14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

AUDIO MASKING AND TIME-FREQUENCY EXPANSIONS

Bernard Mulgrew

School of Engineering Electronics
Institute for Digital Communications
The University of Edinburgh
Edinburgh, Scotland, UK
Email: B.Mulgrew@ed.ac.uk

ABSTRACT

An alternative mechanism for audio masking is postulated. This mechanism is derived as a solution to the classic problem of representing a signal as a linear combination of basis functions which are only approximately orthogonal and hence are prone to leakage. This mechanism involves augmenting each basis function or filter with an auxiliary filter. In this combined detection/estimation process the instantaneous amplitude output of the auxiliary filter sets the masking threshold for the basis filter. No interconnection between basis functions is required to compute this masking threshold. For a gammatone filter bank the auxiliary filter is formed from the cascade of the gammatone itself and a single zero notch filter. The single zero (in the z-plane) has the same frequency as the centre frequency as the gammatone filter and is at a radius dependent on its bandwidth.

1. INTRODUCTION

Perceptual audio masking phenomena are well characterized and have been very successfully applied to the coding of speech and music signals [1]. The field continues to advance with recent interest in perceptually based sinusoidal coding [2]. However the question that does not appear to have been addressed is what, in a signal processing context, is the purpose of these masking phenomena. From a coding perspective there is no need to answer the question. The objective is to mimic the operation of the ear in as accurate and as computationally efficient manner as possible. However if we are to exploit these masking phenomena in signal estimation tasks such a blind source separation [3] we need to understand the phenomena at a more fundamental theoretical level. Several computational auditory models are available (e.g. [4] and [5]) but they tend to include nonlinear elements that render any further analysis of their effect on the signal difficult or even intractable. In particular if a masking processes is used as a preprocessor to independent component analysis (ICA) techniques, a linearized interpretation of the masking process has clear advantages both with respect to successful operation of the ICA (which usually assumes a linear mixing model) and any analysis of the effects of masking on ICA.

It is clear from many studies that the primary function of the ear is to perform some form of time-frequency analysis. The latter is often formulated as a linear combination of vectors from a fixed dictionary. In the language of [6], a signal M-vector $\mathbf{y}$ is described as a linear combination of N vectors or atoms $\{\underline{\phi}_i\}_{i=1}^N$ and a white noise vector $\mathbf{n}$:

$$\mathbf{y} = \sum_{i=1}^{N} \alpha_i \underline{\phi}_i + \mathbf{n} \qquad (1)$$

Each atom $\underline{\phi}_i$ is associated with a particular co-ordinate in the time-frequency plane, i.e.: (t_i, ω_i) and the representation may be overcomplete in that $N > M$. The choice of a particular family of atoms defines the form of analysis e.g. short-term Fourier transform (SDFT), Gabor, Gabor wavelets etc. and these may be either frames or bases as appropriate.

For practical reasons, atoms are not placed at every possible point in the time-frequency plane but rather are placed according to some sampling or tiling strategy of which the simplest is a uniform rectangular grid. The spacing of atoms can be used to control the theoretical properties of the representation e.g. to ensure orthogonality of the vectors or to provide well conditioned frames. The orthogonality property is particularly attractive since it would lead to sparse representations if the signal was indeed a linear combination of atoms placed exactly at grid points. However, real signals are unlikely to be composed of atoms which conveniently fall onto grid points. In this case, the representation will not be sparse even if the atoms on the grid points form an orthogonal set. The most common manifestation of this problem is the leakage or spectral smearing associated with the DFT of a sinewave that does not have an integer number of cycles within the analysis window.

In this paper we consider approximately orthogonal complex-phasor-based expansions such as the DFT and a gammatone filter bank. We show that each atom (basis function or filter) can be used to derive an auxiliary filter whose output provides an approximate upper-bound on these leakage or co-channel interference terms and can thus be used to set a detection threshold at the output of the atom of interest. The output of this detector controls a gate with which to sample the response of the atom of interest to the signal. Essentially we demonstrate that masking phenomena are an observable characteristic of an approximately orthogonal expansion that has been designed to combine the functions of detection and estimation as in [7] but without the added computational expense of estimating signal statistics associated with [7].

In section II the generic method is outlined and illustrated with a simple application in spectral analysis. In section III the same technique is applied to a gammatone-filter-based time-frequency analysis system and the straightforward form of the auxilary filter is highlighted.

2. MASKING OF NON-ORTHOGONAL COMPONENTS

A variety of solutions to identifying the weights, $\{\alpha_i\}_{i=1}^N$, associated with the expansion of (1) have been suggested in the literature from matching pursuit [6] to orthogonalized match-

ing pursuit to Wiener filtering to projection filters [8]. The simplest solution is provided by the matched filter which is optimal when the basis vectors are orthogonal. The matched filter estimates, $\{\hat{\alpha}_i\}$, of the weights $\{\alpha_i\}$ are given by:

$$\hat{\alpha}_i \quad = \quad \mathbf{w}_i^H \mathbf{y}$$

where $\mathbf{w}_i = \underline{\phi}_i/(\underline{\phi}_i^H \underline{\phi}_i)$.

We seek a vector (or filter) $\underline{\psi}_i$ that is easily computed from $\underline{\phi}_i$ but which is orthogonal to it. By definition this vector will not respond to any component of $\underline{\phi}_i$ present in the observed signal $\mathbf{y}$. Formally $\underline{\psi}_i$ is defined as a solution to: $\underline{\psi}_i^H \underline{\phi}_i = 0$. We note that this is not a bi-orthogonality condition. To ease implementation and incorporate simplifying properties from the outset, we postulate the use of an $N \times N$ circulant shift matrix $\mathbf{C}$ with which to construct $\underline{\psi}_i$ from $\underline{\phi}_i$:

$$\underline{\psi}_i \quad \propto \quad \underline{\phi}_i - k_i \mathbf{C} \underline{\phi}_i$$

where the complex gain k_i term ensures that the orthogonality condition is satisfied:

$$k_i^* \quad = \quad \frac{\underline{\phi}_i^H \underline{\phi}_i}{(\mathbf{C}\underline{\phi}_i)^H \underline{\phi}_i} \tag{2}$$

Applying the observed signal $\mathbf{y}$ to this new filter and exploiting the orthogonality property we have:

$$\begin{aligned} z_i &= \underline{\psi}_i^H \mathbf{y} \\ &= \underline{\psi}_i^H \left\{ \sum_{j=1}^{N} \alpha_j \underline{\phi}_j + \mathbf{n} \right\} \\ &= \sum_{j=1, j\neq i}^{N} \alpha_j \{\underline{\psi}_i^H \underline{\phi}_j\} + \underline{\psi}_i^H \mathbf{n} \end{aligned}$$

noting the restriction on the summation. Ideally the response of this new filter would have the form:

$$z_i^o \quad = \quad \frac{1}{|\underline{\phi}_i|^2} \{ \sum_{j=1, j\neq i}^{N} \alpha_j \{\underline{\phi}_i^H \underline{\phi}_j\} + \underline{\phi}_i^H \mathbf{n} \}$$

and provide a measure of the co-channel interference induced by the signal itself and the non-orthogonality of the basis functions. However this would imply a contradiction since it would require $\underline{\phi}_i$ and $\underline{\psi}_i$ to be one and the same.

2.1 Stochastic Component

The filter output z_i is the sum of a deterministic component (or mean) $\sum_{j=1, j\neq i}^{N} \alpha_j \{\underline{\psi}_i^H \underline{\phi}_j\}$ and a zero-mean stochastic component $\underline{\psi}_i^H \mathbf{n}$. Consider the latter first. The variance of z_i is:

$$\mathrm{var}(z_i) \quad = \quad |\underline{\psi}_i|^2 \sigma_n^2 \propto \{|k_i|^2 - 1\} |\underline{\phi}_i|^2 \sigma_n^2$$

Similarly the variance (or noise component) of the matched filter output is:

$$\mathrm{var}(\hat{\alpha}_i) \quad = \quad |\mathbf{w}_i|^2 \sigma_n^2 = \frac{\sigma_n^2}{|\underline{\phi}_i|^2}$$

By equating these two variances, we define $\underline{\psi}_i$ exactly as

$$\underline{\psi}_i \quad = \quad \frac{\underline{\phi}_i - k_i \mathbf{C} \underline{\phi}_i}{\sqrt{\{|k_i|^2 - 1\}} |\underline{\phi}_i|^2}$$

and since $\mathrm{var}(z_i) = \mathrm{var}(\hat{\alpha}_i)$ we can use the output of the new filter to estimate the variance of the noise component present in the output of the matched filter. Thus we can set *a detection or masking threshold* for the presence of the desired component α_i in the matched filter output $\hat{\alpha}_i$. If the noise is Gaussian then this test would also have a constant false alarm rate (CFAR) property.

2.2 Deterministic Component

Returning to the deterministic component, we seek to satisfy the inequality:

$$|z_i| \quad \geq \quad |z_i^o|$$

as tightly as possible so we can set a threshold at the output of the matched filter. Invoking the magnitude inequality, this will be satisfied provided:

$$|\underline{\psi}_i^H \underline{\phi}_j| \quad \geq \quad \frac{|\underline{\phi}_i^H \underline{\phi}_j|}{|\underline{\phi}_i|^2} = |\mathbf{w}_i^H \underline{\phi}_j|, \forall j, j \neq i \tag{3}$$

and

$$|\underline{\psi}_i^H \mathbf{n}| \quad \geq \quad \frac{|\underline{\phi}_i^H \mathbf{n}|}{|\underline{\phi}_i|^2} = |\mathbf{w}_i^H \mathbf{n}|$$

The use of the magnitude inequality may appear to be over restrictive but it has the advantage of leading to conditions that are signal independent being only a property of the sets of vectors $\{\underline{\phi}_i\}$ and $\{\underline{\psi}_i\}$. The condition on the stochastic component will be met on average from the earlier arguments.

2.3 Algorithm Summary

The algorithm can be summarized in the following manner. The complex gain terms k_i and the scaling factor $\sqrt{\{|k_i|^2 - 1\}} |\underline{\phi}_i|^2$ are pre-computed at each i as they are not signal dependent. Then for each i do the following:

1. response to basis vector

$$r_i \quad = \quad \underline{\phi}_i^H \mathbf{y}$$

2. matched filter output

$$\hat{\alpha}_i \quad = \quad r_i / |\underline{\phi}_i|^2$$

3. auxiliary filter output

$$z_i \quad = \quad \frac{r_i - k_i^* \underline{\phi}_i^H (\mathbf{C}^H \mathbf{y})}{\sqrt{\{|k_i|^2 - 1\}} |\underline{\phi}_i|^2}$$

4. masked output

$$\beta_i \quad = \quad \hat{\alpha}_i \mathrm{I}(|\hat{\alpha}_i| - |z_i|)$$

where the indicator function $\mathrm{I}(x) = 1$ when $x > 0$ and is zero otherwise.

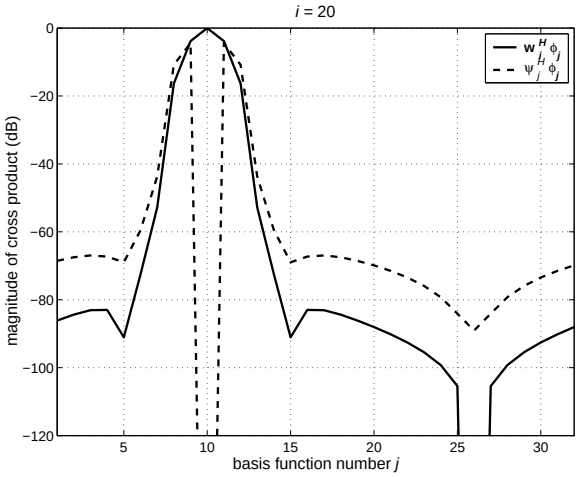

Figure 1: Response of 10th matched filter $\mathbf{w}_{10}$ (solid line) and masking vector $\underline{\psi}_{10}$ (dotted line)to other basis vectors $\underline{\phi}_j$

2.4 Examples

Figure 1 illustrates how well these conditions can be met for a windowed 32-point DFT. The basis vectors are constructed from a 32-point DFT matrix whose columns have been weighted with a Hamming window. The solid curve shows the response of the matched filter $\mathbf{w}_{10}$ to the other basis vectors $\underline{\phi}_j$. It is clear that this basis set is only approximately orthogonal. The dotted curve shows the response of the auxiliary or masking vector $\underline{\psi}_{10}$ to the same basis vectors. The expected null response at $i = 10$ is evident. For this example the response of the auxiliary filter is greater than the response of the matched filter to all other basis functions. Thus the magnitude of the output of the auxiliary filter forms an upper bound on the non-orthogonal interference present in the matched filter output and provides a suitable "pass/reject" threshold for the matched filter output.

It is straightforward to configure the algorithm of II.C for spectral analysis. Figure 2 illustrates the results for the analysis of 4 sinewaves in white noise. The basis vectors are constructed from the first 16 rows a of 32×32 discrete Fourier transform matrix $\mathbf{F_{32}}$. The columns of the resultant rectangular matrix are weighted with a 16-point Hamming window to form the 32 candidate basis vectors. The matched filter output shown is identical to that which we would be obtained using a 16-point Hanning window, zero-padding and a 32-point DFT. The masking threshold is the output of the auxiliary filters at each frequency of interest. The two closely-spaced sine waves at A are not resolvable by this matched filter bank but they are detected as the matched filter output is greater than the masking threshold. Of greater interest is the component at C which is detected despite there being no evidence of a peak in the matched filter response. Also of note are the lack of detections in the region to the left of A and to the right of B. The white noise components in these regions have been masked by the presence of the sinewaves at A, B and C. The resultant "masked output" response is considerably sparser than the "matched filter" outputs.

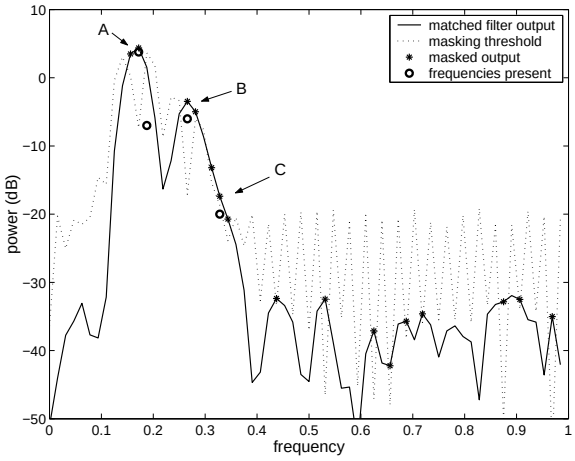

Figure 2: Spectral Analysis: four sine waves in white noise at -30dB

3. TIME-FREQUENCY ANALYSIS

Recently [9], it was shown that a gammatone filter bank is all that is needed to compute the audio masking threshold. If we use a complex gammatone filter bank to define the basis vector $\underline{\phi}_i$, its lth element would be:

$$\phi_i(l) \quad = \quad e^{j\omega_i(l-l_i)}\gamma_i(l-l_i) \tag{4}$$

where $\gamma_i(l)$ is a gamma distribution of the form:

$$\gamma_i(l) \quad = \quad l^{\gamma-1}e^{-B_i l}$$

and B_i is the bandwidth which is a constant fraction of the centre frequency ω_i. Thus the ith basis vector is associated with a co-ordinate (l_i, ω_i) in the time-frequency plane. The temporal extent of the gamma distribution is proportional to the time constant $1/B_i$ and is usually much smaller than the length of the data record and hence the basis vectors. Thus, apart from basis vectors associated with time indices at the beginning and end of the data record, the circulant shift operation is identical to a simple shift or time delay operation. Thus $\mathbf{C}\underline{\phi}_i$ can be replaced by $\mathbf{S}\underline{\phi}_i$ where ever it appears - $\mathbf{S}$ being a simple non-circulant shift matrix.

As is usual practice all the matched filters associated with a particular centre frequency ω_i can be realized by applying the time domain signal $y(l)$ to a single gammatone filter with impulse response:$g_i(l) = e^{j\omega l}\gamma_i(l)$ and transfer function $G_i(z)$ and recording the output signal. The transfer function of the associated auxiliary filter is given by: $G_i(z)(1-k^*z^{-1})$ which is a cascade of the gammatone filter and a notch filter with a single zero at $z = k^*$. Combining (2) and (4) we have:

$$k_i^* \quad = \quad \frac{e^{j\omega_i}\sum_l \gamma_i^2(l-l_i)}{\sum_l \gamma_i(l-l_i)\gamma_i(l-l_i-1)}$$

Thus, since the gamma distribution is a positive real function, the complex gain term has the form $k_i^* = |k_i|e^{j\omega_i}$. The notch filter has a notch in the centre of the band at frequency ω_i. The magnitude $|k_i|$ controls the proximity of the zero to the unit circle in the z-plane and hence the depth of the notch. One element of the resultant filter bank is illustrated in Figure 3.

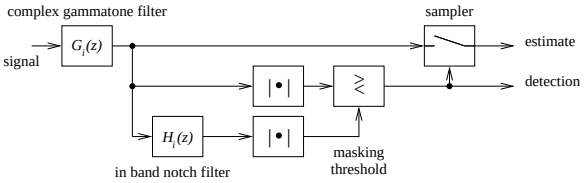

Figure 3: Model for audio masking

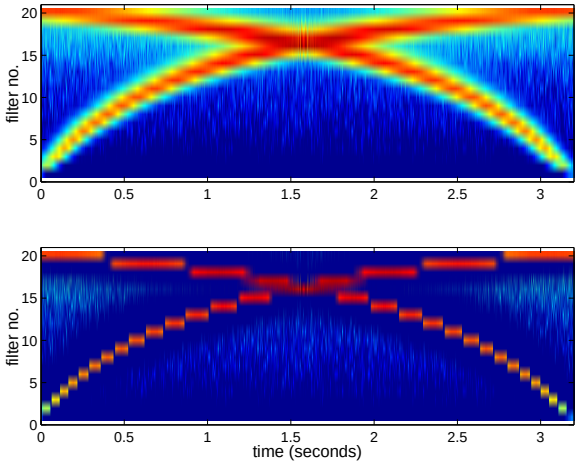

Figure 4: Response to linear frequency ramps (chirps): gammatone filterbank (upper); gammatone filter bank with masking model of Section II.C (lower)

Initial results have indicated that this linearized model exhibits many of the masking phenomena described in the psychoacoustics literature [10] such as *temporal masking, simultaneous masking and the precedence effect*. In addition it also exhibits the *peak-sampling* characteristics that are necessary to preserve frequency resolution [4]. Some of the properties of this linearized model are illustrated in Figure 4 where the sum of two linear frequency chirps are applied to a gammatone filter bank that employs 20 filters to cover a frequency range up to 6kHz. One sine wave ramps up in frequency from 0 Hz and the other ramps down from 6kHz. White noise is added at a level of 10 dB below the signal. The upper plots shows the response of a gammatone filter bank alone and the lower lower curve illustrates the effect of adding the masking model of Section II.C. The most noticeable effects are: (i) clearer definition of the signal in the time/frequency plane; (ii) removal of the noise in a time/frequency band around the signal. Figure 5 illustrates a similar analysis of a short single-note jazz guitar phrase with a cymbal crash at the end. Here again the effect of the masking is to provide a more localized and sparser decomposition of the signal in the time/frequency plane.

4. CONCLUSIONS

In this paper we have postulated an alternative linear mechanism for audio masking. This mechanism involves augmenting each basis function or filter with an auxiliary filter. In this combined detection/estimation process the instantaneous amplitude output of the auxiliary filter sets the masking threshold for the basis filter. No interconnection between basis functions is required to compute this masking threshold. For a gammatone filter bank the auxiliary filter is formed from the cascade of the gammatone itself and a single zero notch filter. The single zero (in the z-plane) has the same frequency as the centre frequency as the gammatone filter and is at a radius dependent on its bandwidth. Future work will involve an attempt to test and calibrate this method with measured data.

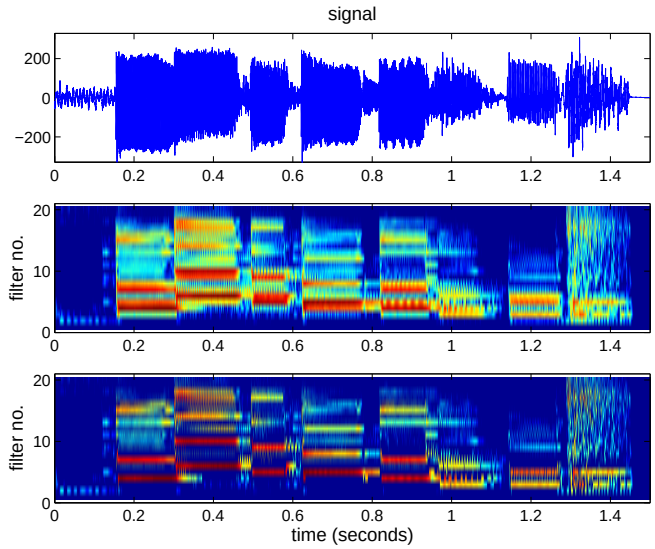

Figure 5: Jazz guitar: signal (upper); gammatone filterbank (middle); gammatone filter bank with masking model of Section II.C (lower)

Acknowledgment

The author acknowledges the support of the Royal Academy of Engineering for this work.

REFERENCES

[1] T. Painter and A. Spinias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–513, Apr 2000.

[2] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Applied Signal Processing*, vol. 9, no. 1, pp. 1334–1349, Jun 2005.

[3] R. Guddeti and B. Mulgrew, "Perceptually motivated blind source separation of convolutive mixtures," in *IEEE International Conference on Acoustics Speech and Signal Processing*, March 18-23, 2005, pp. 273 – 276.

[4] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Information Theory*, vol. 38, no. 2, pp. 824–839, Mar 1992.

[5] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *Acoustical Society of America Journal*, vol. 99, pp. 3615–3622, June 1996.

[6] S. Mallat and Z. Zhang, "Matching pursuits with time-

frequency dictionaries," *IEEE Trans Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

[7] E. Aboutanios and B. Mulgrew, "A STAP algorithm for radar target detection in heterogeneous environments," in *Proceedings IEEE/SP 13th Workshop on Statistical Signal Processing*, Bordeaux, France, July 2005.

[8] F. Hlawatsch and W. Kozek, "Time-frequency projection filters and time-frequency signal expansions," *IEEE Trans Signal Processing*, vol. 42, no. 12, pp. 3321 – 3334, Dec 1994.

[9] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new pychoacoustal masking model for audio coding applications," in *IEEE International Conference on Acoustics Speech and Signal Processing*, vol. 2, Orlando, 2002, pp. 1805–1808.

[10] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, Massachusetts: The MIT Press, 1997.