# A perfectly invertible and perceptually motivated time-frequency transform for audio representation, analysis and synthesis

Thibaud Necciari

joint work with Peter Balazs, Nicki Holighaus, and Peter L. Søndergaard

Acoustics Research Institute, Vienna

ESI12 Workshop, December 3–7, 2012, Vienna

*Time-frequency methods for the applied sciences*

ÖAW
Österreichische Akademie der Wissenschaften

ARI

# Context: Analysis-Synthesis of Sound Signals.

- Audio processing techniques like sound design, audio coding, or speech & music processing require tools to:
  - analyse (represent, extract relevant features...)
  - process
  - re-synthesize sounds
- Standard tools = time-frequency (TF) transforms

# Context: Analysis-Synthesis of Sound Signals.

- Audio processing techniques like sound design, audio coding, or speech & music processing require tools to:
  - analyse (represent, extract relevant features...)
  - process
  - re-synthesize sounds
- Standard tools = time-frequency (TF) transforms
- Humans listeners = main receivers of speech & music signals
- Intuition: Account for auditory perception in signal analysis = TF transform that approximates the auditory TF resolution

# Context: Analysis-Synthesis of Sound Signals.

- Audio processing techniques like sound design, audio coding, or speech & music processing require tools to:
  - analyse (represent, extract relevant features...)
  - process
  - re-synthesize sounds
- Standard tools = time-frequency (TF) transforms
- Humans listeners = main receivers of speech & music signals
- Intuition: Account for auditory perception in signal analysis = TF transform that approximates the auditory TF resolution

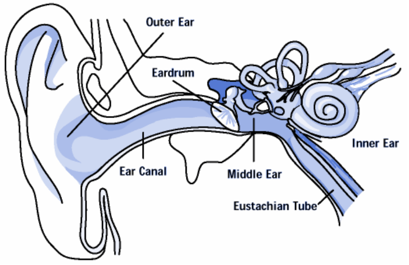
Ideal transform properties:

- Invertibility
- Computational efficiency
- Adaptable redundancy

# The Auditory Resolution.

## 1. Frequency domain: The Auditory Filters.

= Ability to resolve sinusoidal components in complex sounds.



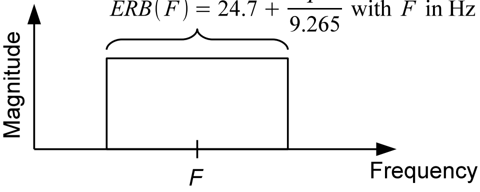

Peripheral filtering ≡ bank of bandpass filters = auditory filters

# The Auditory Resolution.

## 1. Frequency domain: The ERB Scale [Moore & Glasberg, 1983].

**ERB** = **E**_quivalent_ **R**_ectangular_ **B**_andwidth_

$$ERB(F) = 24.7 + \frac{F}{9.265} \text{ with } F \text{ in Hz}$$



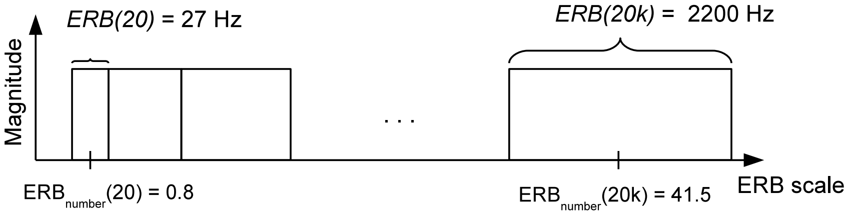

$$ERB_{\text{number}}(F) = 9.265 \ln\left(1 + \frac{F}{228.8455}\right)$$

- distribution of filters:
  - $\approx$ linear at low frequencies ($F < 500$ Hz)
  - logarithmic at high frequencies ($F > 2$ kHz)
- $ERB(F) \approx$ constant-Q only at high frequencies

# The Auditory Resolution.

## 1. Frequency domain: The ERB Scale [Moore & Glasberg, 1983].

**ERB** = **E***quivalent* **R***ectangular* **B***andwidth*




- distribution of filters:
  - $\approx$ linear at low frequencies ($F < 500$ Hz)
  - logarithmic at high frequencies ($F > 2$ kHz)
- $ERB(F) \approx$ constant-Q only at high frequencies

# The Auditory Resolution.

## 2. Temporal domain.

$=$ Ability to detect rapid changes in sounds over time.

- Frequency partition into filters
  $\rightsquigarrow$ Time windows with frequency-dependent lengths
- **Windows' length = temporal resolution**
- Windows' shape is well approximated by Gaussians with [van Schijndel *et al.*, 1999]:
  - bandwidth $\approx ERB(F)$
  - temporal width $\approx$ 4 periods of $F$, *e.g.*,
    4 ms **@** 1 kHz, 1 ms **@** 4 kHz

# Perceptually Motivated TF Representations.

## State-of-the-Art.

- Auditory models [Plack *et al.*, 2002; Meddis *et al.*, 2012]
  ✓ Useful to gain insights into auditory processing
  X Not invertible, large parameter sets, computationally demanding

# Perceptually Motivated TF Representations.

## State-of-the-Art.

- Auditory models [Plack *et al.*, 2002; Meddis *et al.*, 2012]
  ✓ Useful to gain insights into auditory processing
  X Not invertible, large parameter sets, computationally demanding
- Constant-Q transforms [Philippe *et al.*, 1999; Velasco *et al.*, 2011]
  ✓ Near-perfect or perfect reconstruction
  X Approximate the auditory resolution only at high frequencies, large concentration of filters at low frequencies

# Perceptually Motivated TF Representations.

## State-of-the-Art.

- Auditory models [Plack *et al.*, 2002; Meddis *et al.*, 2012]
  ✓ Useful to gain insights into auditory processing
  X Not invertible, large parameter sets, computationally demanding
- Constant-Q transforms [Philippe *et al.*, 1999; Velasco *et al.*, 2011]
  ✓ Near-perfect or perfect reconstruction
  X Approximate the auditory resolution only at high frequencies, large concentration of filters at low frequencies
- Auditory filterbanks (gammatone, frequency warping) [Smith & Abel, 1999; Hohmann, 2002; Irino & Patterson, 2006]
  ✓ Approximate well the auditory resolution
  X No or only approximate reconstruction

# Goal of the Study.

Achieve a linear TF transform featuring:

- **perceptually motivated** TF resolution
- **perfect reconstruction**
- adaptable resolution and redundancy, *i.e.*,
  - adjustable frequency channels (number of sub-bands)
  - adjustable down-sampling factors

# Goal of the Study.

Achieve a linear TF transform featuring:

- **perceptually motivated** TF resolution
- **perfect reconstruction**
- adaptable resolution and redundancy, *i.e.*,
  - adjustable frequency channels (number of sub-bands)
  - adjustable down-sampling factors

## Proposed approach:

- Use frame theory and the non-stationary Gabor transform (NSGT) [cf. presentation by Peter Balazs] to develop a NSGT matched to the ERB scale
- "ERBlet transform" = *non-uniform auditory filterbank*

# Outline.

1. Underlying concept: The non-stationary Gabor transform
2. ERBlet implementation
3. Simulations
4. Conclusions & perspectives

# Outline.


1. Underlying concept: The non-stationary Gabor transform
2. ERBlet implementation
3. Simulations
4. Conclusions & perspectives

# The Non-Stationary Gabor Transform (NSGT).

## Formulation as a Non-Uniform Filterbank [Balazs *et al.*, 2011].

NSG system with resolution evolving across frequency:

$$\mathcal{G}(\mathbf{g}, \mathbf{D}) := (g_{n,k}[l]) = (g_k\left[l - nD_k\right])$$

where

- $l \in \mathbb{Z} =$ time variable
- $n, k \in \mathbb{Z} =$ time and frequency index, resp.
- $\mathbf{g} := (g_k) =$ frequency-dependent filters
- $\mathbf{D} := (D_k) =$ frequency-dependent down-sampling factors

# The NSGT *continued*.

## Frame Theory.

The sequence $(g_{n,k})$ is called a *frame* if the constants $A, B \in \mathbb{R}^{+\star}$ exist that satisfy

$$A\|f\|^2 \leq \sum_{n,k} | \langle f, g_{n,k} \rangle |^2 \leq B\|f\|^2$$

for any signal $f \in \mathbb{R}$.

# The NSGT *continued*.

## Analysis and Synthesis (1/2).

### NSG analysis:

Analysis through the frame operator $\mathbf{S}$ is given by

$$\mathbf{S}f = \sum_{n,k} \langle f, g_{n,k} \rangle \, g_{n,k}.$$

If $\mathbf{S}$ is invertible, then perfect reconstruction is achieved using the *canonical dual frame*

$$\widetilde{\mathcal{G}(\mathbf{g}, \mathbf{D})} = (\tilde{g}_{n,k}) = \mathbf{S}^{-1}(g_{n,k}).$$

### NSG synthesis:

$$f = \mathbf{S}^{-1}\mathbf{S}f = \sum_{n,k} \langle f, g_{n,k} \rangle \, \tilde{g}_{n,k}.$$

# The NSGT *continued*.

## Analysis and Synthesis (2/2).

Conditions for "painless" reconstruction:

- $\hat{g}_k = \mathcal{F}(g_k)$ has a bandpass characteristic
- $\mathrm{supp}(\hat{g}_k) = \mathcal{I}_k$ (in samples)
- $D_k$ satisfies $\lceil \frac{L}{D_k} \rceil \geq \mathcal{I}_k$, $L =$ signal length

It follows that the operator $\hat{\mathbf{S}} := \mathcal{F}\,\mathbf{S}\,\mathcal{F}^{-1}$ is diagonal and easily invertible.

# Outline.


1 Underlying concept: The non-stationary Gabor transform

2 ERBlet implementation
- Analysis & dual windows: ERBlets
- Algorithms

3 Simulations

4 Conclusions & perspectives

# ERBlet Design.

## Analysis Windows.

ERBlet transform $= \mathcal{G}(\mathbf{g}, \mathbf{D})$ with $g_k, k = 0 \ldots K$, defined in the frequency domain by

$$\hat{g}_k[m] = \frac{1}{\sqrt{\Gamma_k}}\, e^{-\pi\left[\frac{m-\nu_k}{\Gamma_k}\right]^2}$$

# ERBlet Design.

## Analysis Windows.

ERBlet transform $= \mathcal{G}(\mathbf{g}, \mathbf{D})$ with $g_k, k = 0 \dots K$, defined in the frequency domain by

$$\hat{g}_k[m] = \frac{1}{\sqrt{\Gamma_k}}\, e^{-\pi\left[\frac{m-\nu_k}{\Gamma_k}\right]^2}$$

To obtain filters equidistantly spaced on the ERB scale:

- Let $F_{\text{min}}$, $F_{\text{max}}$ = min, max analysis frequencies, resp.
- Then $E_0 = ERB_{\text{number}}(F_{\text{min}})$ and $E_K = ERB_{\text{number}}(F_{\text{max}})$
- Distribute $K+1$ filters from $E_0$ to $E_K$ with $V$ filters/ERB
- $\rightsquigarrow E_k = E_0 + k/V$ and $K = V\,(E_K - E_0)$.
- $\nu_k = ERB^{-1}_{\text{number}}(E_k)$
- $\Gamma_k = ERB(\nu_k)$

# ERBlet Design.

## Analysis Windows.

ERBlet transform $= \mathcal{G}(\mathbf{g}, \mathbf{D})$ with $g_k, k = 0 \dots K$, defined in the frequency domain by

$$\hat{g}_k[m] = \frac{1}{\sqrt{\Gamma_k}} \, e^{-\pi \left[ \frac{m-\nu_k}{\Gamma_k} \right]^2}$$

To obtain filters equidistantly spaced on the ERB scale:

- Let $F_{\text{min}}$, $F_{\text{max}} =$ min, max analysis frequencies, resp.
- Then $E_0 = ERB_{\text{number}}(F_{\text{min}})$ and $E_K = ERB_{\text{number}}(F_{\text{max}})$
- Distribute $K+1$ filters from $E_0$ to $E_K$ with $V$ filters/ERB
- $\rightsquigarrow E_k = E_0 + k/V$ and $K = V\,(E_K - E_0)$.
- $\nu_k = ERB^{-1}_{\text{number}}(E_k)$
- $\Gamma_k = ERB(\nu_k)$

Windows truncated so that $\text{supp}(\hat{g}_k) = \mathcal{I}_k = \lceil 4\,\Gamma_k \rceil$.

# ERBlet Design.

## Dual Windows.

“Painless case” condition:

*i.e.*, choose $D_k$ such that the number of time positions

$$N_k = \left\lceil \frac{L}{D_k} \right\rceil \geq \lceil 4\,\Gamma_k \rceil.$$

$\rightsquigarrow$ $\hat{\mathbf{S}}$ is diagonal and easily invertible and $\widetilde{g_{n,k}} = \mathcal{F}^{-1}\hat{\mathbf{S}}^{-1}\widehat{g_{n,k}}$.

### “Painless case” condition:

*i.e.*, choose $D_k$ such that the number of time positions

$$N_k = \left\lceil \frac{L}{D_k} \right\rceil \geq \lceil 4\,\Gamma_k \rceil.$$

$\rightsquigarrow$ $\hat{\mathbf{S}}$ is diagonal and easily invertible and $\widetilde{g_{n,k}} = \mathcal{F}^{-1}\hat{\mathbf{S}}^{-1}\widehat{g_{n,k}}$.

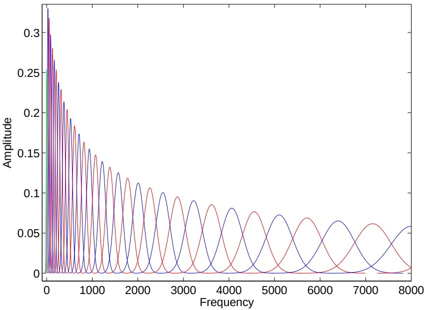
### Otherwise,

if $N_k < \lceil 4\,\Gamma_k \rceil$ then $\hat{\mathbf{S}}$ is not diagonal. We use an iterative method to approximate $\hat{\mathbf{S}}^{-1}$.

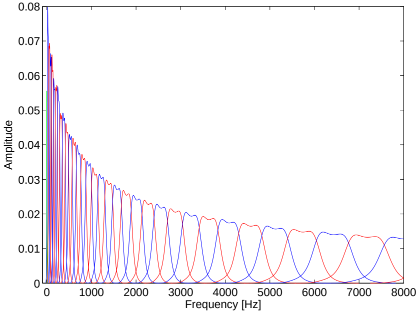
# ERBlet Design.

## Windows Example: Spectral Domain.

Analysis windows



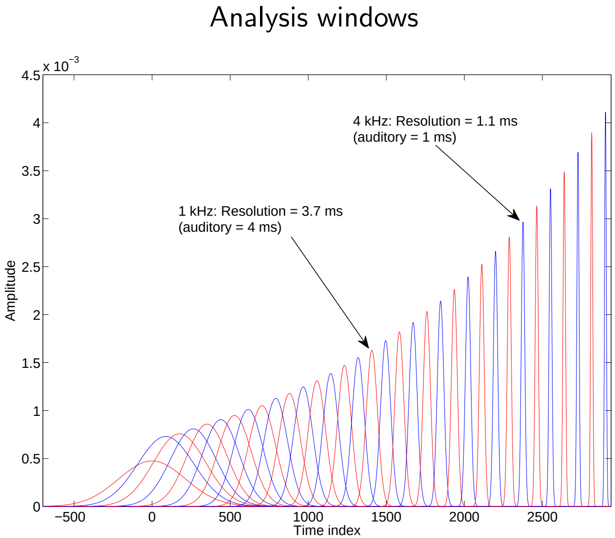

Dual windows




- $F_{\text{min}} = 0$, $F_{\text{max}} = 8$ kHz (Nyquist frequency)
- $V = 1$ filter/ERB ($\equiv$ auditory filterbank)
- $K = 34$ channels

# ERBlet Design.

## Windows Example: Time Domain.



- NSGT with resolution evolving over time available in LTFAT [Søndergaard *et al.*, 2012]: functions `nsdgt.m` and `insdgt.m`
- Applying these algorithms to $\hat{f}$ allows to construct NSGT with resolution evolving over frequency
- ERBlet is determined by 2 parameters: $V$ and $D_k$
  - enable adaptable resolution & redundancy
  - $red = \sum_{k=0}^{K} D_k^{-1}$
  - `erblet.m` and `ierblet.m` soon available in LTFAT

We use a *conjugate gradients* algorithm (CG) to solve the system

$$\hat{\mathbf{S}} f = \sum_{n,k} c_{n,k}\, \widehat{g_{n,k}}.$$

- CG works with Hermitian and positive-definite matrices

We use a *conjugate gradients* algorithm (CG) to solve the system

$$\hat{\mathbf{S}} f = \sum_{n,k} c_{n,k} \, \widehat{g_{n,k}}.$$

- CG works with Hermitian and positive-definite matrices
  - $\hat{\mathbf{S}}$ is Hermitian provided $(g_{n,k})$ is a frame

We use a *conjugate gradients* algorithm (CG) to solve the system

$$\hat{\mathbf{S}} f = \sum_{n,k} c_{n,k} \, \widehat{g_{n,k}}.$$

- CG works with Hermitian and positive-definite matrices
  - $\hat{\mathbf{S}}$ is Hermitian provided $(g_{n,k})$ is a frame
- $\hat{\mathbf{S}} f \equiv \mathbf{S}^{-1} \mathbf{S} f$

We use a *conjugate gradients* algorithm (CG) to solve the system

$$\hat{\mathbf{S}} f = \sum_{n,k} c_{n,k} \widehat{g_{n,k}}.$$

- CG works with Hermitian and positive-definite matrices
  - $\hat{\mathbf{S}}$ is Hermitian provided $(g_{n,k})$ is a frame
- $\hat{\mathbf{S}} f \equiv \mathbf{S}^{-1}\mathbf{S} f$
  - we can use `nsdgt.m` followed by `insdgt.m` instead of $\hat{\mathbf{S}}$

We use a *conjugate gradients* algorithm (CG) to solve the system

$$\hat{\mathbf{S}} f = \sum_{n,k} c_{n,k}\, \widehat{g_{n,k}}.$$

- CG works with Hermitian and positive-definite matrices
  - $\hat{\mathbf{S}}$ is Hermitian provided $(g_{n,k})$ is a frame
- $\hat{\mathbf{S}} f \equiv \mathbf{S}^{-1}\mathbf{S} f$
  - we can use `nsdgt.m` followed by `insdgt.m` instead of $\hat{\mathbf{S}}$
- Since $\widehat{g_k}$ decay fast, $\hat{\mathbf{S}}$ is diagonal dominant and

$$\mathbf{P}(\hat{\mathbf{S}})^{-1}_{m,l} = \begin{cases} \left(\sum N_k |\hat{g}_k|^2\right)^{-1}[m], & \text{if } m = l \\ 0, & \text{else} \end{cases}$$

is a good preconditioner [Balazs *et al.*, 2006].

# Outline.


1. Underlying concept: The non-stationary Gabor transform
2. ERBlet implementation
3. Simulations
   - Iterative reconstruction
   - Signal representation
4. Conclusions & perspectives

# Simulations.

## Overview.

### 2 Experiments:

- Exp. 1: Test the convergence of CG for various redundancies
- Exp. 2: Compare the ERBlet to a standard DGT and a linear gammatone filterbank [Hohmann, 2002]

### Setup:

- Audio material: 2 musical excerpts (5–10 sec) in mono format, sampled at 44.1 kHz, 16 bits/sample
- $F_{\text{min}} = 0$, $F_{\text{max}} = 22.05$ kHz

# Simulations.

## Experiment 1: Convergence of CG.

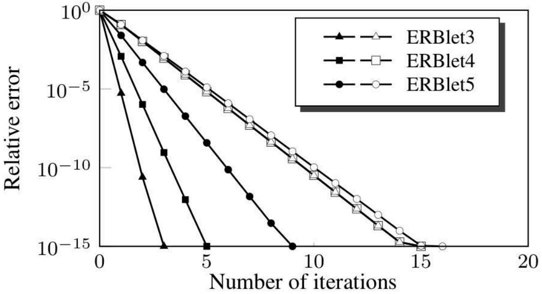



Figure (CG)

| $\mathcal{G}(\mathbf{g}, \mathbf{D})$ | $V$ | $K$ | $N_k$ | $red$ | $B/A$ |
|---|---|---|---|---|---|
| ERBlet3 | 1 | 43 | $\lceil \frac{32\,\Gamma_k}{9} \rceil$ | 3.53 | 1.44 |
| ERBlet4 | 1 | 43 | $\lceil \frac{8\,\Gamma_k}{3} \rceil$ | 2.64 | 1.44 |
| ERBlet5 | 1 | 43 | $\lceil 2\,\Gamma_k \rceil$ | 1.98 | 1.52 |
| ERBlet6 | 1 | 43 | $\lceil \frac{4\,\Gamma_k}{3} \rceil$ | 1.32 | 2.56 |
| ERBlet7 | 1 | 43 | $\lceil \frac{12\,\Gamma_k}{11} \rceil$ | 1.08 | 5.88 |

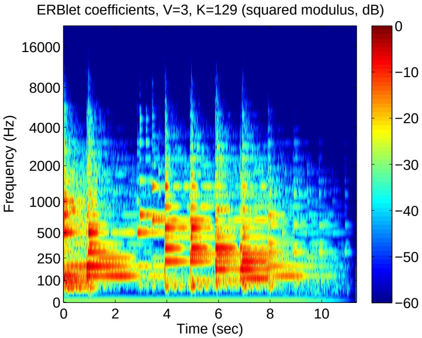
Painless case (reference)

| $V$ | $K$ | $N_k$ | $red$ | $B/A$ |
|---|---|---|---|---|
| 1 | 43 | $\lceil 4\,\Gamma_k \rceil$ | 4.00 | 1.44 |
| 3 | 129 | $\lceil 4\,\Gamma_k \rceil$ | 12.00 | 1.07 |

# Simulations.

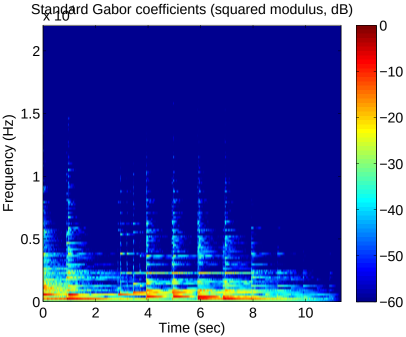
## Experiment 2: ERBlet *vs.* DGT.

ERBlet (dB)




- $B/A = 1.07$ ("painless")
- $red = 12$
- Rel. error $< 10^{-15}$

DGT (dB)

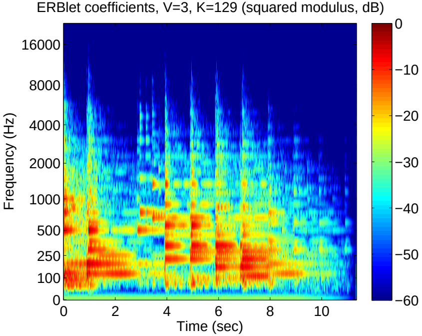



- $B/A = 1.0$
- $red = 11.73$
- Rel. error $< 10^{-15}$

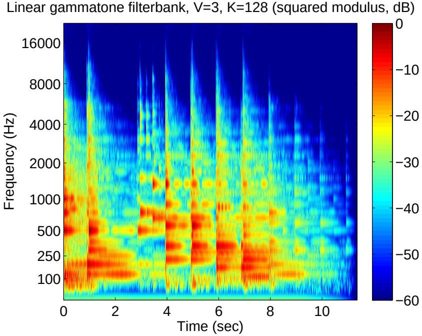
# Simulations.

## Experiment 2: ERBlet *vs.* GFB.

ERBlet (dB)




GFB (dB)



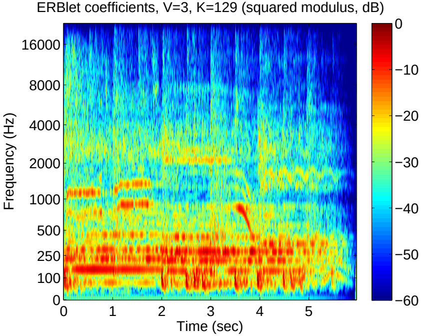

- $B/A = 1.07$ ("painless")
- $red = 12$
- Rel. error $< 10^{-15}$

- $red = 128$
- Rel. error $= 1.3$

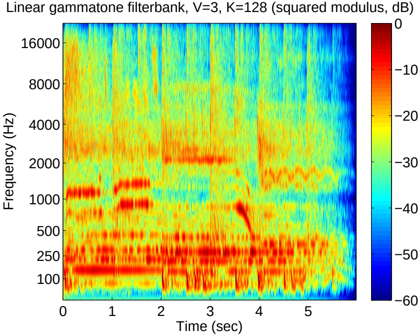
# Simulations.

## Experiment 2: ERBlet *vs.* GFB.

ERBlet (dB)




GFB (dB)




- $B/A = 1.07$ ("painless")
- $red = 12$
- Rel. error $< 10^{-15}$

- $red = 128$
- Rel. error $= 1.4$

# Outline.

1 Underlying concept: The non-stationary Gabor transform

2 ERBlet implementation

3 Simulations

4 Conclusions & perspectives

# Conclusions.

- ERBlet = Linear and perfectly invertible TF transform adapted to human auditory perception
- Adaptable resolution and redundancy
  - Perfect reconstruction achieved using iterative method even using 1 filter/ERB and $red = 1.08$
- Compatible with linear gammatone representation
  - Approximates well the auditory TF resolution
- Soon available in the Matlab/Octave toolbox LTFAT
- New analysis/synthesis tool for audio processing

# Perspectives.

- Include basilar membrane compression and compare with nonlinear gammatone filterbanks [Irino & Patterson, 2006]
- Use windows with Gaussian shapes on the ERB scale, *i.e.*, use a warping function to map linear frequency to ERB scale
- Introduce perceptual sparsity in the transform using recent data on auditory TF masking [Balazs *et al.*, 2010; Necciari, 2010]

Thank you for your attention!

thibaud@kfs.oeaw.ac.at