The ERBlet transform, auditory time-frequency masking and perceptual sparsity

Thibaud Necciari[1]

joint work with P. Balazs[1], B. Laback[1], P. Soendergaard[1,3],
R. Kronland-Martinet[2], S. Meunier[2], S. Savel[2], and S. Ystad[2]

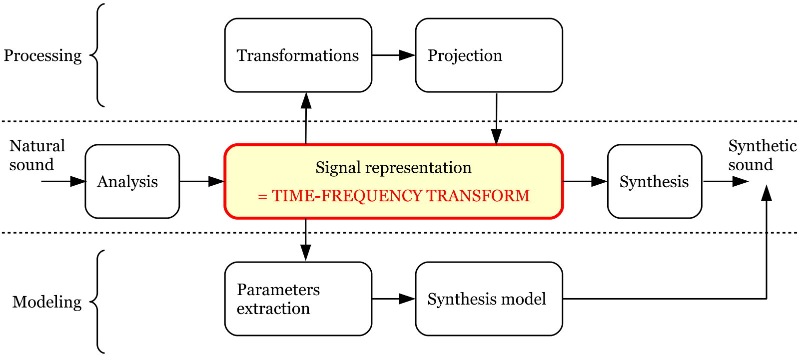
[1]Acoustics Research Institute, Vienna, Austria

[2]Laboratoire de Mécanique et d'Acoustique, Marseille, France

[3]Technical University of Denmark

2nd SPLab Workshop, October 24–26, 2012, Brno

Context: Analysis-Synthesis of Sound Signals.



Idea: Integrate aspects of human auditory perception in the signal representation

Goal of the Study.

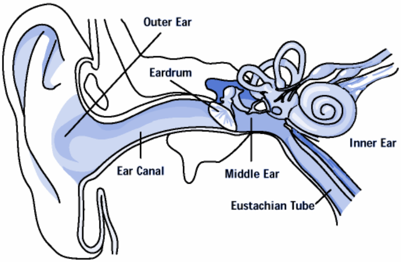
Achieve a **perceptually-motivated** and **invertible** TF transform based on:

1. Properties of TF transforms:
    - Linear
    - Allow perfect reconstruction
    - Adapted to non-stationary signals
2. Results on human auditory perception (psychoacoustics)

Some Aspects of Human Auditory Perception.

1. Spectral Resolution: The Auditory Filters.

= Ability to resolve sinusoidal components in complex sounds.

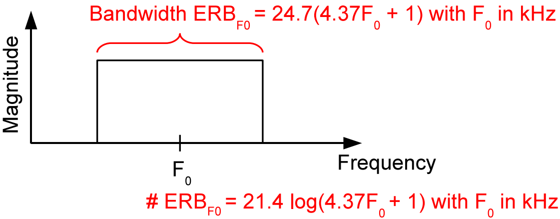


Peripheral filtering ≡ bank of bandpass filters = auditory filters

Some Aspects of Human Auditory Perception.

1. Spectral Resolution: The ERB Scale [Moore & Glasberg, 1983].

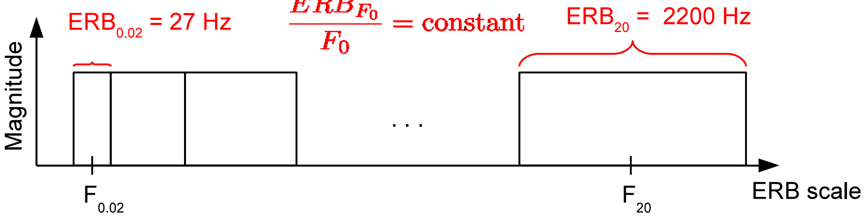
Each auditory filter is characterized by its
ERB = **E***quivalent* **R***ectangular* **B***andwidth*



Each auditory filter is characterized by its

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

$$\frac{ERB_{F_0}}{F_0} = \text{constant}$$



Some Aspects of Human Auditory Perception.

2. Temporal Resolution.

= Ability to detect rapid changes in sounds over time.

- Time axis partitioned into time windows
  (analog to spectral resolution)
- **Windows length = temporal resolution**
- Windows length **= frequency dependent**
  $\approx$ "internal" TF analysis [van Schijndel *et al.*, 1999]
- Windows length $\approx$ **4 periods of center frequency**
  e.g., 4 ms **@** 1 kHz and 1 ms **@** 4 kHz

= Increase in the detection threshold of a sound ("target") in the presence of another sound ("masker").

= Increase in the detection threshold of a sound ("target") in the presence of another sound ("masker").

Measurement

Amount of masking (dB) =

$$\underbrace{\text{masked threshold}}_{\text{Detection threshold of target in presence of the masker}} - \underbrace{\text{absolute threshold}}_{\text{Detection threshold of target in quiet}}$$

= Increase in the detection threshold of a sound ("target") in the presence of another sound ("masker").

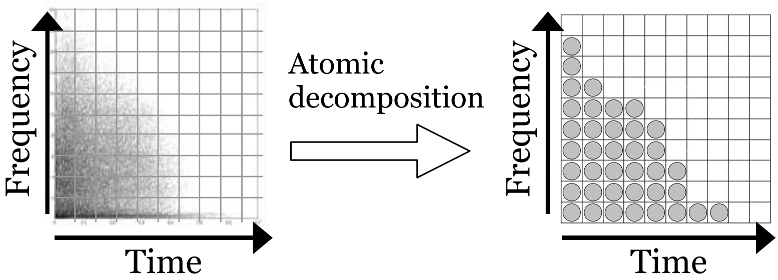
Main parameters:

- Time
- Frequency
- Stimulus duration
- Stimulus level
- Frequency region of the audible spectrum [20 Hz ... 20 kHz]

Some Aspects of Human Auditory Perception.

3. Auditory Masking: Consequence in Signal Representation.

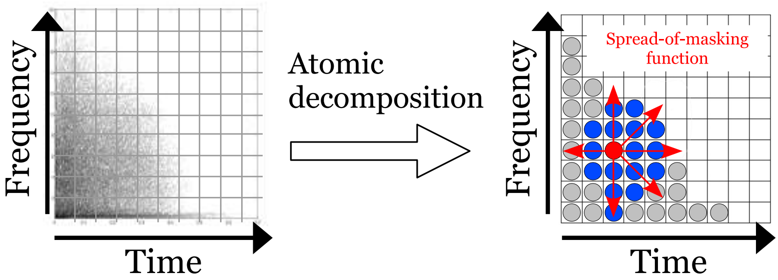
$$s(t) = \underbrace{C_g}_{\text{normalization}} \iint_{\mathbb{R}} STFT(\tau,\omega)\, \underbrace{g_{\tau,\omega}(t)}_{\text{TF atom}}\, \mathrm{d}\tau \mathrm{d}\omega$$



Some Aspects of Human Auditory Perception.

3. Auditory Masking: Consequence in Signal Representation.

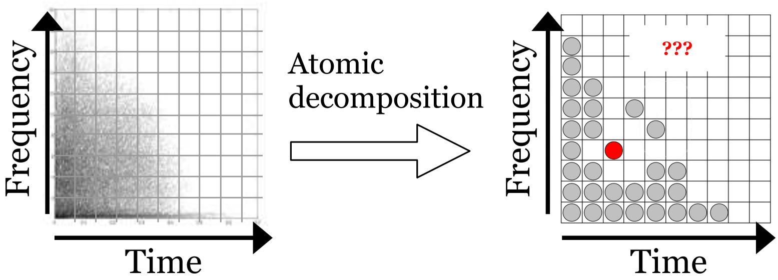
$$s(t) = \underbrace{C_g}_{\text{normalization}} \iint_{\mathbb{R}} STFT(\tau,\omega)\ \underbrace{g_{\tau,\omega}(t)}_{\text{TF atom}}\ \mathrm{d}\tau\mathrm{d}\omega$$



Some Aspects of Human Auditory Perception.

3. Auditory Masking: Consequence in Signal Representation.

$$s(t) = \underbrace{C_g}_{\text{normalization}} \iint_{\mathbb{R}} STFT(\tau,\omega)\, \underbrace{g_{\tau,\omega}(t)}_{\text{TF atom}}\, \mathrm{d}\tau \mathrm{d}\omega$$



- Can we represent only audible atoms?
- If so, **which atoms can be removed?**

Proposed Approach.

To obtain a perceptually-motivated and invertible TF transform:

Proposed Approach.

To obtain a perceptually-motivated and invertible TF transform:

1. Adapt the transform parameters to mimic the auditory TF resolution
   $\hookrightarrow$ **A variable-resolution transform is required!**

Proposed Approach.

To obtain a perceptually-motivated and invertible TF transform:

1. Adapt the transform parameters to mimic the auditory TF resolution
   $\hookrightarrow$ **A variable-resolution transform is required!**
2. Use a psychoacoustic model of TF masking to represent *only the audible* components (perceptual sparsity concept).

Outline.

1. Perceptually-based TF transform: The ERBlet
2. Perceptual sparsity concept: Investigating auditory TF masking
3. Discussion: Combination of ERBlet & perceptual sparsity?

Outline.

1. Perceptually-based TF transform: The ERBlet
   - Concept
   - Implementation
   - Example
2. Perceptual sparsity concept: Investigating auditory TF masking
3. Discussion: Combination of ERBlet & perceptual sparsity?

The *ERBlet Transform*.

Concept.

The non-stationary Gabor transform (NSGT) [Balazs *et al.*, 2011]

- Allows resolution to freely evolve over T and/or F
- We can adapt both
  - The shape of $g(t)$ either in T or F
  - The redundancy
- Perfect reconstruction is achieved if the frame inequality is fulfilled

Idea

Develop a perceptually-motivated NSGT:

- Use NSGT with resolution evolving over frequency to mimic the ERB scale $\hookrightarrow$ The *ERBlet transform*.

- NSGT with resolution evolving over time available in LTFAT [Soendergaard, 2010]: function `nsdgt.m`
- Applying `nsdgt` on the Fourier transform of $s(t) \mapsto \hat{s}(\nu)$ allows to construct NSGT with resolution evolving over frequency ($=$ constant-Q NSGT in [Velasco *et al.*, 2011] but with $\neq$ functions)

ERBlet Implementation.

1. Analysis Functions.

- NSGT with resolution evolving over time available in LTFAT [Soendergaard, 2010]: function `nsdgt.m`
- Applying `nsdgt` on the Fourier transform of $s(t) \mapsto \hat{s}(\nu)$ allows to construct NSGT with resolution evolving over frequency (= constant-Q NSGT in [Velasco *et al.*, 2011] but with $\neq$ functions)

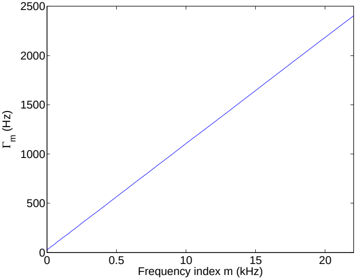
Analysis functions (Gaussian windows):

$$\hat{h}_m(\nu) = \frac{1}{\sqrt{\Gamma_m}} e^{-\pi \left(\frac{\nu}{\Gamma_m} \right)^2}$$

where

- m = frequency index
- $\Gamma_m = ERB_m$ (in Hz)

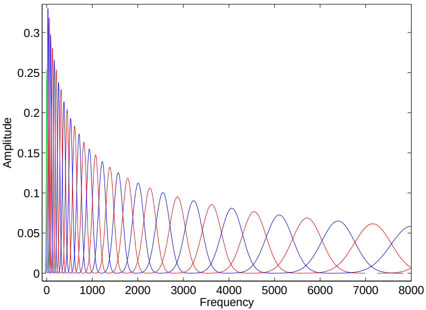
$$\Gamma_m = \mathrm{f}(m)$$



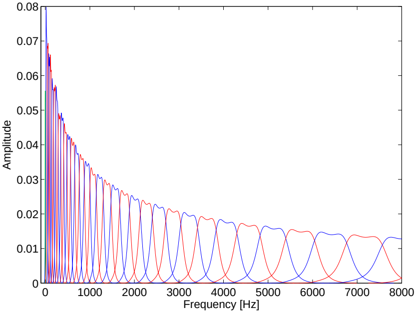
ERBlet Implementation.

2. Spectral Resolution.

Analysis windows



Dual windows

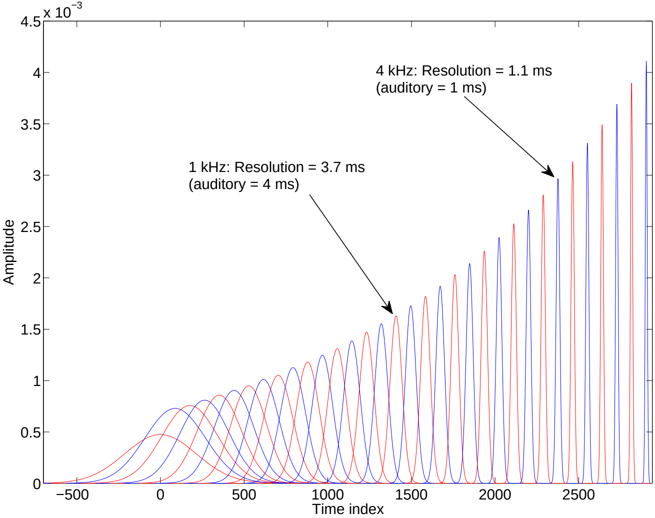


- 1 window/ERB ($\equiv$ auditory filterbank); 34 channels **@** 8 kHz, 49 channels **@** 22 kHz

ERBlet Implementation.

3. Temporal Resolution.

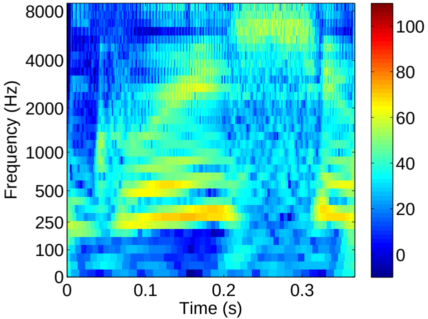
Analysis windows, time



ERBlet Example.

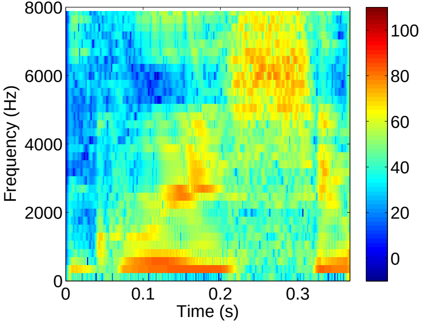
LTFAT Speech Test Signal “greasy”.

ERBlet (dB SPL)



- Frame bounds ratio $= 1.5$
- Redundancy ≈ 4
- Reconstruction error $< 10^{-16}$

Standard Gabor (dB SPL)



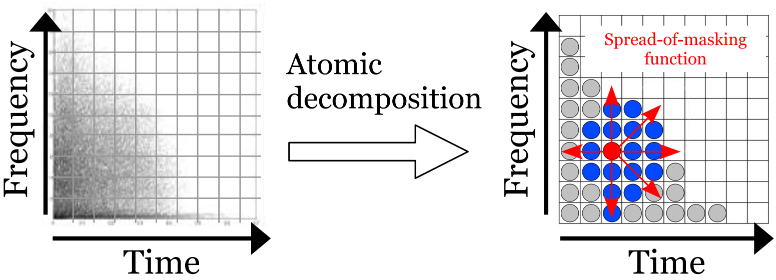
- Frame bounds ratio $= 1$
- Redundancy ≈ 4.6
- Reconstruction error $< 10^{-16}$

Outline.

1. Perceptually-based TF transform: The ERBlet
2. Perceptual sparsity concept: Investigating auditory TF masking
   - Problematic
   - Experimental methods
   - Results
3. Discussion: Combination of ERBlet & perceptual sparsity?

Auditory TF Masking: Problematic.

Which atoms can be removed from the signal representation?



A representation of TF masking for **short <u>and</u> narrowband** signals is required.

Auditory TF Masking: Problematic.

Current masking data are not suitable for prediction of masking between TF atoms

Auditory TF Masking: Problematic.

Current masking data are not suitable for prediction of masking between TF atoms

- Psychoacoustical studies **mostly** focused on **T OR F**

Auditory TF Masking: Problematic.

Current masking data are not suitable for prediction of masking between TF atoms

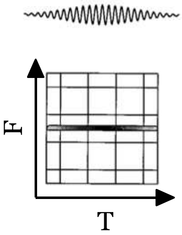
- Psychoacoustical studies **mostly** focused on **T OR F**
- **Very few** studies measured **TF masking**
  [Fastl, 1979; Kidd & Feth, 1981; Soderquist *et al.*, 1981; Moore *et al.*, 2002]

Auditory TF Masking: Problematic.

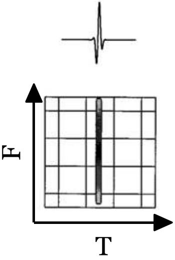
Current masking data are not suitable for prediction of masking between TF atoms

- Psychoacoustical studies **mostly** focused on **T OR F**
- **Very few** studies measured **TF masking** [Fastl, 1979; Kidd & Feth, 1981; Soderquist *et al.*, 1981; Moore *et al.*, 2002]
- These studies used **long-duration** maskers: not compatible with atomic decomposition

Sinusoid



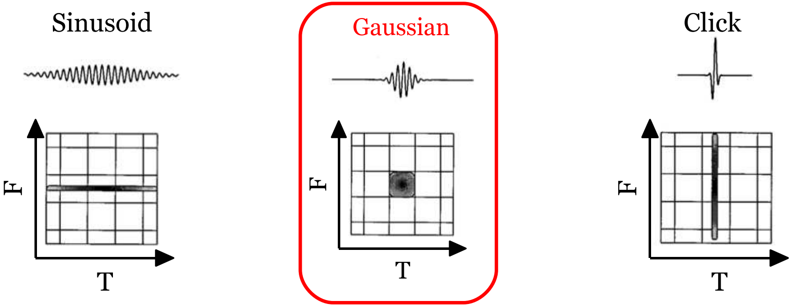
Click



Auditory TF Masking: Problematic.

Current masking data are not suitable for prediction of masking between TF atoms

- Psychoacoustical studies **mostly** focused on **T OR F**
- **Very few** studies measured **TF masking** [Fastl, 1979; Kidd & Feth, 1981; Soderquist *et al.*, 1981; Moore *et al.*, 2002]
- These studies used **long-duration** maskers: not compatible with atomic decomposition



Formula

$$s(t) = A\sqrt{\Gamma}\,\sin\left(2\pi f_0 t + \tfrac{\pi}{4}\right)\, e^{-\pi(\Gamma t)^2}$$

- $f_0 =$ carrier frequency
- $\frac{\pi}{4}$ phase shift: signal energy $=$ independent of f_0
- $\Gamma =$ shape factor of the Gaussian window

Formula

$$s(t) \;=\; A\,\sqrt{\Gamma}\,\sin\left(2\pi f_0 t + \tfrac{\pi}{4}\right)\, e^{-\pi(\Gamma t)^2}$$

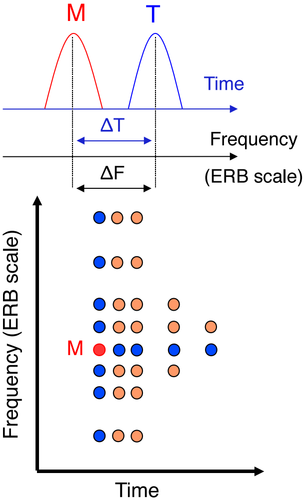
- $f_0 =$ carrier frequency
- $\frac{\pi}{4}$ phase shift: signal energy $=$ independent of f_0
- $\Gamma =$ shape factor of the Gaussian window

Spectro-temporal characteristics

- $ERB \Leftrightarrow \Gamma =$ 600 Hz [van Schijndel et al., 1999]
- $ERD \Leftrightarrow \Gamma^{-1} =$ 1.7 ms
- 0-amplitude duration $=$ 9.6 ms

Experimental Methods.

2. Conditions: Stimulus Parameters & Listeners.



- $F_M = 4$ kHz, $L_M = 81$–84 dB SPL
- $\Delta F = 0, \pm 1, \pm 2, \pm 4,$ or $+6$ ERBs
- $\Delta T = 0,\ 5,\ 10,\ 20,$ or 30 ms
- 30 crossed conditions
- 4 normal-hearing listeners

Experimental Methods.

3. Psychoacoustic Procedure for Thresholds Estimation.

- 3-interval forced-choice adaptive procedure
- 1 trial = 3 intervals:
  - Masker alone in 2 intervals
  - Masker + Target in 1 interval, chosen randomly
  - Task: *"Which interval contained the target?"*

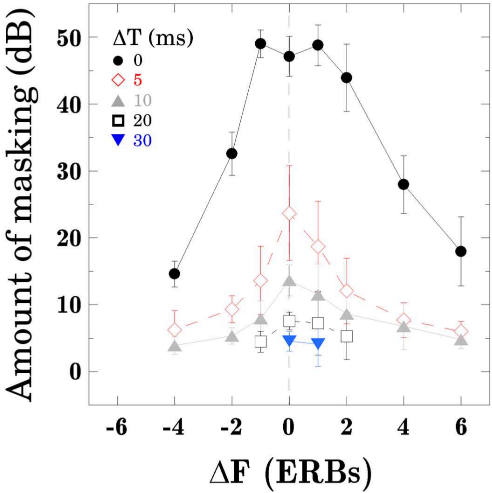
Experimental Methods.

3. Psychoacoustic Procedure for Thresholds Estimation.

- 3-interval forced-choice adaptive procedure
- 1 trial = 3 intervals:
  - Masker alone in 2 intervals
  - Masker + Target in 1 interval, chosen randomly
  - Task: *"Which interval contained the target?"*
- Masker level (L_M) was fixed
- Target level varied adaptively (3$\searrow$ - 1$\nearrow$ rule; 79.4% correct)
- Stimuli monaurally presented to the right ear

Mean Results.

Parameter $= \Delta T$.



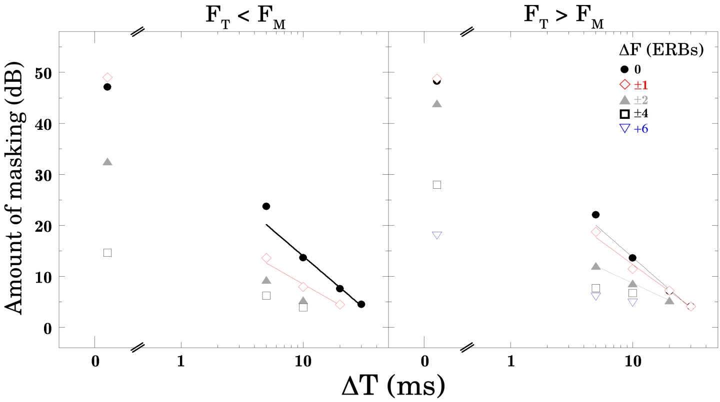
Patterns broaden when $\Delta T \nearrow$

ΔT	Q_{3dB}
0	12
5	3
10	2

[Fastl, 1979;
Kidd & Feth, 1981]

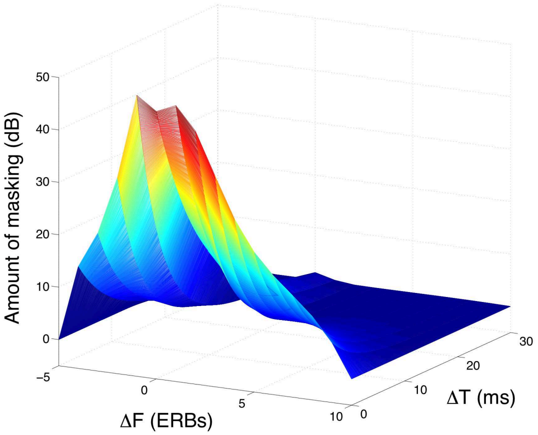
Mean Results.

Parameter $= \Delta F$.



Mean Results Extrapolated.

TF Masking Pattern for One Gaussian TF Atom.



Outline.

Computation of wavelet filters (frequency domain)

$$\hat{g}_a(\omega) = \sqrt{a}\hat{g}(a\omega)$$

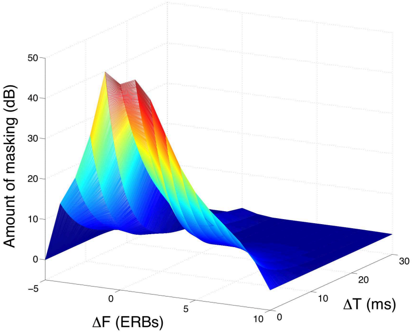
with “mother wavelet” (compatibility with experiments)

$$\hat{g}(\omega) = \frac{1}{2j\sqrt{\Gamma}} e^{-\pi\left(\frac{\omega-\omega_0}{\Gamma}\right)^2}$$

- $a > 1 =$ *scale factor* (compression only)
- $\Gamma = \alpha f_0 = \alpha \frac{\omega_0}{2\pi}$
- $\alpha = 0.15$
- $f_0 =$ frequency of mother wavelet ($f_0 =$ 16.5 kHz)
- Analysis in [30 Hz . . . 20 kHz]

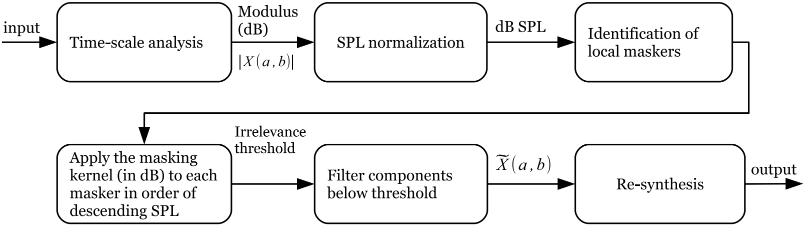
Use the measured TF masking pattern as a *masking kernel*

$$\mathcal{M}(\Delta T, \Delta F)$$



Previous Implementation with Wavelets.

3. Implementation of the Masking Kernel.



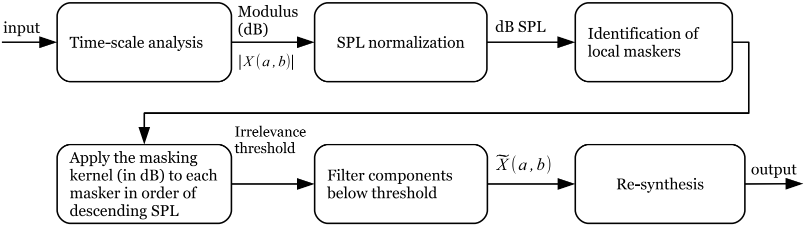
1. Identification of local maskers

$$\Omega_M = \{|X(a,b)| \geq Tq(a,\cdot) + 60\} \quad \text{(dB SPL)}$$

where $Tq(a)$ = threshold in quiet function [Terhardt, 1979]

Previous Implementation with Wavelets.

3. Implementation of the Masking Kernel.



2. Apply $\mathcal{M}(a,b)$ to each masker

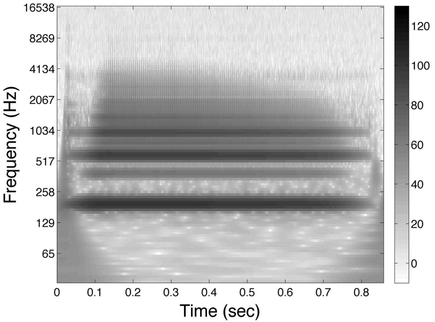
$$\widetilde{X}_g(a,b) = \begin{cases} X_g(a,b) & \text{if } |X_g(a,b)| \geq Tq(a,\cdot) + \mathcal{M}(a,b) \\ 0 & \text{otherwise} \end{cases}$$

until Ω_M is empty (iterate in descending SPL).

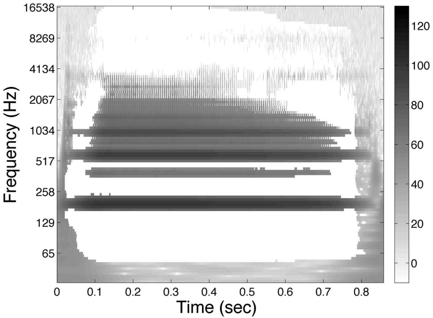
Previous Implementation with Wavelets.

4. Result (Test with Clarinet Note A3).

$$|X_g(a,b)|$$



$$|\widetilde{X}_g(a,b)|$$



50% components removed *but audible problems* at reconstruction due to removal of TF components.

Current limitations

- Reproducing kernel $\rightsquigarrow$ Tricky to remove atoms
  - ✓ Re-encode inaudible atoms like in audio codecs (mp3)‽

Extension to ERBlet.

Future Works.

Current limitations

- Reproducing kernel $\rightsquigarrow$ Tricky to remove atoms
  - ✓ Re-encode inaudible atoms like in audio codecs (mp3)‽
- Highly redundant representation $\rightsquigarrow$ **masking overestimation** and **high computational cost**
  - ✓ Change representation‽ $\Rightarrow$ ERBlet!

Extension to ERBlet.

Future Works.

Current limitations

- Reproducing kernel $\rightsquigarrow$ Tricky to remove atoms
  - ✓ Re-encode inaudible atoms like in audio codecs (mp3)‽
- Highly redundant representation $\rightsquigarrow$ **masking overestimation** and **high computational cost**
  - ✓ Change representation‽ $\Rightarrow$ ERBlet!
- Masking kernel for one atom
  - ✓ Use an analytic TF masking model‽
  - ✓ Incorporate level effects (✓ data collected)
  - ✓ Additivity of TF masking (✓ data collected)

Conclusions.

Conclusions.

- *ERBlet*: Linear and invertible TF transform adapted to human auditory perception $\leadsto$ New analysis/synthesis tool for the audio processing community

Conclusions.

- *ERBlet*: Linear and invertible TF transform adapted to human auditory perception $\rightsquigarrow$ New analysis/synthesis tool for the audio processing community
- *New psychoacoustic data* on auditory TF masking for **one** and **multiple atoms** $\rightsquigarrow$ Crucial for the development of an efficient TF masking model

Conclusions.

- *ERBlet*: Linear and invertible TF transform adapted to human auditory perception $\rightsquigarrow$ New analysis/synthesis tool for the audio processing community
- *New psychoacoustic data* on auditory TF masking for **one** and **multiple atoms** $\rightsquigarrow$ Crucial for the development of an efficient TF masking model

Next steps

1. Design an analytic TF masking model
2. Investigate the perceptual sparsity criterion: Combine Step 1. and the ERBlet
3. Calibrate & validate the new transform using perceptual listening tests

Thank you for your attention.

`thibaud@kfs.oeaw.ac.at`

Further reading:

P. Balazs *et al.*
Theory, implementation and applications of nonstationary Gabor frames.
J. Comput. Appl. Math. 236(6):1481, 2011.

T. Necciari *et al.*
Perceptual optimization of audio representations based on time-frequency masking data for maximally-compact stimuli.
AES 45th conference, Helsinki, 2012.

Acknowledgments: Work partly funded by Égide, the ANR, and WWTF.