Structured sparsity

Sevag Hanssian

MUMT 622, Winter 2021

March 25, 2021

2021-03-24

Structured sparsity

Structured sparsity

Sevag Hanssian

MUMT 622, Winter 2021

March 25, 2021

Structured sparsity

- Main paper: Kowalski and Torrésani[1]
- Same year, same authors: Kowalski and Torrésani[2]
- Same techniques: Siedenburg and Doerfler[3]

[1] Matthieu Kowalski and Bruno Torrésani. "Sparsity and persistence: Mixed norms provide simple signal models with dependent coefficients". In: *Signal Image and Video Processing* 3 (Sept. 2009). DOI: 10.1007/s11760-008-0076-1.

[2] Matthieu Kowalski and Bruno Torrésani. "Structured Sparsity: from Mixed Norms to Structured Shrinkage". In: *SPARS09-Signal Processing with Adaptive Sparse Structured Representations* 53 (Apr. 2009).

[3] Kai Siedenburg and Monika Doerfler. "Structured sparsity for audio signals". In: *Proc. of the 14th International Conference on Digital Audio Effects (DAFx-11), Paris, France*. Jan. 2011.

Structured sparsity
- Main paper: Kowalski and Torrésani[1]
- Same year, same authors: Kowalski and Torrésani[2]
- Same techniques: Siedenburg and Doerfler[3]

Sparsity

Sparsity seems to be a particularly efficient guiding principle in view of a number of tasks such as signal compression, denoising, image deblurring, blind source separation, ... The guiding principle may be summarized as follows: for most signal classes, it is possible to find a basis or a dictionary of elementary building blocks (or atoms) with respect to which all (or most) signals in the class may be expanded, so that when the expansion is truncated in a suitable way, high precision approximations are obtained even when very few terms are retained.[4]

[4] Kowalski and Torrésani, 2009.

Sparsity

Sparsity seems to be a particularly efficient guiding principle in view of a number of tasks such as signal compression, denoising, image deblurring, blind source separation, ... The guiding principle may be summarized as follows: for most signal classes, it is possible to find a basis or a dictionary of elementary building blocks (or atoms) with respect to which all (or most) signals in the class may be expanded, so that when the expansion is truncated in a suitable way, high precision approximations are obtained even when very few terms are retained.[4]

[4] Kowalski and Torrésani, 2009.

LASSO shrinkage

Tibshirani[5] proposed a new technique for linear regression: LASSO (least absolute shrinkage and selection operator). It "minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients."[6]

Figure: Lasso is an acronym and also an analogy for cattle ranchers[7]

[5] Robert Tibshirani. "Regression shrinkage selection via the LASSO". In: *Journal of the Royal Statistical Society Series B* 73 (June 2011), pp. 273–282. DOI: 10.2307/41262671.

[6] `https://statweb.stanford.edu/~tibs/lasso.html`

[7] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN: 1498712169.

LASSO shrinkage

Tibshirani[5] proposed a new technique for linear regression: LASSO (least absolute shrinkage and selection operator). It "minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients."[6]

Figure: Lasso is an acronym and also an analogy for cattle ranchers[7]

[5] Robert Tibshirani. "Regression shrinkage selection via the LASSO". In: *Journal of the Royal Statistical Society Series B* 73 (June 2011), pp. 273–282. DOI: 10.2307/41262671.

[6] https://statweb.stanford.edu/~tibs/lasso.html

[7] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN: 1498712169.

Similar ideas – basis pursuit

> *Basis pursuit (BP) is a principle for decomposing a signal into an "optimal" superposition of dictionary elements, where optimal means having the smallest l1 norm of coefficients among all such decompositions*[8]

Least Squares Optimization with L1 Regularization:[9] Estimating Least Squares parameters subject to an L1 penalty was presented and popularized independently under the names Least Absolute Selection and Shrinkage Operator (LASSO) in Tibshirani[10] and Basis Pursuit Denoising in Chen, Donoho, and Saunders[11]

---

[8] Mallat and Zhang, 1993.

[9] Mark Schmidt. *Least Squares Optimization with L1-Norm Regularization*. 2005. URL: https://www.cs.ubc.ca/~schmidtm/Documents/2005_Notes_Lasso.pdf.

[10] Tibshirani, "Regression shrinkage selection via the LASSO".

[11] S.S. Chen, D. Donoho, and Michael Saunders. "Atomic decomposition by basis pursuit[J]". In: *Siam Review* 43 (Jan. 2001), pp. 33–61.

Similar ideas – basis pursuit

Basis pursuit (BP) is a principle for decomposing a signal into an "optimal" superposition of dictionary elements, where optimal means having the smallest l1 norm of coefficients among all such decompositions[8]

Least Squares Optimization with L1 Regularization:[9] Estimating Least Squares parameters subject to an L1 penalty was presented and popularized independently under the names Least Absolute Selection and Shrinkage Operator (LASSO) in Tibshirani[10] and Basis Pursuit Denoising in Chen, Donoho, and Saunders[11]

[8] Mallat and Zhang, 1993.
[9] Mark Schmidt. *Least Squares Optimization with L1-Norm Regularization*. 2005. URL: https://www.cs.ubc.ca/~schmidtm/Documents/2005_Notes_Lasso.pdf.
[10] Tibshirani, "Regression shrinkage selection via the LASSO".
[11] S.S. Chen, D. Donoho, and Michael Saunders. "Atomic decomposition by basis pursuit[J]". In: *Siam Review* 43 (Jan. 2001), pp. 33–61.

L1 and L2 norm

Problem to solve:

- Given $\mathbf{y} \in \mathbb{R}^T$ be a noisy representation of a signal $\mathbf{s} \in \mathbb{R}^T$
- Let $\mathcal{D}$ denote a fixed dictionary for $\mathbb{R}^T$, and denote by $A \in \mathbb{R}^{T \times N}$ the matrix whose columns are the vectors from dictionary $\mathcal{D}$
- We assume $\mathbf{s}$ has a sparse expansion in $\mathcal{D}$
- We want to estimate $\mathbf{s}$ from $\mathbf{y}$
- LASSO approach: $\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1$
- $A\widehat{\mathbf{x}}$ is the estimate of $\mathbf{y}$, λ is a fixed parameter
- L1 norm is the $|| \cdot ||_1$ expression
- Generally, Lp norm is $|| \cdot ||_p^p$, so L2 norm is $|| \cdot ||_2^2$
- These are loss functions for the approximation. Can apply a mixed norm in two dimensions $\mathcal{P}_{p,q}$, e.g., time and frequency

L1 and L2 norm

Problem to solve:

- Given $\mathbf{y} \in \mathbb{R}^T$ be a noisy representation of a signal $\mathbf{s} \in \mathbb{R}^T$
- Let $\mathcal{D}$ denote a fixed dictionary for $\mathbb{R}^T$, and denote by $A \in \mathbb{R}^{T \times N}$ the matrix whose columns are the vectors from dictionary $\mathcal{D}$
- We assume $\mathbf{s}$ has a sparse expansion in $\mathcal{D}$
- We want to estimate $\mathbf{s}$ from $\mathbf{y}$
- LASSO approach: $\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{R}^N}{\operatorname{argmin}} \|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1$
- $A\widehat{\mathbf{x}}$ is the estimate of $\mathbf{y}$, λ is a fixed parameter
- L1 norm is the $|| \cdot ||_1$ expression
- Generally, Lp norm is $|| \cdot ||_p^p$, so L2 norm is $|| \cdot ||_2^2$
- These are loss functions for the approximation. Can apply a mixed norm in two dimensions $\mathcal{P}_{p,q}$, e.g., time and frequency

Mixed norm in time and frequency

In the context of time-frequency dictionaries, a natural step beyond classical sparsity approaches is the introduction of sparsity criteria which take into account the two-dimensionality of the time-frequency representations used. Mixed norms on the coefficient arrays make it possible to enforce sparsity in one domain and diversity and persistence in the other domain.[12]

Sparsity of coefficients may be enforced by ℓ^1-regression, also known as the *Lasso* [2]. Given a noisy observation $y = s + e$ in $\mathbb{C}^L$ it finds

$$\hat{c} = \arg\min_{c \in \mathbb{C}^P} \frac{1}{2}\|y - \Phi c\|_2^2 + \lambda \Psi(c) \tag{2}$$

with penalty term $\Psi(\cdot) = \|\cdot\|_1$ and $\lambda > 0$. Since the sequence $c_{k,j}$ is ordered along two dimensions for Gabor frames, the ℓ^1-prior Ψ in (2) may be replaced by a two-dimensional mixed norm $\ell^{p,q}$ which acts differently on groups (indexed by g in the sequel, may be either time or frequency) and their members (indexed by m):

$$\Psi(c) = \|c\|_{p,q} = \left(\sum_g \left(\sum_m |c_{g,m}|^p \right)^{q/p} \right)^{1/q} \tag{3}$$

[12] Kai Siedenburg and Monika Doerfler. "Structured sparsity for audio signals". In: *Proc. of the 14th International Conference on Digital Audio Effects (DAFx-11), Paris, France*. Jan. 2011.

Mixed norm in time and frequency

In the context of time-frequency dictionaries, a natural step beyond classical sparsity approaches is the introduction of sparsity criteria which take into account the two-dimensionality of the time-frequency representations used. Mixed norms on the coefficient arrays make it possible to enforce sparsity in one domain and diversity and persistence in the other domain.[12]

Sparsity of coefficients may be enforced by ℓ^1-regression, also known as the *Lasso* [2]. Given a noisy observation $y = s + e$ in $\mathbb{C}^L$ it finds

$$\hat{c} = \arg\min_{c \in \mathbb{C}^P} \frac{1}{2}\|y - \Phi c\|_2^2 + \lambda \Psi(c) \tag{2}$$

with penalty term $\Psi(\cdot) = \|\cdot\|_1$ and $\lambda > 0$. Since the sequence $c_{k,j}$ is ordered along two dimensions for Gabor frames, the ℓ^1-prior Ψ in (2) may be replaced by a two-dimensional mixed norm $\ell^{p,q}$ which acts differently on groups (indexed by g in the sequel, may be either time or frequency) and their members (indexed by m):

$$\Psi(c) = \|c\|_{p,q} = \left(\sum_g \left(\sum_m |c_{g,m}|^p \right)^{q/p} \right)^{1/q} \tag{3}$$

[12] Kai Siedenburg and Monika Doerfler. "Structured sparsity for audio signals". In: *Proc. of the 14th International Conference on Digital Audio Effects (DAFx-11), Paris, France*. Jan. 2011.

Elitist and Group LASSO

Two new and strikingly different shrinkage operators, E-LASSO and G-LASSO:

1. Elitist-LASSO is the solution $\widehat{\mathbf{x}}$ of problem $\mathcal{P}_{1,2}$, using $l_{1,2}$ coefficient penalty. Each coefficient is shrunk individually, but the corresponding threshold depends on its 1D neighborhood; can be understood as an elitist group shrinkage, since most members of a given group are thresholded, and only the *emerging* (or best) coefficients of each group remain, or "only the best coefficients are chosen per group"[13]
2. Group-LASSO is the solution $\widehat{\mathbf{x}}$ of problem $\mathcal{P}_{2,1}$, using $l_{2,1}$ coefficient penalty. A 1D group of coefficients is either globally retained or discarded. This may be understood as a united group shrinkage, since the same threshold applies to all members of a given group; "one keeps entire groups of coefficients, namely the most energetic groups"

E-LASSO promotes sparsity within groups of coefficients instead of sparsity across groups like G-LASSO

[13] Kowalski and Torrésani, 2009.

Elitist and Group LASSO

Two new and strikingly different shrinkage operators, E-LASSO and G-LASSO:

1. Elitist-LASSO is the solution $\widehat{\mathbf{x}}$ of problem $\mathcal{P}_{1,2}$, using $l_{1,2}$ coefficient penalty. Each coefficient is shrunk individually, but the corresponding threshold depends on its 1D neighborhood; can be understood as an elitist group shrinkage, since most members of a given group are thresholded, and only the *emerging* (or best) coefficients of each group remain, or "only the best coefficients are chosen per group"[13]
2. Group-LASSO is the solution $\widehat{\mathbf{x}}$ of problem $\mathcal{P}_{2,1}$, using $l_{2,1}$ coefficient penalty. A 1D group of coefficients is either globally retained or discarded. This may be understood as a united group shrinkage, since the same threshold applies to all members of a given group; "one keeps entire groups of coefficients, namely the most energetic groups"

E-LASSO promotes sparsity within groups of coefficients instead of sparsity across groups like G-LASSO

[13] Kowalski and Torrésani, 2009.

Group-LASSO

Group-LASSO:[14] solve the problem of finding signifiance maps (i.e. locations of significant coefficients) in the transform domain

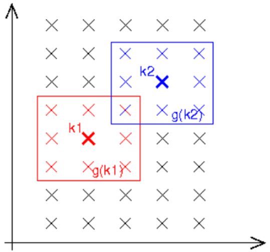


Figure: Group-LASSO regression on two overlapping groups[15]

[14] Matthieu Kowalski and Bruno Torrésani. "Sparsity and persistence: Mixed norms provide simple signal models with dependent coefficients". In: *Signal Image and Video Processing* 3 (Sept. 2009). DOI: 10.1007/s11760-008-0076-1.

[15] Matthieu Kowalski and Bruno Torrésani. "Structured Sparsity: from Mixed Norms to Structured Shrinkage". In: *SPARS09-Signal Processing with Adaptive Sparse Structured Representations* 53 (Apr. 2009).

Structured sparsity

2021-03-24

└Group-LASSO

Group-LASSO
Group-LASSO:[14] solve the problem of finding signifiance maps (i.e. locations of significant coefficients) in the transform domain
Figure: Group-LASSO regression on two overlapping groups[15]
[14] Matthieu Kowalski and Bruno Torrésani. "Sparsity and persistence: Mixed norms provide simple signal models with dependent coefficients". In: Signal Image and Video Processing 3 (Sept. 2009). DOI: 10.1007/s11760-008-0076-1.
[15] Matthieu Kowalski and Bruno Torrésani. "Structured Sparsity: from Mixed Norms to Structured Shrinkage". In: SPARS09-Signal Processing with Adaptive Sparse Structured Representations 53 (Apr. 2009).

Tonal/transient separation with Group-LASSO

WMDCT[16] + Group-LASSO – “audioshrink”[17]

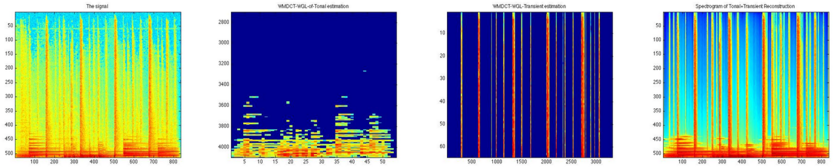


Figure: Audioshrink for tonal/transient separation in jazz music

Use 2 WMDCT transforms (wide + narrow window) + Group-LASSO to shrink input signal into significant coefficients in “time” and “frequency” groups. Example: 🔊 mix, 🔊 tonal, 🔊 transient

[16] Kai Siedenburg and Monika Doerfler. “Structured sparsity for audio signals”. In: *Proc. of the 14th International Conference on Digital Audio Effects (DAFx-11), Paris, France*. Jan. 2011.

[17] `https://homepage.univie.ac.at/monika.doerfler/StrucAudio.html`, `https://ltfat.github.io/doc/demos/demo_audioshrink.html`

2021-03-24

Structured sparsity

└Tonal/transient separation with Group-LASSO

Tonal/transient separation with Group-LASSO

WMDCT[16] + Group-LASSO – “audioshrink”[17]

Figure: Audioshrink for tonal/transient separation in jazz music

Use 2 WMDCT transforms (wide + narrow window) + Group-LASSO to shrink input signal into significant coefficients in “time” and “frequency” groups. Example: 🔊 mix, 🔊 tonal, 🔊 transient

[16] Kai Siedenburg and Monika Doerfler. “Structured sparsity for audio signals”. In: Proc. of the 14th International Conference on Digital Audio Effects (DAFx-11), Paris, France. Jan. 2011.

[17] https://homepage.univie.ac.at/monika.doerfler/StrucAudio.html, https://ltfat.github.io/doc/demos/demo_audioshrink.html

WMDCT tonal/transient dictionaries

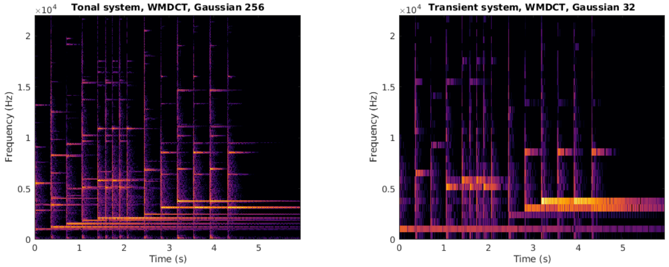


Figure: WMDCT + Group-LASSO shrink tonal/transient separation[18]

2 Gabor systems + `franagrouplasso`:

> *franagrouplasso(F,f,lambda) solves the group LASSO regression problem in the time-frequency domain: minimize a functional of the synthesis coefficients defined as the sum of half the l_2 norm of the approximation error and the mixed l_1/l_2 norm of the coefficient sequence, with a penalization coefficient lambda.*

[18] `https://ltfat.github.io/doc/frames/franagrouplasso.html`, `https://ltfat.github.io/doc/demos/demo_audioshrink.html`

WMDCT tonal/transient dictionaries
Figure: WMDCT + Group-LASSO shrink tonal/transient separation[18]
2 Gabor systems + franagrouplasso:
franagrouplasso(F,f,lambda) solves the group LASSO regression problem in the time-frequency domain: minimize a functional of the synthesis coefficients defined as the sum of half the l_2 norm of the approximation error and the mixed l_1/l_2 norm of the coefficient sequence, with a penalization coefficient lambda.
[18] https://ltfat.github.io/doc/frames/franagrouplasso.html, https://ltfat.github.io/doc/demos/demo_audioshrink.html

Lasso comparison for tonal/transient separation

Tonal layer is expected to be sparsely represented in the frequency domain, with emergent frequencies that may evolve slowly with time (i.e. almost horizontal lines of large MDCT coefficients). Lasso choices:

- E-LASSO (sparse within group) with the time label as group label
- G-LASSO (sparse across groups) with the frequency label as group label - **bad choice** because of the slow evolution in time of frequencies

Transient layer is expected to be sparse in time, but spread out in the frequency domain. Lasso choices:

- E-LASSO (sparse within group) with the frequency label as group label
- G-LASSO (sparse across groups) with the time label as group label – **good choice** because transients are sharply time-localized

└Lasso comparison for tonal/transient separation

Lasso comparison for tonal/transient separation

Tonal layer is expected to be sparsely represented in the frequency domain, with emergent frequencies that may evolve slowly with time (i.e. almost horizontal lines of large MDCT coefficients). Lasso choices:

- E-LASSO (sparse within group) with the time label as group label
- G-LASSO (sparse across groups) with the frequency label as group label - **bad choice** because of the slow evolution in time of frequencies

Transient layer is expected to be sparse in time, but spread out in the frequency domain. Lasso choices:

- E-LASSO (sparse within group) with the frequency label as group label
- G-LASSO (sparse across groups) with the time label as group label – **good choice** because transients are sharply time-localized

Transform to dictionary

Mallat and Zhang[19] *proposed a novel sparse signal expansion scheme based on the selection of a small subset of functions from a general overcomplete dictionary of functions. Chen, Donoho and Saunders*[20] *published their influential paper on the Basis Pursuit, and the two works signalled the beginning of a fundamental move from transforms to dictionaries for sparse signal representation … The terminological change enclosed the idea that a signal was allowed to have more than one description in the representation domain, and that selecting the best one depended on the task*[21]

[19] S.G. Mallat and Zhifeng Zhang. "Matching Pursuits with Time-Frequency Dictionaries". In: *Trans. Sig. Proc.* 41.12 (Dec. 1993), pp. 3397–3415. ISSN: 1053-587X. DOI: 10.1109/78.258082.

[20] S.S. Chen, D. Donoho, and Michael Saunders. "Atomic decomposition by basis pursuit[J]". In: *Siam Review* 43 (Jan. 2001), pp. 33–61.

[21] Ron Rubinstein, Alfred Bruckstein, and Michael Elad. "Dictionaries for Sparse Representation Modeling". In: *Proceedings of the IEEE* 98 (July 2010), pp. 1045–1057. DOI: 10.1109/JPROC.2010.2040551. URL: https://www.cs.technion.ac.il/~ronrubin/Publications/dictdesign.pdf.

Transform to dictionary

Mallat and Zhang[19] *proposed a novel sparse signal expansion scheme based on the selection of a small subset of functions from a general overcomplete dictionary of functions. Chen, Donoho and Saunders*[20] *published their influential paper on the Basis Pursuit, and the two works signalled the beginning of a fundamental move from transforms to dictionaries for sparse signal representation … The terminological change enclosed the idea that a signal was allowed to have more than one description in the representation domain, and that selecting the best one depended on the task*[21]

[19] S.G. Mallat and Zhifeng Zhang. "Matching Pursuits with Time-Frequency Dictionaries". In: *Trans. Sig. Proc.* 41.12 (Dec. 1993), pp. 3397–3415. ISSN: 1053-587X. DOI: 10.1109/78.258082.
[20] S.S. Chen, D. Donoho, and Michael Saunders. "Atomic decomposition by basis pursuit[J]". In: *Siam Review* 43 (Jan. 2001), pp. 33–61.
[21] Ron Rubinstein, Alfred Bruckstein, and Michael Elad. "Dictionaries for Sparse Representation Modeling". In: *Proceedings of the IEEE* 98 (July 2010), pp. 1045–1057. DOI: 10.1109/JPROC.2010.2040551. URL: https://www.cs.technion.ac.il/~ronrubin/Publications/dictdesign.pdf.

Sparsity and entropy

Sparsity and entropy are complementary concepts

Given a signal x, its transform w, and probability mass function p:

- *Sparsity, compressibility:* the property of concentrating most of the energy of x in few coefficients of w
- *Entropy, uncertainty:* the property of not concentrating most of the probability mass of x in few atoms of p

For a given signal x, the uncertainty (randomness) of its elements defines the compressibility (compactness) of its coefficients w in a given domain[22]

[22] Giancarlo Pastor et al. *Mathematics of Sparsity and Entropy: Axioms, Core Functions and Sparse Recovery*. 2015. arXiv: 1501.05126 [cs.IT].

Sparsity and entropy

Sparsity and entropy are complementary concepts

Given a signal x, its transform w, and probability mass function p:

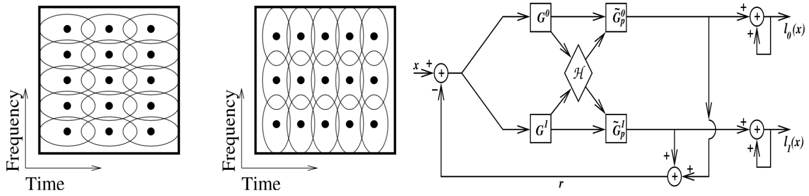
- *Sparsity, compressibility:* the property of concentrating most of the energy of x in few coefficients of w
- *Entropy, uncertainty:* the property of not concentrating most of the probability mass of x in few atoms of p

For a given signal x, the uncertainty (randomness) of its elements defines the compressibility (compactness) of its coefficients w in a given domain[22]

[22] Giancarlo Pastor et al. *Mathematics of Sparsity and Entropy: Axioms, Core Functions and Sparse Recovery*. 2015. arXiv: 1501.05126 [cs.IT].

Time-Frequency Jigsaw Puzzle

1. Create time-frequency "super-tiles" by superimposing a large window + small window Gabor analysis
2. Use Rényi entropy to set coefficients to zero where sound has more entropy than random white noise



(a) TF supertiles (b) Entropy criterion

Figure: TF Jigsaw Puzzle tonal/transient separation[23]

[23] Florent Jaillet and Bruno Torrésani. "Time-Frequency Jigsaw Puzzle: adaptive multiwindow and multilayered Gabor expansions". In: *International Journal of Wavelets, Multiresolution and Information Processing* 05 (Mar. 2007). DOI: 10.1142/S0219691307001768.

Time-Frequency Jigsaw Puzzle
1. Create time-frequency "super-tiles" by superimposing a large window + small window Gabor analysis
2. Use Rényi entropy to set coefficients to zero where sound has more entropy than random white noise
(a) TF supertiles (b) Entropy criterion
Figure: TF Jigsaw Puzzle tonal/transient separation[23]
[23] Florent Jaillet and Bruno Torrésani. "Time-Frequency Jigsaw Puzzle: adaptive multiwindow and multilayered Gabor expansions". In: International Journal of Wavelets, Multiresolution and Information Processing 05 (Mar. 2007). DOI: 10.1142/S0219691307001768.

- i.e. good tonal/good transient
- high entropy = indicating sound is poorly represented

TFJigsaw

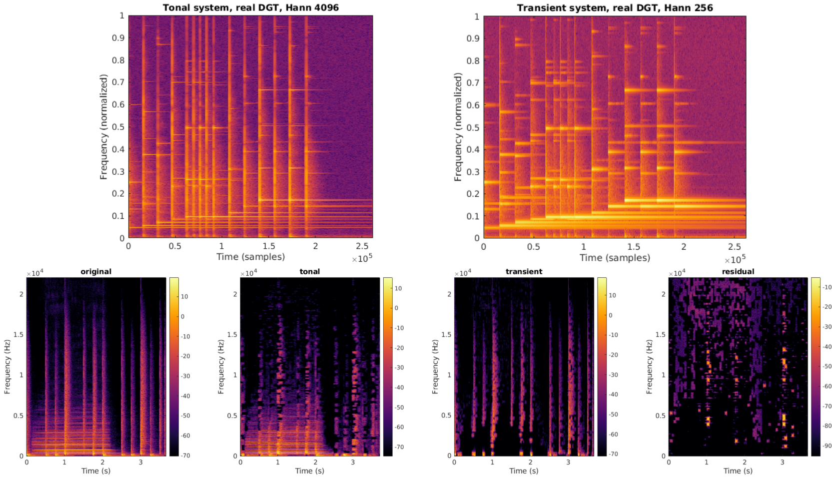


Figure: TF Jigsaw Puzzle tonal/transient separation[24]

[24] https://ltfat.github.io/doc/sigproc/tfjigsawsep.html, https://github.com/ltfat/ltfat/blob/master/demos/demo_tfjigsawsep.m

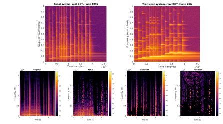


Figure: TF Jigsaw Puzzle tonal/transient separation[24]

[24] https://ltfat.github.io/doc/sigproc/tfjigsawsep.html, https://github.com/ltfat/ltfat/blob/master/demos/demo_tfjigsawsep.m