PERCEPTUAL MATCHING PURSUIT WITH GABOR DICTIONARIES AND TIME-FREQUENCY MASKING

Gilles Chardon, Thibaud Necciari, and Peter Balazs

Acoustics Research Institute
Austrian Academy of Sciences
Wohllebengasse 12–14, A-1040 Vienna, Austria

ABSTRACT

This paper describes a method to obtain a perceptually relevant sparse representation of a sound signal. Based on matching pursuit (MP) and recent psychoacoustic data on time-frequency masking measured with Gabor atoms, a perceptual matching pursuit (PMP) algorithm is proposed. To obtain a good match between the masking model and the signal representation, a dictionary of Gabor atoms with variable sizes is chosen for MP. In the proposed method, the signal is first decomposed using MP and the masking model is applied on the resulting set of atoms. This allows for isolating the masked components from the residual. Experimental results show that exploiting time-frequency masking allows to remove more atoms than using only spectral masking. Additionally, accounting for masking effects between atoms of different sizes and at different times allows for sparser representations. The objective evaluation of the proposed PMP algorithm indicates imperceptible distortions.

Index Terms— matching pursuit, auditory masking, sparse representations.

1. INTRODUCTION

This study addresses the combination of sparse representation of sounds and perceptual masking models. Sparse representations extracts relevant information from the signal and describes it with a minimal amount of data. In the context of audio processing, it is desirable that these representations take human auditory perception into account and allow reconstruction, not with a low error (say, error in the ℓ_2 norm), but with a controlled amount of perceived distortion (see e.g. [1]).

In the context of signal representation, time-frequency (TF) representations like the Gabor or Wavelet transforms have become standard tools. They allow to decompose signals into a set of elementary functions called "TF atoms" with good TF localization and achieve perfect reconstruction if the transform parameters are chosen appropriately (e.g., [2]). The set of TF atoms generally consists of scaled, translated and modulated versions of a single window function $g(t) \in L^2(\mathbb{R})$, where $L^2(\mathbb{R})$ is the Hilbert space of complex valued functions. In other words, the generation of the set follows a fixed preset rule and does not necessarily adapt to the signal structures or the auditory perception (note, however, that auditory-based TF transforms have been proposed in, e.g., [3,4]). Moreover, TF transforms usually yield redundant representations. Methods based on sparsity, such as matching pursuits (MPs), use this redundancy to decompose signals into a set of functions that are iteratively selected among a large dictionary of waveforms [5]. MPs yield sparse signal representations that allow interpreting the signal structures. MPs have found many applications in audio processing, especially in parametric audio coding (e.g., [6,7]). To account for auditory perception in MP, psychoacoustic criteria (e.g., hearing threshold in quiet, auditory masking or loudness models) can be included in the process; this is usually referred to as "perceptual MP" (PMP, see Sec. 2).

In this work, we propose a new PMP algorithm that combines a dictionary of Gabor atoms with various window sizes and a *matched* TF masking model for Gabor atoms [8]. To isolate the masked components from the residual, we apply the masking model *after* MP. We show that using a TF masking model allows to eliminate more atoms than using only a simple spectral masking model. The TF model is also extended to the case of multiple atom lengths, allowing to combine dictionaries with various atom lengths and auditory masking.

2. PRIOR WORK

The general idea of PMP is the following: Adaptively find the perceptually most significant (i.e., *audible*) components in the signal to obtain a perceptually relevant sparse representation. This is usually achieved by including a masking model in MP. Most PMP algorithms exploit only spectral masking [6, 9–13]. TF masking is exploited in [14, 15] using mod-

This work was supported by the Austrian Science Fund (FWF) START-project FLAME ("Frames and Linear Operators for Acoustical Modeling and Parameter Estimation"; Y 551-N13) and the FWF-ANR project POTION ("Perceptual Optimization of Time-Frequency Representations and Coding"; I 1362-N30).

els that are based on a simple superposition of spectral and temporal masking functions measured for stimuli that do not have good TF localizations (typically, long-lasting sinusoids or noise bands). In [8], it was shown that such simple models do not provide an accurate representation of the measured TF masking function for TF atoms. Thus, it is possible that the amount of components identified as "masked" in current PMPs is underestimated. Moreover, in most PMPs the psychoacoustic criteria are considered at each iteration of MP. Consequently, the masked components are mixed with the residual. For signal analysis purposes, a separation of the original signal into a relevant, masked and residual part is of high interest. To isolate the masked components from the residual, the selection of components has to be performed after MP [6, 11].

In most of the cited approaches, the MP algorithm was applied separately on short-length frames. It is however, as shown in [6], more efficient to use MP with a dictionary based on MDCT of multiple lengths. In [6], a masking model was considered, but applied only between atoms sharing the same length. We will follow a similar approach with Gabor atoms of various lengths and show that applying the masking model between atoms of different sizes yields sparser representations.

3. PROPOSED METHOD

The basic idea of our method is the following: First, perform a MP decomposition using a redundant dictionary of Gabor atoms. Second, apply a TF masking model on the set of atoms selected by MP to keep only the audible atoms.

3.1. Preliminaries

The relationship between the linear and the psychoacoustic ERB frequency scale is given by [16]

$$ERB_{\text{num}}(\nu) = 9.265 \ln\left(1 + \frac{\nu}{228.8455}\right) \tag{1}$$

with ν expressed in Hz.

3.2. Time-frequency masking model

To accurately predict the audibility of each atom in a TF decomposition, it is important to have a TF masking model that is valid for such atoms. In [8], psychoacoustic experiments were conducted using Gaussian TF atoms of the form

$$s_i(t) = \sin\left(2\pi\nu_i t + \frac{\pi}{4}\right) e^{-\pi(\Gamma t)^2} \tag{2}$$

where ν_i is the center frequency (in Hz), Γ is a shape factor that controls the duration and bandwidth of $s_i(t)$ and the $\pi/4$ phase shift allows maintaining the energy constant $\forall\, \nu_i$. For all signals Γ was fixed at 600 s^{-1}. The amounts of masking (AM, in dB) were measured for several masker-target combinations

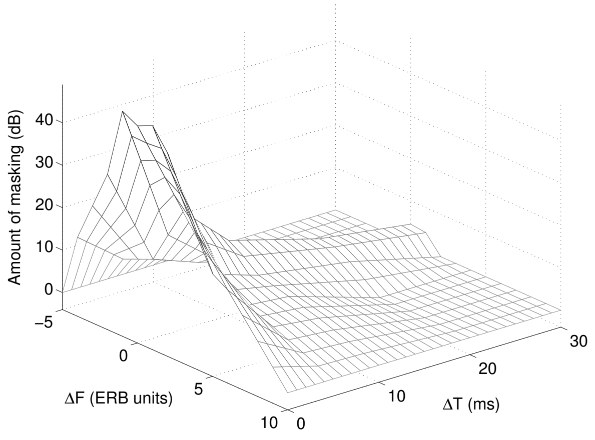


Fig. 1. Mean TF masking data plotted in the TF plane [8] for a masker at level 60 dB.

$$S_{\Delta F,\Delta T}(t) = s_M(t) + s_T(t - \Delta T)$$

with $\Delta F = \nu_T - \nu_M$. The frequency of the masker signal $s_M(t)$ was fixed at ν_M = 4000 Hz and ν_T varied from 2521 to 7835 Hz. Using Eq. (1), the ν_T values correspond to $\Delta F \in [-4; +6]$ ERB units relative $ERB_{\text{num}}(\nu_M)$. ΔT varied from 0 to 30 ms. The mean results are plotted in Fig. 1. These results can be described by the following model

$$AM(\Delta T, \Delta F) = C(\Delta F)\, e^{-\Delta T/\lambda(\Delta F)} \tag{3}$$

where C describes the spectral spread of masking at ΔT = 0 and λ is a time constant that characterizes the frequency-dependent temporal decay of forward masking. Based on polynomial fits of the experimental data, C and λ are defined by [8]

$$C(\Delta F) = \begin{cases} +11.1\,\Delta F + 58.0 & \text{if} \quad \Delta F < 0 \\ -6.4\,\Delta F + 55.4 & \text{if} \quad \Delta F \geq 0 \end{cases}$$

and

$$\lambda(\Delta F) = \begin{cases} +0.43\,\Delta F^3 + 3.4\,\Delta F^2 + & \\ +7.1\,\Delta F + 8.7 & \text{if} \quad \Delta F < 0 \\ -0.05\,\Delta F^3 + 0.75\,\Delta F^2 - & \\ -3.2\,\Delta F + 8.8 & \text{if} \quad \Delta F \geq 0 \end{cases}$$

Noteworthy, in Eq. (3), ΔF is defined in ERB units and ΔT in ms.

length	N_{FFT}	shift
128 (3 ms)	8192	64
256 (6 ms)	8192	64
512 (11 ms)	8192	64
1024 (23 ms)	8192	64
2048 (46 ms)	8192	64

Table 1. Parameters (in samples) used for the construction of the dictionary in MPTK.

3.3. Time-frequency (Gabor) dictionary for MP

To obtain a perfect match between the TF masking model and the TF representation, we opt for a dictionary of complex TF atoms (also called Gabor dictionary) of the form

$$g_{a,\nu,\Gamma}(t) = K\, e^{-\pi(\Gamma(t-a))^2} e^{2i\pi\nu t} \qquad (4)$$

where K is a normalization constant. Following [6], we use Gabor atoms with multiple window lengths. This allows a sparser representation of the signal, as such a dictionary can represent both harmonic components and transients in an efficient way.

3.4. Algorithm formulation

The MP algorithm builds a sparse representation of the signal s in the following way:

1. initialize the residual $r_0 = s$ and build the dictionary using a set of parameters (a_l, ν_l, Γ_l).

2. compute the correlations between the residual and the atoms of the dictionary, $c_l = | < r_n, g_{a_l,\nu_l,\Gamma_l} > |$

3. select the atom maximizing the correlation and remove it from the residual, $r_{n+1} = r_n - < r_n, g_{a_l,\nu_l,\Gamma_l} > g_{a_l,\nu_l,\Gamma_l}$.

4. repeat from 2. until the number of iterations N is reached.

The output of MP is a list of parameters (length, frequency, amplitude and phase) of the atoms identified by the algorithm. The signal can be re-synthesized by summing these atoms. Once the list of identified atoms is obtained, the atoms are ordered by decreasing amplitude and the masking model in Eq. (3) is applied to each atom starting with the greatest amplitude. All atoms with a length equal or smaller than that of the atom considered as "masker" and below the masking level are removed from the set. Since the masking model is applied after MP, we can use a standard implementation of MP such as MPTK [17].

% atoms				
removed	PMP F 1	PMP F 2	PMP TF 1	PMP TF 2
maderna	35	46	40	51
vega	36	53	54	66

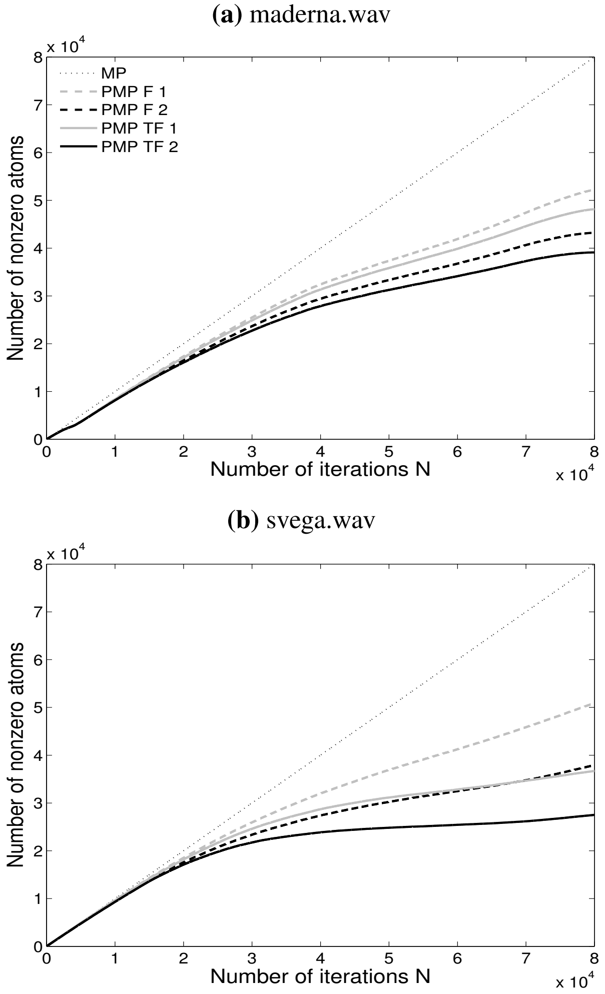
Table 2. Percentages of atoms removed from the set of atoms identified by MP after 80 000 iterations for each variant of PMP.

4. RESULTS

To evaluate the performance of the proposed method, we ran the PMP algorithm on two musical excerpts sampled at 44.1 kHz: the piano concerto from Bruno Maderna (length = 3 s) and Suzanne Vega (4 s). The parameters used in MPTK for the MP decomposition are listed in Tab. 1. The total number of iterations N was fixed at 80 000. This number ensures a perfect signal reconstruction using MP (SNR = -78 dB for vega, -37 dB for maderna). Moreover, to evaluate the contribution of temporal masking, we tested four variants of PMP. In the first variant, PMP F 1, only spectral masking (i.e., $AM(0, \Delta F)$, see Eq. (3)) between atoms sharing the same length was exploited, as was done in [6]. In the second variant, PMP F 2, spectral masking across lengths was considered. More specifically, each atom with a duration Γ_l can mask atoms with durations $\Gamma_{k \neq l} \leq \Gamma_l$. Similarly, the mono- vs. across-length effect was investigated in the two other variants, PMP TF 1 and PMP TF 2, but using the full TF masking model in Eq. (3). The results are shown in Fig. 2, which displays the number of nonzero atoms as a function of N for MP (dotted line), PMP F 1 (gray dashed line), PMP F 2 (black dashed line), PMP TF 1 (gray solid line) and PMP TF 2 (black solid line). Table 2 lists percentages of atoms removed from the set of atoms identified by MP after 80 000 iterations for each variant of PMP. Including temporal masking allowed to remove 5–10% more atoms than using only spectral masking (PMP TF 2 *vs.* PMP F 2). Accordingly, the results in [14] showed that the amount of selected components can be reduced by about 5% if temporal masking is exploited. Interestingly, by allowing long atoms to mask shorter ones, the amount of selected atoms could be reduced by more than 10% (PMP TF 2 *vs.* PMP TF 1).

In Figure 3, the amplitude of the atoms selected by MP are plotted. The circles indicate the atoms removed using PMP TF 2. These results show that the TF masking model removes atoms that are identified in the first iterations of the algorithm.

The quality of the approximations was assessed using the PEMO-Q model [18]. The results are given in Tab. 3 for the sparse approximation given by MP and the result of the TF masking model applied to this representation. Applying the masking model degrades slightly the quality but, in any case, the "Objective Difference Grade" (ODG) values are all between 0 and -1, i.e., between *imperceptible* and *perceptible*


(a) maderna.wav

(b) svega.wav

Fig. 2. Number of nonzero atoms as a function of N for MP and the four PMP variants for (a) B. Maderna and (b) S. Vega.

but not annoying, according to [18].

Example soundfiles (results of MP with and without masking, as well as the residuals and the masked components) are accessible at
`http://www.kfs.oeaw.ac.at/ICASSP2014_PMP`.

5. CONCLUSIONS

In this paper, we demonstrated the combination of a sparse representation of sounds based on Gabor dictionaries and a TF masking model. Compared to previous work, we show that considering the masking between atoms of different sizes and at different times yields sparser representations.

Possible improvements to this work include the use of the ERBLET transform [4], a transformation adapted to perception, to generate the dictionaries, and a refined masking model to take the additivity of masking into account for multiple maskers. Finally, as the sparse representation and masking stages are independent, MP can be replaced by any other sparse approximation algorithm such as Iterative Hard Thresholding [19] or Basis Pursuit [20].

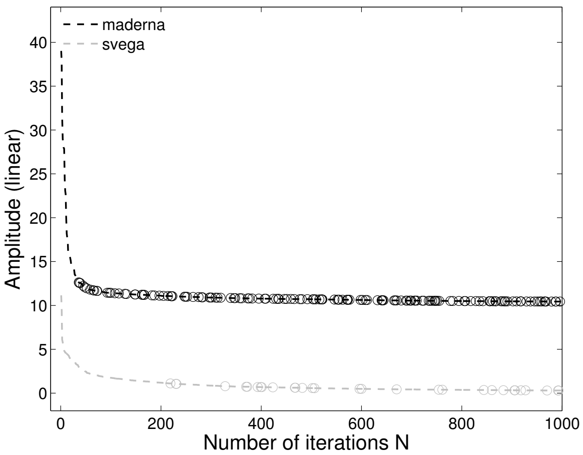


Fig. 3. Linear amplitudes of the atoms in MP as a function of N for maderna (black curve) and svega (gray curve). Circles indicate atoms removed using PMP TF 2. For clarity, the curve for maderna was shifted up by 10 and the results are shown only for the first 1000 iterations.

		PSM	ODG
svega	MP	1	0
	PMP	0.999	-0.1303
maderna	MP	0.9995	-0.1609
	PMP	0.9985	-0.2953

Table 3. Results of the PEMO-Q evaluation for MP and PMP with TF masking and N = 80 000.

6. REFERENCES

[1] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, “Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking,” *IEEE Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 34–49, 2010.

[2] O. Rioul and M. Vetterli, “Wavelets and signal processing,” *IEEE Signal Process. Mag.*, vol. 8, pp. 14–38, October 1991.

[3] M. D. Abolhassani and Y. Salimpour, “A human auditory tuning curves matched wavelet function,” in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2008)*, Vancouver, Canada, August 20–24 2008, pp. 2956–2959.

[4] T. Necciari, P. Balazs, N. Holighaus, and P. Søndergaard, “The ERBlet transform: An auditory-based time-frequency representation with perfect reconstruction,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, May 2013, IEEE, pp. 498–502.

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

[6] E. Ravelli, G. Richard, and L. Daudet, “Union of MDCT bases for audio coding,” *IEEE Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1361–1372, 2008.

[7] M.D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M.E. Davies, “Sparse representations in audio and music: From coding to source separation,” *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, June 2010.

[8] T. Necciari, *Auditory time-frequency masking: Psychoacoustical measures and application to the analysis-synthesis of sound signals*, Degree of Doctor of Acoustics, Aix-Marseille University, France, October 2010.

[9] T. S. Verma and T. H. Y. Meng, “Sinusoidal modeling using frame-based perceptually weighted matching pursuits,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP 1999)*, Phoenix, Arizona, USA, March 1999, IEEE, vol. 2, pp. 981–984.

[10] R. Heusdens, R. Vafin, and W. B. Kleijn, “Sinusoidal modeling using psychoacoustic-adaptive matching pursuits,” *IEEE Signal Process. Lett.*, vol. 9, no. 8, pp. 262–265, August 2002.

[11] H. Purnhagen, N. Meine, and B. Edler, “Sinusoidal coding using loudness-based component selection,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, Orlando, FL, USA, May 13–17 2002, IEEE, vol. 2, pp. 1817–1820.

[12] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jepsen, and S.H. Jensen, “A perceptual model for sinusoidal audio coding based on spectral integration,” *EURASIP J. Appl. Sig. Proc.*, vol. 9, pp. 1292–1304, 2005.

[13] P. Vera-Candeas, N. Ruiz-Reyes, and F. López-Ferreras, “Bark scale-based perceptual matching pursuit for improving sinusoidal audio modeling,” *Digit. Signal Process.*, vol. 19, no. 2, pp. 229–240, 2009.

[14] R. Vafin, S. V. Andersen, and W. B. Kleijn, “Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP 2000)*, Istanbul, Turkey, June 2000, vol. 2, pp. 901–904.

[15] R. Pichevar, H. Najaf-Zadeh, L. Thibault, and H. Lahdili, “Auditory-inspired sparse representation of audio signals,” *Speech Commun.*, vol. 53, no. 5, pp. 643–657, May–June 2011.

[16] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[17] S. Krstulovic and R. Gribonval, “MPTK: Matching Pursuit made tractable,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, Toulouse, France, May 2006, IEEE, vol. 3, pp. 496–499.

[18] R. Huber and B. Kollmeier, “PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception,” *IEEE Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1902–1911, November 2006.

[19] T. Blumensath and M. E. Davies, “Iterative thresholding for sparse approximations,” *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.

[20] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.