Sinusoidal Modeling Using Psychoacoustic-Adaptive Matching Pursuits

Richard Heusdens, Renat Vafin, and W. Bastiaan Kleijn, *Fellow, IEEE*

***Abstract*—In this letter, 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 that can be used for assigning the dictionary elements to the individual segments in a rate-distortion optimal way. 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.**

***Index Terms*—Audio/speech coding, matching pursuit, psychoacoustics, sinusoidal modeling.**

I. Introduction

SINUSOIDAL CODING has proven to be an efficient technique for the purpose of coding speech signals [1], [2, Ch. 4, pp. 121–174]. More recently, it was shown that this method can also be exploited for low-rate audio coding [3]–[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 sinusoidal parameters can be estimated with a number of methods, including spectral peak-picking and analysis-by-synthesis. We focus on the matching-pursuit algorithm [7] that is 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 that 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

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

Manuscript received July 30, 2001; revised May 13, 2002. This research was supported by Philips Research and the Technology Foundation STW, Applied Science Division of NWO, and the Technology Programme of the Ministry of Economics Affairs. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Steven L. Gay.

R. Heusdens is with the Department of Mediamatics, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: R.Heusdens@its.tudelft.nl).

R. Vafin and W. B. Kleijn are with the Department of Speech, Music, and Hearing, Royal Institute of Technology, 100 44 Stockholm, Sweden.

Publisher Item Identifier 10.1109/LSP.2002.802999.

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

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

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 (1). We will refer to this method as the weighted matching pursuit (WMP) 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 accounted for by a norm on the signal space. The norm changes at each iteration. In contrast to the WMP algorithm, this new psychoacoustic-adaptive matching pursuit (PAMP) algorithm has the desired property that if the signal is a scaled version of one of the dictionary elements, this element is always selected. Moreover, in the new algorithm, the norm of the residual signal converges exponentially to zero when the number of iterations approaches infinity.

II. Psychoacoustic-Adaptive Matching Pursuit

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. Hence, to represent an audio/speech signal without audible artifacts at the lowest possible bit rate, we have to shape the residual signal spectrum such that it equals the frequency-masking threshold. More generally, if we allow some audible distortion, we assume, in line with models of partial loudness [8], that distortions are integrated over the entire spectrum. This motivates us to define a perceptual distortion measure as

$$\|Rx\|^2 = \int_0^1 \hat{a}(f)|(w\hat{R}x)(f)|^2 df \tag{2}$$

where $\hat{}$ 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 select to be the inverse of the masking threshold. By doing so, regions in which the auditory system is less sensitive will contribute less to the total distortion as compared with regions in which the auditory system is more

sensitive. The distortion measure (2) 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

$$\langle x, y \rangle = \int_0^1 \hat{a}(f)(\hat{wx})(f)(\hat{wy})^*(f)df \tag{3}$$

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 (the signal expansion) that changes with each iteration. Thus, 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 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 the Appendix).

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

$$\|R^m x\|_{\hat{a}_m} \leq 2^{-\lambda m}\|x\|_{\hat{a}_0} \tag{4}$$

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 since $\hat{a}_{m-1}$ is the reciprocal of the frequency-masking threshold used at iteration m, 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.

To see how the PAMP algorithm performs, let us consider the case where the dictionary $\mathcal{D} = (g_\gamma)_{\gamma\in\Gamma}$ consists of complex exponentials

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

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

$$\langle R^m x, g_\gamma \rangle = \frac{1}{\sqrt{N}} \int_0^1 \hat{a}_m(f)(w\hat{R}^m x)(f)\hat{w}^*(f-\gamma)df. \tag{5}$$

For the case $N \to \infty$, the function $\hat{w}$ becomes a δ-function, or Dirac, and (5) reduces to

$$\langle R^m x, g_\gamma \rangle = \frac{1}{\sqrt{N}} \hat{a}_m(\gamma)(\hat{R^m x})(\gamma).$$

Hence, the matching pursuit algorithm selects $g_{\gamma_{m+1}} \in \mathcal{D}$ such that

$$|\langle R^m x, g_{\gamma_{m+1}} \rangle| = \frac{1}{\sqrt{N}} \sup_{\gamma\in\Gamma} |\hat{a}_m(\gamma)(\hat{R^m x})(\gamma)|. \tag{6}$$

Since $\hat{a}_m$ is the reciprocal of the masking threshold at iteration $m+1$, we conclude, therefore, that for $N \to \infty$, the PAMP selects the exponential located where the ratio of the power spectrum and the masking threshold is largest, which is consistent with the way the human auditory system works [9], [10]. Therefore, the PAMP and the WMP algorithm give identical results for infinite window length.

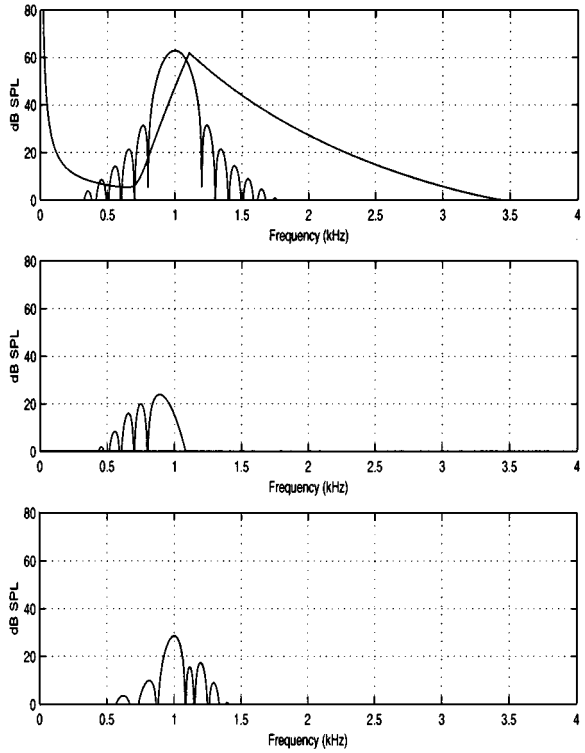

Fig. 1. Example of selecting sinusoidal components using the WMP (middle plot) and PAMP (lower plot) algorithms.

The PAMP method has advantages over the WMP 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 PAMP method will select g_γ as desired (from the Cauchy–Schwarz 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 WMP method. Fig. 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) $|\langle x, g_\gamma \rangle|^2 = 1/N|(\hat{wx})(f)|^2$ and the masking threshold. The middle subplot shows the projection energy for the WMP 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 (3). The steep slope of the masking threshold around $f = 1$ KHz causes the WMP algorithm to select a suboptimal solution, whereas the PAMP algorithm correctly selects a $f = 1$ KHz sinusoid.

A second advantage of the PAMP method is that it discriminates between main lobes and side lobes in a spectrum of a sum of (windowed) sinusoids, as shown in Figs. 2 and 3. The upper and lower plots of Fig. 2 show the results for the WMP and PAMP methods, respectively, for a rectangularly 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 WMP method has selected a component at 3.8 KHz corresponding to a side lobe. This contrasts with the PAMP

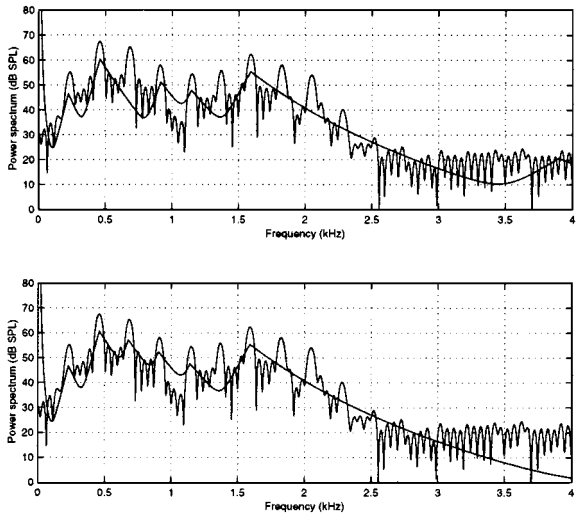

Fig. 2. Selection of six sinusoidal components using the WMP (upper plot) and PAMP (lower plot) algorithms for a 20-ms long-voiced speech fragment.

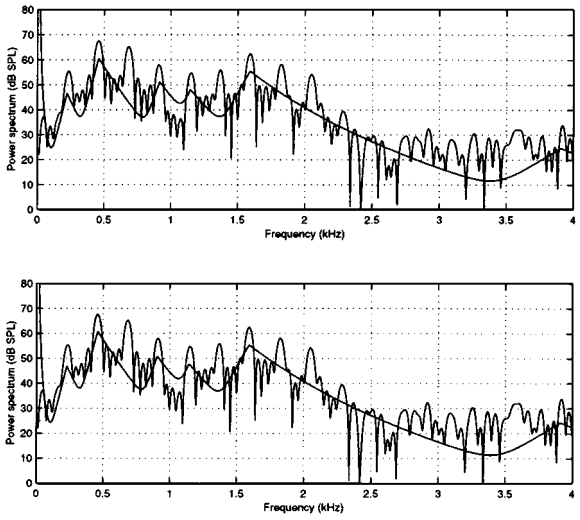

Fig. 3. Selection of six sinusoidal components using the WMP (upper plot) and PAMP (lower plot) algorithms for a 20-ms long-voiced speech fragment plus zero-mean white Gaussian noise.

method that selected 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 Fig. 3. In this case, both methods select the same spectral peaks. It should be noted, however, that the frequency estimation of the spectral peaks is still better with the PAMP, especially at low frequencies where the slope of the masking curve is steep around the peaks.

III. Experimental Results

In this section, we present results obtained by computer simulations and listening tests with audio and speech signals. The signals are mono, sampled at 48 KHz, where each sample is represented by 16 bits. The test excerpts are a harpsichord solo, Suzanne Vega, contemporary pop music, castanets, clean German male speech, and clean English female speech. We used a dictionary consisting of real-valued sinusoids. The analysis/synthesis was done on a segment-by-segment basis using a 50%-overlap 21.3-ms Hanning window.

TABLE I
Results of the Listening Test. The Left Column Indicates the Excerpt, the Right Column the Percentage of Preference for the PAMP Method Over the WMP Method

excerpt	duration (s)	preference for PAMP (%)
harpsichord	17.3	94
Suzanne Vega	10.5	100
pop music	15.7	90
castanets	15.0	56
male speech	8.8	100
female speech	7.5	100

To compare performance of the PAMP and WMP methods, each segment was modeled by 25 sinusoids. We performed a subjective listening test in which signal triplets OAB were presented to the listeners. Here, O is the original signal; A or B is the modeled signal using the WMP method; and B or A is the modeled signal using the PAMP method. The task of the listener was to indicate which signal (A or B) is closer to the original. For each test excerpt, the triplets OAB were presented five times, and the position of the modeled signal using the WMP and PAMP methods was changed randomly each time. Ten listeners participated in the test among which five were experienced (the authors not included). The results averaged over all listeners are shown in Table I. Except for the castanet excerpt, the PAMP method performs significantly better than the WMP method.

Appendix
Proof of Theorem 1

A. Proof

Since the dictionary $\mathcal{D} = (g_\gamma)_{\gamma\in\Gamma}$ is complete, there exists, at each iteration, $\alpha > 0$ such that for any $x \in \mathcal{H}$

$$|\langle R^{m-1}x, g_{\gamma_m}\rangle| \geq \alpha \|R^{m-1}x\|. \tag{7}$$

Using (7) and the orthogonality relation between $R^m x$ and g_{γ_m}, we conclude that at each iteration

$$\|R^m x\|_{\hat{a}_{m-1}} \leq (1-\alpha^2)^{1/2}\|R^{m-1}x\|_{\hat{a}_{m-1}}. \tag{8}$$

Next, assume that $\hat{a}_m(f) = \hat{a}_{m-1}(f) + \hat{\varepsilon}(f)$ for all f. We conclude, using (2), that for all $x \in \mathcal{H}$

$$\begin{aligned}\|R^m x\|^2_{\hat{a}_m} &= \|R^m x\|^2_{\hat{a}_{m-1}} + \int_0^1 \hat{\varepsilon}(f)|(w\hat{R^m}x)(f)|^2 df \\ &\leq \|R^m x\|^2_{\hat{a}_{m-1}},\end{aligned}$$

if and only if $\hat{\varepsilon}(f) \leq 0$ for all $f \in [0,1)$. Combining this with (8), we have that

$$\|R^m x\|_{\hat{a}_m} \leq (1-\alpha^2)^{1/2}\|R^{m-1}x\|_{\hat{a}_{m-1}}$$

so that for $2^{-\lambda} = (1-\alpha^2)^{1/2} < 1$ we readily obtain (4), which completes the proof. ■

REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[2] ——, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995.

[3] M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. ICASSP*, vol. 3, Munich, Germany, May 1997, pp. 2037–2040.

[4] J. Nieuwenhuijse, R. Heusdens, and E. F. Deprettere, "Robust exponential modeling of audio signals," in *Proc. ICASSP*, vol. 6, Seattle, WA, May 1998, pp. 3581–3584.

[5] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. AES 17th Int. Conf.*, Florence, Italy, Sept. 1999, pp. 244–250.

[6] T. S. Verma and T. H. Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. ICASSP*, vol. 2, Phoenix, AZ, May 1999, pp. 981–984.

[7] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

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

[9] E. Zwicker and H. Fastl, "Springer Series in Information Sciences," in *Psychoacoustics*, 2nd ed. Berlin, Germany: Springer-Verlag, 1999.

[10] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. Orlando, FL: Academic, 1997.