SINUSOIDAL MODELING OF AUDIO AND SPEECH USING PSYCHOACOUSTIC-ADAPTIVE MATCHING PURSUITS

Richard Heusdens

Dept. of Mediamatics,
Delft University of Technology,
2628 CD Delft, The Netherlands,
email: R.Heusdens@its.tudelft.nl

Renat Vafin and W. Bastiaan Kleijn

Dept. of Speech, Music and Hearing,
KTH (Royal Institute of Technology),
S-10044 Stockholm, Sweden,
email:{renat,bastiaan}@speech.kth.se

ABSTRACT

In this paper, we propose a segment-based matching pursuit algorithm where the psychoacoustical properties of the human auditory system are taken into account. Rather than scaling the dictionary elements according to auditory perception, we define a psychoacoustic-adaptive norm on the signal space which can be used for assigning the dictionary elements to the individual segments in a rate-distortion optimal manner. The new algorithm is asymptotically equal to signal-to-mask ratio based algorithms in the limit of infinite analysis window length. However, the new algorithm provides a significantly improved selection of the dictionary elements for finite window length.

1. INTRODUCTION

Sinusoidal coding has proven to be an efficient technique for the purpose of coding speech signals [1, 2]. More recently, it was shown that this method can also be exploited for low-rate audio coding [3, 4, 5, 6]. To account for the time-varying nature of the signal, the sinusoidal analysis/synthesis is done on a segment-by-segment basis, with each segment being modeled as a sum of sinusoids. The sinusoid parameters have been selected with a number of methods, including spectral peak-picking and analysis-by-synthesis. We focus on the matching pursuit algorithm [7], a particular analysis-by-synthesis method.

Matching pursuit approximates a signal by a finite expansion into elements (functions) chosen from a redundant dictionary. Let $\mathcal{H}$ be a Hilbert space and let $\mathcal{D} = (g_\gamma)_{\gamma\in\Gamma}$ be a complete dictionary of unit-norm elements in $\mathcal{H}$ ($\mathcal{H}$ is the closed linear span of the dictionary elements). The matching pursuit algorithm is a greedy iterative algorithm which projects a signal $x \in \mathcal{H}$ onto the dictionary element g_γ that best matches the signal and subtracts this projection to form a residual signal to be approximated in the next iteration. Let $R^{m-1}x$ denote the residual signal after iteration $m-1$. At iteration m, the algorithm decomposes $R^{m-1}x$ as

$$R^{m-1}x = \langle R^{m-1}x, g_{\gamma_m}\rangle g_{\gamma_m} + R^m x, \tag{1}$$

where $g_{\gamma_m} \in \mathcal{D}$ such that

$$|\langle R^{m-1}x, g_{\gamma_m}\rangle| = \sup_{\gamma\in\Gamma} |\langle R^{m-1}x, g_\gamma\rangle|. \tag{2}$$

The orthogonality of $R^m x$ and g_{γ_m} implies

$$\|R^{m-1}x\|^2 = |\langle R^{m-1}x, g_{\gamma_m}\rangle|^2 + \|R^m x\|^2.$$

To account for human auditory perception, the unit-norm dictionary elements can be scaled [6], which is equivalent to scaling the inner products in (2). We will refer to this method as the *weighted matching pursuit algorithm*. While this method performs well, it will be shown below that it does not provide a consistent selection measure for elements of finite time support and for elements in different signal segments.

To address these issues, we introduce a matching pursuit algorithm where psychoacoustical properties are taken into account by defining a proper norm on the signal space. The norm changes at each iteration. In contrast to the weighted matching pursuit algorithm, this new *psychoacoustic-adaptive matching pursuit* algorithm has the desired property that if the signal is identical to one of the dictionary elements, this element is always selected. Moreover, it is able to discriminate between peaks originating from true sinusoidal components and peaks originating from side lobes of the analysis window function. Using this new algorithm, the norm of the residual signal will converge exponentially to zero when the number of iterations approaches infinity.

This paper is organized as follows. In Section 2 we introduce the psychoacoustic-adaptive matching pursuit algorithm and discuss its relation to signal-to-mask ratio based algorithms. Next, in Section 3, we show that our newly proposed algorithm can be implemented efficiently by using Fourier transforms. Finally, in Section 4, we draw some conclusions.

2. PSYCHOACOUSTIC-ADAPTIVE MATCHING PURSUITS

Ignoring time-domain masking phenomena, signal distortion becomes audible when the log power spectrum of the residual signal Rx exceeds the log frequency masking threshold, or equivalently, when the ratio of the power spectrum and the masking threshold exceeds unity. This motivates us to define a perceptual distortion measure as

$$\|Rx\|^2 = \int_0^1 \hat{a}(f)|(\widehat{wRx})(f)|^2 df, \tag{3}$$

where ˆ indicates the Fourier transform operation, w is a window defining the signal segment, and $\hat{a}$ is a weighting function representing the sensitivity of the human auditory system which we

The research is supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

will generally select to be the inverse of the masking threshold. The distortion measure (3) defines a norm on $\mathcal{H}$ if $\hat{a}(f)$ is positive and real for all $f \in [0,1)$ and $wx \neq 0$ for all $x \in \mathcal{H}$. The norm is induced by the inner product

$$\langle x, y\rangle = \int_0^1 \hat{a}(f)(\widehat{wx})(f)(\widehat{wy})^*(f)\,df, \qquad (4)$$

facilitating the use of the distortion measure in selecting the best matching dictionary element in a matching pursuit algorithm.

The masking threshold is based on the reconstructed signal which changes with each iteration. Therefore, the norm on $\mathcal{H}$ must be adapted with each iteration. Let $\hat{a}_{m-1}$ be the weighting function used at iteration m and let $\|\cdot\|_{\hat{a}_{m-1}}$ denote the corresponding norm. We thus minimize $\|R^m x\|_{\hat{a}_{m-1}}$ at iteration m, update $\hat{a}_{m-1}$ to $\hat{a}_m$ using the newly chosen dictionary element, and then minimize $\|R^{m+1}x\|_{\hat{a}_m}$ in the next iteration. The convergence properties of this algorithm are described by the following theorem (proven in [8]):

Theorem 1 *There exists a* $\lambda > 0$ *such that for all* $m > 0$

$$\|R^m x\|_{\hat{a}_m} \leq 2^{-\lambda m}\|x\|_{\hat{a}_0},$$

if and only if for all $m > 0$, $\hat{a}_m(f) \leq \hat{a}_{m-1}(f)$ *for all* $f \in [0,1)$.

Note that if $\hat{a}_{m-1}$ is the reciprocal of the frequency masking threshold at iteration m, then the condition $\hat{a}_m(f) \leq \hat{a}_{m-1}(f)$ for all $f \in [0,1)$ is satisfied since the masking threshold increases with the iteration number.

Let us consider the case where the dictionary $\mathcal{D} = (g_\gamma)_{\gamma\in\Gamma}$ consists of complex exponentials,

$$g_\gamma = \frac{1}{\sqrt{N}} e^{i2\pi\gamma n}, \quad n = 0, \ldots, N-1, \qquad (5)$$

for $\gamma \in [0,1)$. To find the best matching exponential at iteration m, we compute the inner products of $R^{m-1}x$ and the dictionary elements,

$$\langle R^{m-1}x, g_\gamma\rangle = \frac{1}{\sqrt{N}} \int_0^1 \hat{a}_{m-1}(f)(\widehat{wR^{m-1}x})(f)\hat{w}^*(f-\gamma)\,df. \qquad (6)$$

For the case $N \rightarrow \infty$, the function $\hat{w}$ becomes a δ-function, or Dirac, and (6) reduces to

$$\langle R^{m-1}x, g_\gamma\rangle = \frac{1}{\sqrt{N}}\hat{a}_{m-1}(\gamma)(\widehat{R^{m-1}x})(\gamma). \qquad (7)$$

Hence, the matching pursuit algorithm selects $g_\gamma \in \mathcal{D}$ such that

$$|\langle R^{m-1}x, g_{\gamma_m}\rangle| = \frac{1}{\sqrt{N}} \sup_{\gamma\in\Gamma} |\hat{a}_{m-1}(\gamma)(\widehat{R^{m-1}x})(\gamma)|.$$

Note that since (6) reduces to a simple scaling of the dictionary elements for $N \rightarrow \infty$, the psychoacoustic-adaptive and weighted matching pursuit methods give identical results. If $\hat{a}_{m-1}$ is the reciprocal of the masking threshold at iteration m, both procedures select the exponential located where the absolute difference of the log residual signal spectrum and the log masking threshold is largest.

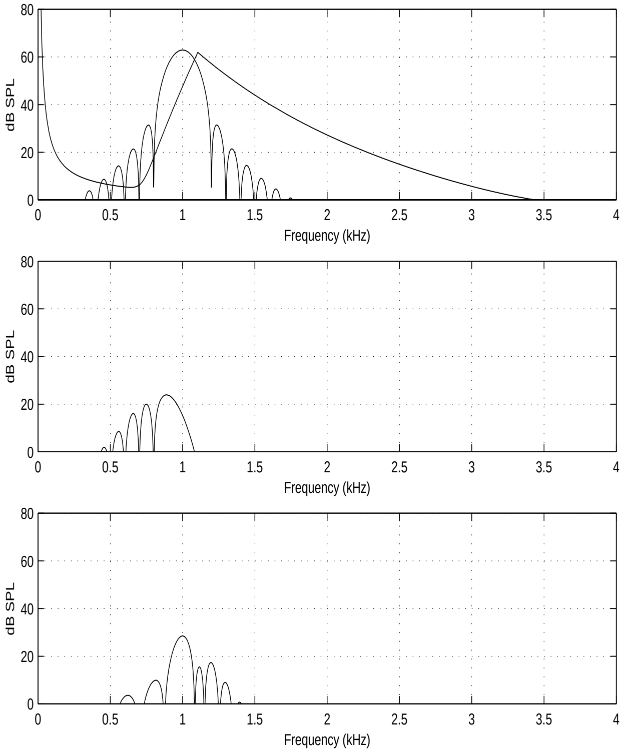

Figure 1: *Example of selecting sinusoidal components using the weighted (middle plot) and psychoacoustic-adaptive (lower plot) matching pursuit algorithm.*

The psychoacoustic-adaptive matching pursuit method has advantages over the weighted matching pursuit method when the signal segment is of finite length. To see this, we first take the signal segment to be a scaled version of one of the dictionary elements, say $x = \alpha g_\gamma$. The psychoacoustic-adaptive matching pursuit method will select g_γ as desired (from the Cauchy-Schwartz inequality we have that $|\langle x, g_\gamma\rangle| \leq \|x\|\|g_\gamma\| = \|x\|$, with equality if and only if x and g_γ are linearly dependent). This is not true for the weighted matching pursuit method. Figure 1 illustrates an example where the original signal contains two sinusoids, at 1 and 1.1 kHz, respectively, with the residual signal after one iteration consisting of the $f = 1$ kHz sinusoid. The upper plot shows the projection energy (in the l_2-sense) $|\langle x, g_\gamma\rangle|^2 = \frac{1}{N}|(\widehat{wx})(f)|^2$ and the masking threshold. The middle subplot shows the projection energy for the weighted matching pursuit algorithm, which corresponds to the signal-to-mask ratio (the difference between the log residual signal spectrum and the log masking threshold of the upper subplot). The lower subplot shows the projection energy $|\langle x, g_\gamma\rangle|^2$ according to the inner product defined by (4). The steep slope of the masking threshold around $f = 1$ kHz causes the weighted matching pursuit algorithm to select a suboptimal solution, whereas the psychoacoustic-adaptive matching pursuit algorithm correctly selects a $f = 1$ kHz sinusoid.

A second advantage of the psychoacoustic-adaptive matching pursuit method is that it discriminates between main lobes and side lobes in a spectrum of a sum of (windowed) sinusoids, as is illustrated in Figures 2 and 3. The upper and lower plots of Figure 2 show the results for the weighted and psychoacoustic-adaptive matching pursuit methods, respectively, for a rectangu-

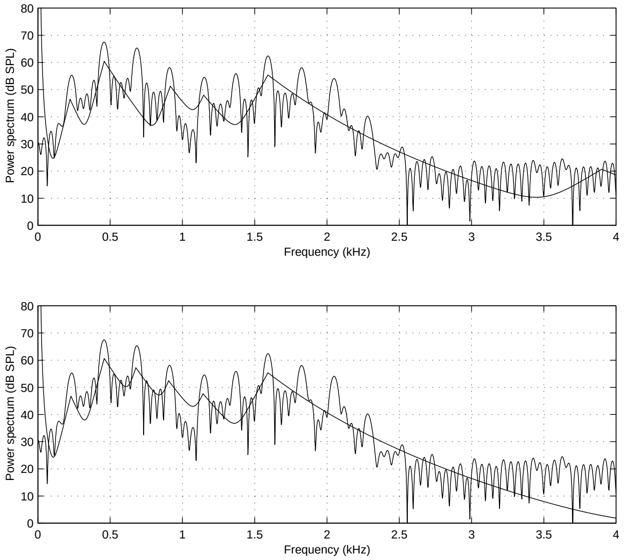

Figure 2: *Selection of 6 sinusoidal components using the weighted (upper plot) and psychoacoustic-adaptive (lower plot) matching pursuit algorithm for a 20 ms long voiced speech fragment.*

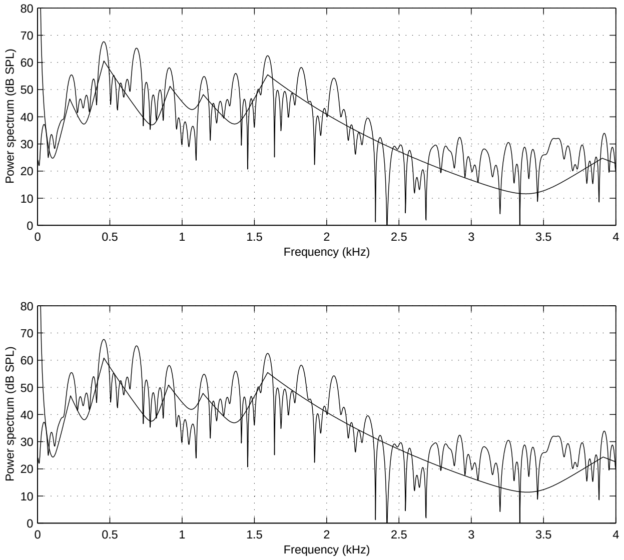

Figure 3: *Selection of 6 sinusoidal components using the weighted (upper plot) and psychoacoustic-adaptive (lower plot) matching pursuit algorithm for a 20 ms long voiced speech fragment plus zero-mean white Gaussian noise.*

larly windowed input signal (20 ms of voiced speech sampled at 8 kHz). The plots show the power spectrum of the input signal and the masking threshold after selecting six sinusoidal components. The weighted matching pursuit method selects a component at 3.8 kHz corresponding to a side lobe. This in contrast to the psychoacoustic-adaptive matching pursuit method which selects a peak corresponding to a true sinusoidal component. To show that this difference does not result from a preference of selecting low-frequency components, we added zero-mean white Gaussian noise to the 20 ms speech fragment in the example of Figure 3. In this case both methods perform similarly.

3. EFFICIENT IMPLEMENTATION

In general, the size of the dictionary used in the matching pursuit is large so that the computational complexity of computing the inner products in (2) becomes considerable. Since $\hat{a}$ changes at each iteration, there does not exist a simple updating formula, except for cases where $\hat{a}$ is independent of the iteration m. In that case we have, by taking the inner products with g_γ on both the left and right-hand side of (1), that

$$\langle R^m x, g_\gamma \rangle = \langle R^{m-1}x, g_\gamma \rangle - \langle R^{m-1}x, g_{\gamma_m} \rangle \langle g_{\gamma_m}, g_\gamma \rangle .$$

Hence, the only computations required for this update are the computations of the inner products $\langle g_{\gamma_m}, g_\gamma \rangle$, which can be computed beforehand and stored in memory.

Notwithstanding the existence of an efficient updating formula in some cases, the computational complexity of computing the inner products $\langle R^m x, g_\gamma \rangle$ can be large for unstructured dictionaries. However, if we take the dictionary elements as defined by (5), we can reduce the computational load by using the Fourier transform to compute the inner products. Indeed, we have that

$$\begin{aligned} \langle R^m x, g_\gamma \rangle &= \frac{1}{\sqrt{N}} \int_0^1 \hat{a}_m(f)(w\hat{R^m}x)(f)\hat{w}^*(f-\gamma)df \\ &= \frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} \left(\int_0^1 \hat{a}_m(f)(w\hat{R^m}x)(f)e^{i2\pi fn}df \right) \\ &\qquad \cdot w^*(n)e^{-i2\pi\gamma n}. \end{aligned} \quad (8)$$

Hence, to compute $\langle R^m x, g_\gamma \rangle$ for all γ, we first compute the Fourier transform of $wR^m x$, multiply the result by $\hat{a}_m$, compute the inverse Fourier transform of this product, multiply this result by w^*, and finally compute the Fourier transform of the result thus obtained to get the desired result. By doing so, (8) can be computed using three Fourier transforms.

In many cases, like modeling of audio and speech signals, we want to model the signals by real-valued sinusoids, rather than complex exponentials. Since a real-valued sinusoid can be expressed as a sum of a complex exponential and its complex conjugate, we can select the best matching sinusoid by using a dictionary consisting of complex exponentials only. By doing so, we have to find the best matching set of dictionary elements, say (g_γ^*, g_γ). In order to find the optimal set (g_γ^*, g_γ), we not only have to compute the inner products $\langle R^m x, g_\gamma \rangle$, but the inner products $\langle g_\gamma^*, g_\gamma \rangle$ as well [3, 6].

Unfortunately, the computations of $\langle g_\gamma^*, g_\gamma \rangle$ do not have an efficient implementation in terms of Fourier transforms. However, since the dictionary elements g_γ and g_γ^* have a sparse interaction, that is, $\langle g_\gamma^*, g_\gamma \rangle \approx 0$ except for γ close to 0 or 1/2, it is in general sufficient to compute these inner product for a limited set of γs only. This is illustrated by the result of Figure 4. The upper plot shows the projection energy for approximating a 20 ms long

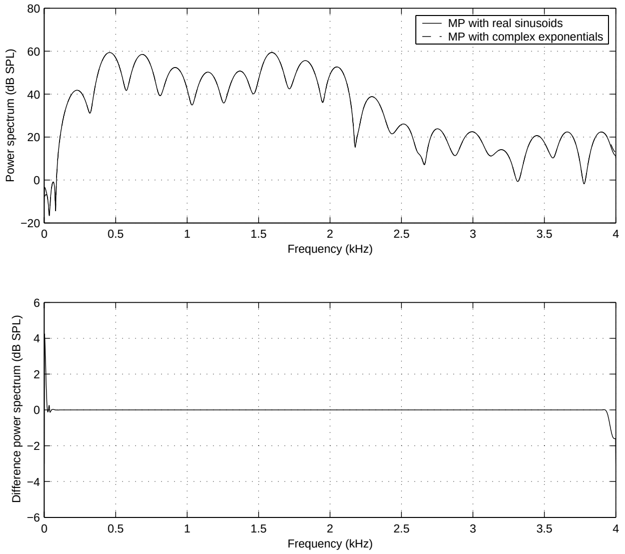

Figure 4: *Difference of using a dictionary consisting of real-valued sinusoids or complex exponentials. The upper plot shows the projection energy of both situations. The lower plot shows the difference (in dBs).*

Hanning windowed voiced speech fragment (sampling frequency 8 kHz) with real-valued sinusoids (solid line) and complex exponentials (dashed line). The latter one can be obtained by setting $\langle g_\gamma^*, g_\gamma \rangle = 0$ for all γ, so that the difference of the two methods, which is depicted in the lower subplot of Figure 4, gives information about how the dictionary elements g_γ and g_γ^* interact. Clearly, these differences are almost zero, except for frequencies aroud 0 and 4 kHz. We, therefore, conclude that the computational complexity for finding the best matching set (g_γ^*, g_γ) is of the same order of magnitude as the one for finding the best matching exponential g_γ, which is three Fourier transform operations.

4. CONCLUSIONS

We proposed a segment-based matching pursuit algorithm which takes psychoacoustical properties into account. Rather than scaling the dictionary elements according to auditory perception, we define a psychoacoustic-adaptive norm on the signal space which can be used for assigning the dictionary elements to the individual segments in a rate-distortion optimal manner. The norm changes with each iteration because of the change in auditory perception. We showed that, if the analysis window length approaches infinity, the algorithm becomes equal to algorithms which select components based on signal-to-mask ratios. For finite window lengths the algorithms behave differently. In contrast to signal-to-mask ratio based methods, the new method selects the correct element when the signal under consideration is a scaled version of a particular dictionary element. Moreover, we showed that the proposed method discriminates between peaks originating from true sinusoidal components and peaks originating from side lobes of the analysis window function, and that the computational complexity is of the order of magnitude of three Fourier transform operations.

5. REFERENCES

[1] R.J. McAulay and T.F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. on ASSP*, vol. 34, no. 4, pp. 744–754, August 1986.

[2] R.J. McAulay and T.F. Quatieri, “Sinusoidal coding,” in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Eds., chapter 4, pp. 121–174. Elsevier Science Publishers, Amsterdam, 1995.

[3] M. Goodwin, “Matching pursuit with damped sinusoids,” in *Proceedings ICASSP’97*, Munich, Germany, May 1997, vol. 3, pp. 2037–2040.

[4] J. Nieuwenhuijse, R. Heusdens, and E.F. Deprettere, “Robust exponential modeling of audio signals,” in *Proceedings ICASSP’98*, Seattle, Washington, USA, May 1998, vol. 6, pp. 3581–3584.

[5] K. Vos, R. Vafin, R. Heusdens, and W.B. Kleijn, “High-quality consistent analysis-synthesis in sinusoidal coding,” in *Proceedings of the AES 17th International Conference*, Florence, Italy, September 1999, pp. 244–250.

[6] T.S. Verma and T.H.Y. Meng, “Sinusoidal modeling using frame-based perceptually weighted matching pursuits,” in *Proceedings ICASSP’99*, Phoenix, Arizona, USA, May 1999, vol. 2, pp. 981–984.

[7] S.G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[8] R. Heusdens, R. Vafin, and W.B. Kleijn, “Sinusoidal modeling using psychoacoustic-adaptive matching pursuits,” November 2000, Submitted to IEEE Signal Processing Letters.