1 : n_i + L_k + FL_k),)
\]
Similarly, the backward masking affects the global masking pattern in critical band \( k \) as follows
\[
M_k(n_i - 0.005f_k : n_i - 1) = \max(M_k(n_i - 0.005f_k : n_i - 1),)
\]
We have considered the masking effects caused by any extracted kernel in two adjacent critical bands. According to [8] a single masker produces an asymmetric linear masking pattern in the Bark domain, with a slope of -27 dB/Bark for the lower frequency side and a level-dependent slope for the upper frequency side. The slope for the upper frequency side is given by
\[
s_u = -24 - \frac{230}{f} + 0.2L_m
\]
where \( f \) is the masker frequency and \( L_m \) is the masker level in dB. We have used this approach to calculate the masking effects caused by each selected kernel in the two immediate neighboring critical bands.

2.2. PERCEPTION-BASED SIGNAL DECOMPOSITION

In matching pursuit, at each iteration the value and position of the maximum of the cross correlation of the residual signal and each kernel is found. The kernel which has the highest correlation with the residual signal is identified. The maximum value of the cross correlation and its position are determined. Prior to finding the maximum value for each correlation function, the values below the masking threshold are set to zero. In other words, the correlation at any time index
is taken into consideration if the sensation level is above the associated masking threshold at that time index,

\[
\frac{A^2(n)G_k^2}{QT_k} > 10^{(M(n)/10)},
\]

\[
|A(n)| > \sqrt{QT_k} \frac{10^{(M(n)/10)}}{G_k}.
\]

As such, only audible kernels are extracted and the masked values in the correlation sequences will be discarded. By doing so the noise spectrum (i.e., residual spectrum) is shaped and a higher noise level is allowed as long as it is inaudible. Figure 1 shows the power spectrum of a frame of an audio signal and also the spectra for the residual for the matching pursuit and PMP algorithms. As is seen, the PMP algorithm shapes the noise spectrum and therefore produces higher quality audio for the same number of extracted kernels.

3. COMPARISON OF PMP WITH MP

We have used ITR-R PEAQ for an objective comparison of MP with PMP. The test materials were monaural audio sequences, sampled at 44.1 kHz, with 16-bit resolution. In order to assure transparent reconstruction of audio sequences, the number of iterations for each file was set such that either MP or/and PMP achieve a PEAQ score above -1.

Table 1. Mean subjective and objective scores for a few audio files processed with MP and PMP.

<table>
<thead>
<tr>
<th>Audio Material</th>
<th>Mean Subj. Score MP</th>
<th>Mean Subj. Score PMP</th>
<th>PEAQ Obj. Score MP</th>
<th>PEAQ Obj. Score PMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Susan Vega</td>
<td>-0.7500</td>
<td>-0.9667</td>
<td>-0.593</td>
<td>-0.330</td>
</tr>
<tr>
<td>Trumpet</td>
<td>-0.9667</td>
<td>-0.4333</td>
<td>-1.809</td>
<td>-0.791</td>
</tr>
<tr>
<td>Orchestra</td>
<td>-0.7667</td>
<td>-0.4000</td>
<td>-1.239</td>
<td>-0.915</td>
</tr>
<tr>
<td>Harpsichord</td>
<td>-0.5000</td>
<td>-0.3667</td>
<td>-1.699</td>
<td>-0.502</td>
</tr>
<tr>
<td>Bagpipe</td>
<td>-0.4000</td>
<td>-0.2000</td>
<td>-0.765</td>
<td>-0.867</td>
</tr>
<tr>
<td>Glockenspiel</td>
<td>-0.2000</td>
<td>-0.333</td>
<td>-0.925</td>
<td>-1.266</td>
</tr>
<tr>
<td>Plucked Strings</td>
<td>-0.4667</td>
<td>-0.2667</td>
<td>-1.050</td>
<td>-0.299</td>
</tr>
</tbody>
</table>

As is seen in Table 1, PMP achieves higher scores for the majority of test materials.

3. COMPARISON OF PMP WITH MP

In order to verify the objective scores, we conducted a semi-formal listening test to evaluate the quality of the test signals. Six subjects took part in a “triple stimulus hidden reference” test and listened to the audio materials (presented in Table 1) over the headphone in a quiet room. The CRC SEAQ software was used in the test which allowed the listener to seamlessly switch among the three stimuli. In each trial, the stimuli “A” was always the reference stimulus known by the subject. Two other stimuli, “B” or “C”, were either a hidden reference, identical to “A”, or a synthesized version of the same audio material. None of “B” or “C” was known to the subject. The listener had to identify the synthesized version (either “B” or “C”) and to grade its quality relative to that of the reference on “A”. The grading scale was continuous from 1 (very annoying) to 5 (no difference between the reference and the synthesized file). During the listening test, each subject was free to take as much time as required on any trial, switching freely among the three stimuli as often as desired.

The average subjective scores for MP and PMP were 4.4214 and 4.5762, and the standard deviations of the scores were 0.2522 and 0.2612 respectively. Although the confidence intervals for the subjective scores are overlapping, the majority of the test materials received higher subjective scores for PMP, which is consistent with the objective evaluation. For two test materials, i.e. Susan Vega and Glockenspiel, PMP received slightly lower scores that will be investigated to enhance the algorithm.

The audio materials in the listening test were high quality, i.e. transparent or near transparent replica of the
reference files. However, a listening test at lower bitrates (below or near transparent audio quality) would make a better comparison of the two different algorithms.

4. CONCLUDING REMARKS

We have introduced a Perceptual Matching Pursuit (PMP) algorithm for audio coding. A masking model has been developed and integrated into the matching pursuit algorithm to account for the characteristics of the hearing system. PMP selects an audible kernel at each iteration and stops decomposing the audio signal once there is no audible part left in the residual. We have used ITU-R PEAQ for objective evaluation of audio materials decomposed by PMP and by matching pursuit. Objective scores for PMP increase by up to 1 unit. A semi-formal listening test has verified the objective scores and shown the perceptual superiority of PMP over the matching pursuit algorithm at a fixed number of iterations.

Since PMP selects only audible kernels, it would better reproduce low energy parts of audio signals, which might not be well-presented in a matching pursuit-based system. However, the accuracy of the perceptual matching pursuit is highly affected by the auditory masking model used in the algorithm. Any inaccuracy of the masking model would cause overmasking of audible parts of audio signals.

This work shows that the integration of auditory masking into the matching pursuit algorithm enhances the quality of the reconstructed signal at a fixed bitrate, or equivalently reduces the number of selected kernels for the same audio quality. However, more refinement of the auditory masking model is required to increase the coding efficiency while maintaining high quality reconstruction of audio signals.

5. REFERENCE


APPENDIX A: DETERMINATION OF THE TONALITY INDEX

That is known from psychoacoustic experiments that noise-like and tonal maskers (with the same power) produce different masking thresholds. The effectiveness of a noise masker exceeds that of a tonal masker by up to 20 dB. This different behavior has been exploited in simultaneous masking models [9][10]. With the same
reasoning, we have adapted the masking offset value to the characteristic of audio signals in different critical bands. In fact for many sounds such as speech a strong tonal structure can be found in the low frequency part of the spectrum, while no tonal behavior is observed at high frequencies. This leads us to make the masking pattern adaptive to the local behavior of the spectrum in each critical band.

We have investigated different methods to identify tonal structure in audio spectrum. In steady-state parts of an audio signal, identification of tonal tracks (through inter-frame sinusoidal track continuity) is the most accurate way to confirm a tonal structure. However this method fails to identify short tones (10-20 msec). Hence, in order to avoid missing tonal behavior, we just apply a peak-picking method on a spectrum representing 1024 audio samples (23.2 msec at 44100 Hz sampling rate).

We determine the tonality level in each critical band in a frame of 1024 samples. A frame of the audio signal is multiplied by a Hanning window, followed by taking the DFT of 1024 points. The first 512 spectral components are grouped into 25 critical bands. The peaks in the spectrum are found and associated with the corresponding critical band. If there is no spectral peak found in a critical band, its tonality index is set to zero.

For a peak in a critical band, we take the peak value and the higher magnitude from the two adjacent frequency bins. The peak value and the magnitude in the adjacent frequency bin are assumed to be produced by a sinusoid. In order to verify this hypothesis, these values are compared with the normalized spectrum of a pure sinusoid windowed by a Hanning window. The reason we use a 3rd-order polynomial for this prediction is that the adjacent bin with a smaller magnitude is more than one frequency bin away from the position of the maximum magnitude in the audio spectrum. Therefore, we have fit one side of the prototype spectral magnitude (up to 1.5 bins away from the position of the peak to a 3rd-order polynomial in the log domain. The 3rd-order fit is very accurate and represented by the following coefficients

$$C_{p3} = [-2.2088 \ -2.8538 \ -0.9984 \ 0.0606].$$

In using these coefficients, $C_{p3}(4)$ is shifted by $A_{\text{max}}$, and the position of the maximum magnitude is considered as the origin.

We also predict the phase at the three frequency bins (bin with the peak magnitude and two adjacent bins). It can be shown that spectral phase of the prototype spectrum varies linearly around the location of the maximum magnitude with a slope of $\pi$ per bin (see Appendix….). Using this fact, we can predict the spectral phase at the three frequency bins. Prior to phase prediction at those bins, we need an estimate of the phase at the location of the maximum magnitude. This phase can be estimated from the spectral phase at the two neighboring frequency bins as follows

$$P_{\text{max}} = P_p + (P_{\text{adj}} - P_p)|\Delta_k|,$$

The spectral phase at the frequency bin with the peak magnitude and the two adjacent bins are predicted as follows

$$\Delta_k = \frac{A_p - A_{\text{adj}} - C_{p2}(1)}{2C_{p2}(1)},$$

where $A_p$ and $A_{\text{adj}}$ are the peak and the magnitude at the adjacent frequency bin, $A_{\text{max}}$ is the maximum magnitude and $k_{\text{max}}$ is the index to the position of the maximum magnitude in the frequency domain (around the selected spectral peak). Once the maximum peak and its location are found, we predict the spectral magnitude in the frequency bin with the peak magnitude and the two adjacent bins. In doing so, we predict those magnitudes from a 3rd-order polynomial that has been fit to one side of the prototype magnitude spectrum (i.e. spectrum of a pure sinusoid windowed by a Hanning window).
\[ \tilde{P}_p = P_{\text{max}} - \pi(k_p - k_{\text{max}}), \]
\[ \tilde{P}_1 = P_{\text{max}} - \pi(k_p - 1 - k_{\text{max}}), \]
\[ \tilde{P}_2 = P_{\text{max}} - \pi(k_p + 1 - k_{\text{max}}). \]

We calculate the prediction relative error by comparing the predicted values with the spectral values at the three bins

\[ \xi = \sum_{m=1}^{3} \left| \frac{X(k_m) - \tilde{X}(k_m)}{X(k_m)} \right|. \]

This error is zero for a perfect sinusoid. The predictability is defined as

\[ \rho = \max(1 - \xi, 0). \]

A tonality index is defined as

\[ \tau = \sum_{i=1}^{I} \frac{\rho_i E_i}{E_T} \]

where \( I \) is the number of peaks in a critical band, \( E_i \) is the energy in three frequency bins around peak \( i \), and \( E_T \) is the total energy in the critical band.

The tonality index is 1 if all the peaks in a critical band are representing perfect sinusoids. Since the tonality level of some short tones may be under-estimated, we take the maximum value of the tonality index and the average value of the tonality index in three successive frames for the same critical band.

**APPENDIX B**

In this appendix we show that if a frame of the input signal is multiplied by a Hanning window with certain characteristics, the phase difference between the DFT in two frequency bins surrounding the frequency of the input signal is \( \pi \).

The N-point Hanning window is expressed as follows,

\[ w(n) = 0.5 \left( 1 - \cos \left( \frac{2\pi n}{N} \right) \right), \quad n = 0, ..., N - 1. \]

The DFT of the Hanning window is given by

\[ W(k) = 0.5 \delta(k) - 0.25\delta(k-1), \]

where \( \delta(.) \) denotes the Dirac delta function.

It is obvious from the DFT of the Hanning window, the phase difference between the two adjacent frequency bins is \( \pi \).

Similarly, it can be shown that this phase relationship holds for any window functions with the following characteristics,

\[ w(0) = 0, \quad w(n) = w(N - n), \quad n = 1, ..., N - 1, \]
\[ w(n) < w \left( \frac{N}{2} - n \right), \quad n = 1, ..., \frac{N}{4} - 1. \]