Audio inpainting using structured sparsity

Pavel Rajmic
(joint work with C. Wiesmeyr, V. Mach, and N. Holighaus)

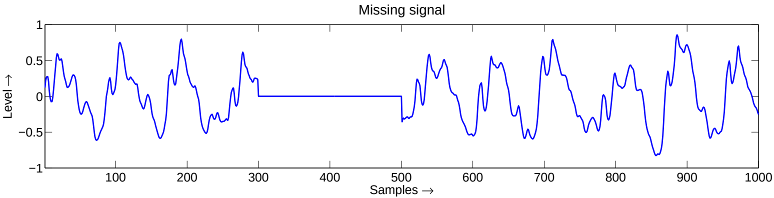
June 3, 2014 @ Strobl

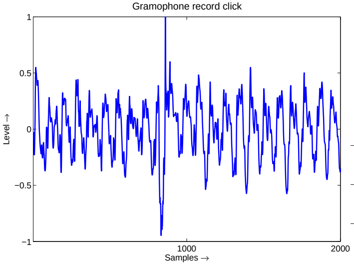
Motivation

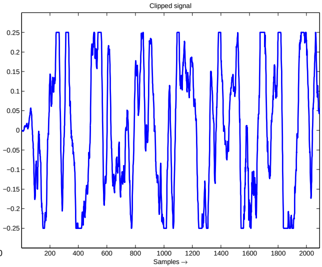
- Digital restoration of the phonograph cylinders recordings *[Mach 2012]*
- Wax cylinders more than 100 years old
- Joint project of Czech, Austrian and Slovak Academies of Sciences, 2009–2012
- Severe corruptions $\rightarrow$ delete block and try to interpolate

Motivation

- Signal loss; clicks; clipping (saturation)
- Applications: gramophone records, wax cylinders, magnetic tapes, packet loss in VoIP, munching removal. . .





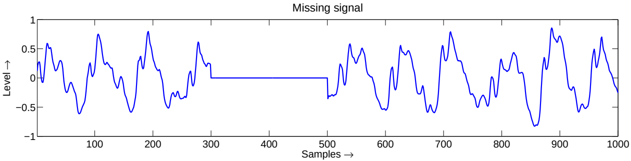


Outline

- Problem formulation
- Common methods
- Audio inpainting utilizing **sparsity and structured sparsity**
- Experiments
- Remarks, open problems

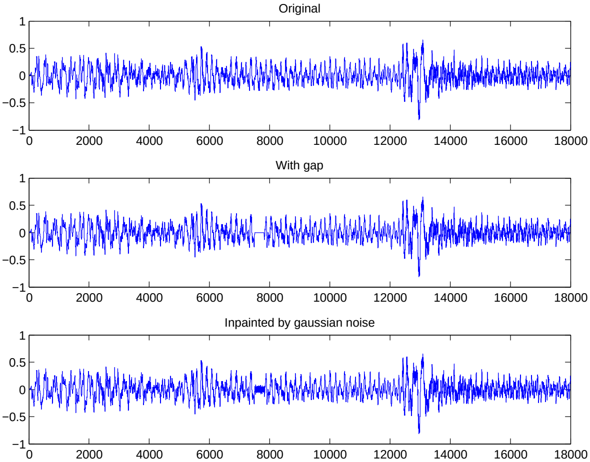
Audio Inpainting – Problem Statement

- Original signal in time domain (suppose we know it): $\mathbf{y}$
- Reliable samples: $\mathbf{y}^{\mathrm{r}} = \mathbf{M}^{\mathrm{r}}\mathbf{y}$, with $\mathbf{M}^{\mathrm{r}}$ masking operator (projection matrix containing zeros on the diagonal)
- Using only information from reliable part $\mathbf{y}^{\mathrm{r}}$
- the goal is to approximate the missing data in the "gap": $\mathbf{y}^{\mathrm{m}} = \mathbf{M}^{\mathrm{m}}\mathbf{y}$
- term *inpainting* comes from image processing field



Audio Inpainting – Which gaps lengths are relevant?

- Inpaint by noise: it works for very short gaps
- We concentrate on real gaps, length up to tens of miliseconds



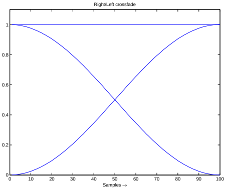
Autoregression-based methods and related

- Older methods based on AR modelling in time domain

$$y[i] = \sum_{j=1}^{k} a_j \, y[i-j] + u[i],$$

- First estimating AR coefficients, then extrapolation from two sides and their crossfading

$$\hat{y}(i) = w_{\mathrm{L}}(i-l)\hat{y}_{\mathrm{L}}(i) + w_{\mathrm{R}}(i-l)\hat{y}_{\mathrm{R}}(i)$$



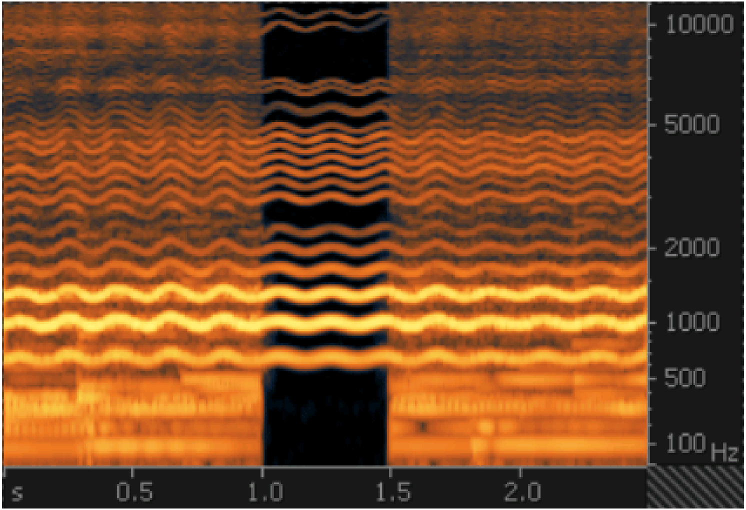
- See *[Janssen 1986]*, *[Etter 1996]* etc.
- Useful in speech inpaiting

Autoregression-based methods and related

Generalized AR approach — sinusoidal modelling, interpolation of partials

- *[Lagrange et al. 2005]*
    - AR modelling of the parameters of partial harmonics, *not* the time samples
    - Consider tremolo (amplitude modulation) or vibrato (frequency modulation)
    - Must cope with pairing of harmonics from the left- and right-side, and treat those which are single
- *[Lukin & Todd 2008]*
    - Adding back the right noise
    - Simultaneous AR parameters estimation for both sides of gap
- depend heavily on the separation of partials

Autoregression-based methods and related

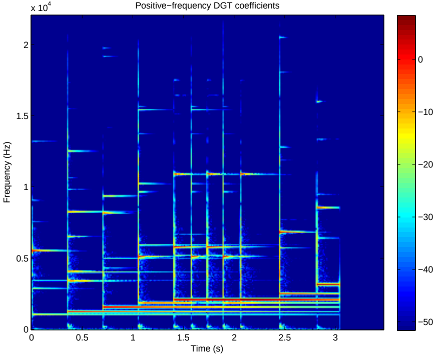


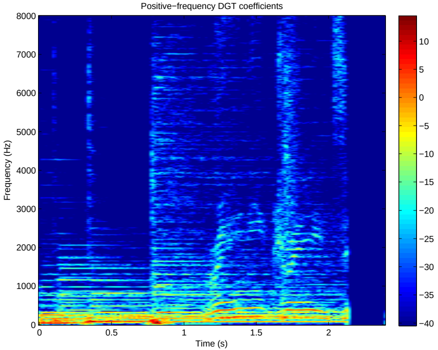
Sparse modelling approach

- Realistic assumption on most audio signals: it is approximately sparse in a time-frequency dictionary:

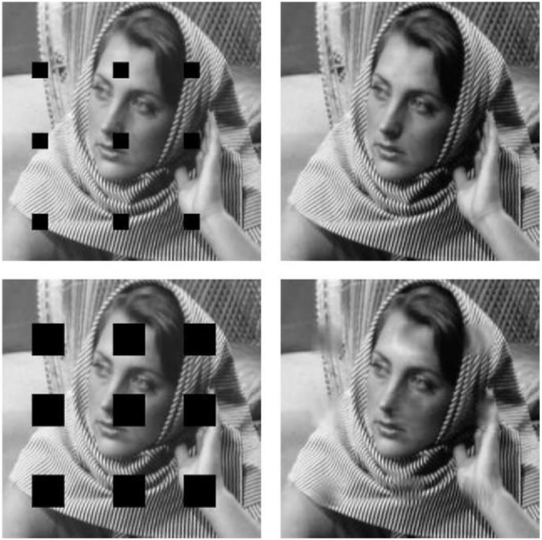
$$\mathbf{y} \approx \mathbf{D}\mathbf{x},$$

atoms as columns of $\mathbf{D}$, and $\|\mathbf{x}\|_0$ is small





Sparse modelling approach – image inpainting



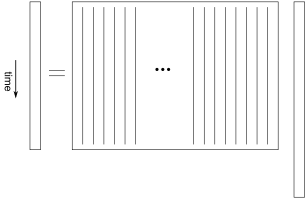
[Elad, 2010]

Sparse modelling approach

- Following such image inpainting approaches,
- sparsity-based audio-inpainting was introduced by *[Adler et al., 2012]*

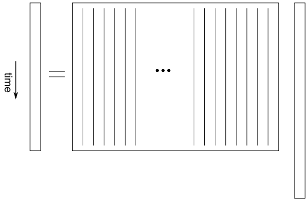
Sparse modelling approach

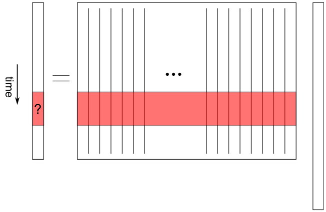
- Following such image inpainting approaches,
- sparsity-based audio-inpainting was introduced by *[Adler et al., 2012]*
- We assume $\mathbf{y} \approx \mathbf{Dx}$ with sparse $\mathbf{x}$



Sparse modelling approach

- Following such image inpainting approaches,
- sparsity-based audio-inpainting was introduced by *[Adler et al., 2012]*
- We assume $\mathbf{y} \approx \mathbf{D}\mathbf{x}$ with sparse $\mathbf{x}$





- The main idea is to
  1. Obtain sparse signal coefficients $\mathbf{x}$ from reliable samples and reduced dictionary $\mathbf{D}^{\mathrm{r}} = \mathbf{M}^{\mathrm{r}}\mathbf{D}$:
  $$\hat{\mathbf{x}} = f(\mathbf{y}^{\mathrm{r}}, \mathbf{D}^{\mathrm{r}})$$
  2. Restore signal using full dictionary
  $$\hat{\mathbf{y}} = \mathbf{D}\hat{\mathbf{x}}$$

Obtaining coefficients

We form an optimization problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t. } \|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2 \leq \delta$$

- solved via Orthogonal Matching Pursuit (like in SMALLbox)
- or relaxing to the convex ℓ_1-norm approximation

ℓ_1-minimization approach

- BPDN, constrained relaxed problem of this type

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \leq \delta,$$

- can be solved as the unconstrained, equivalent one, termed LASSO

$$\arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1,$$

  where λ influences degree of sparse regularization by penalization of high $\|\mathbf{x}\|_1$

ℓ_1-minimization approach

- BPDN, constrained relaxed problem of this type

$$\arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \leq \delta,$$

- can be solved as the unconstrained, equivalent one, termed LASSO

$$\arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1,$$

  where λ influences degree of sparse regularization by penalization of high $\|\mathbf{x}\|_1$

- In our case, the coefficients can be obtained as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1,$$

  and then the completed signal part takes the form

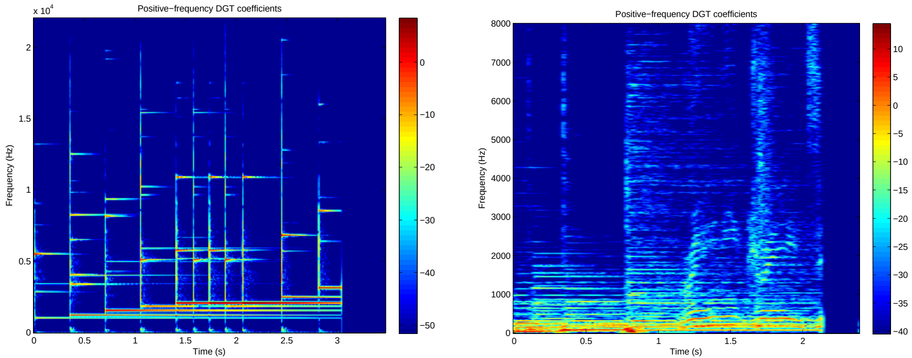
$$\hat{\mathbf{y}}^{\mathrm{m}} = \mathbf{M}^{\mathrm{m}}\mathbf{D}\hat{\mathbf{x}}$$

ℓ_1-minimization – soft thresholding

- Note that $\|\mathbf{x}\|_1 = \sum_{i,j} |x_{i,j}|$ considers *each coefficient independently*
- Obtaining $\hat{\mathbf{x}}$ by iterative thresholding: (F)ISTA *[Beck 2009]*
- includes soft thresholding in each iteration:

$$x_{i,j} \;\leftarrow\; x_{i,j}\left(1-\frac{\lambda}{|x_{i,j}|}\right)^{+}$$

Structured Sparsity



- Spectrogram of natural signals is not only nearly sparse, but also the non-zero coefficients are *structured*:
  - tonal part. . . horizontal groups of coefficients
  - transient part. . . vertical groups
  - natural musical instruments. . . several linked harmonics

Structured Sparsity

- Structured LASSO

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{\mathrm{S}}$$

Structured Sparsity

- Structured LASSO

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{\mathrm{S}}$$

- Structured LASSO using mixed norms

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{p,q}$$

where p represents a within-group penalty and q is across-group penalty

Structured Sparsity

- Structured LASSO

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{\mathrm{S}}$$

- Structured LASSO using mixed norms

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{p,q}$$

  where p represents a within-group penalty and q is across-group penalty

- LASSO: $p = 1,\ q = 1$
- Group-LASSO: $p = 2,\ q = 1$
- Elitist-LASSO: $p = 1,\ q = 2$

Structured Sparsity – Group LASSO soft thresholding

- Group-LASSO: $p = 2$, $q = 1$
- with groups as rows of spectrogram
- $\|\mathbf{x}\|_{2,1} = \sum_i \|\mathbf{x}_{i,:}\|_2$

Structured Sparsity – Group LASSO soft thresholding

- Group-LASSO: $p = 2$, $q = 1$
- with groups as rows of spectrogram
- $\|\mathbf{x}\|_{2,1} = \sum_i \|\mathbf{x}_{i,:}\|_2$
- in (F)ISTA, the thresholding step takes the form

$$x_{i,j} \;\leftarrow\; x_{i,j}\left(1 - \frac{\lambda}{\|\mathbf{x}_{i,:}\|_2}\right)^{+}$$

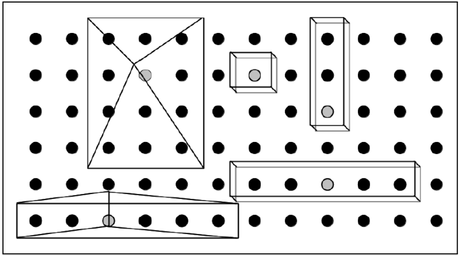
- tiny coefficients can be retained when the group is strong enough

Structured Sparsity – more complex modelling

- shaping groups,
- overlapping groups,
- weighting group members

Structured Sparsity – more complex modelling

- shaping groups,
- overlapping groups,
- weighting group members
- in the audio-processing context: Kowalski, Siedenburg, Dörfler, Torrésani, Bayram . . .



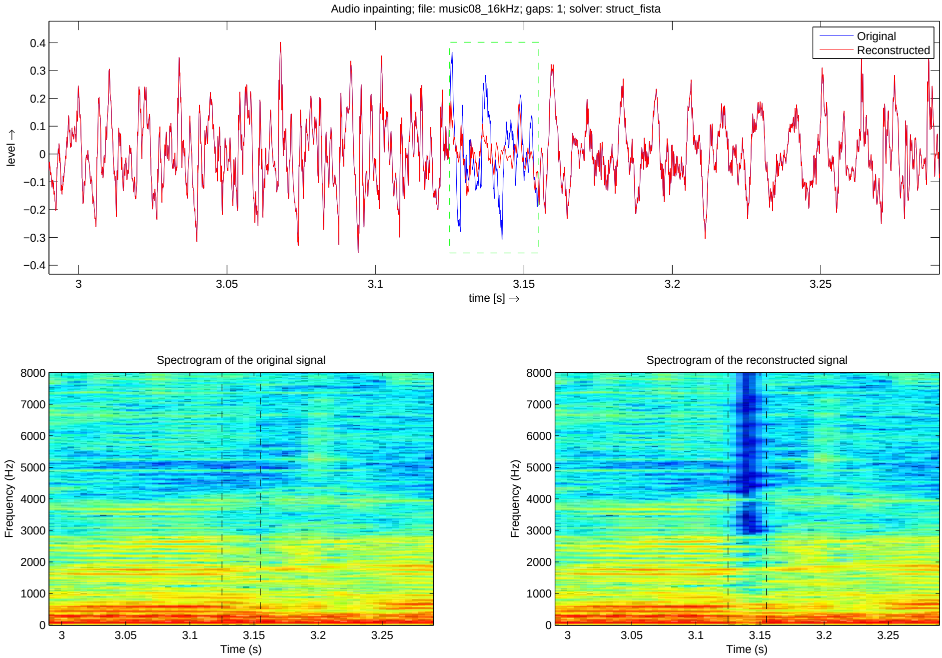
(from *[Siedenburg 2011]*)

Experiment — Single gap inpainting

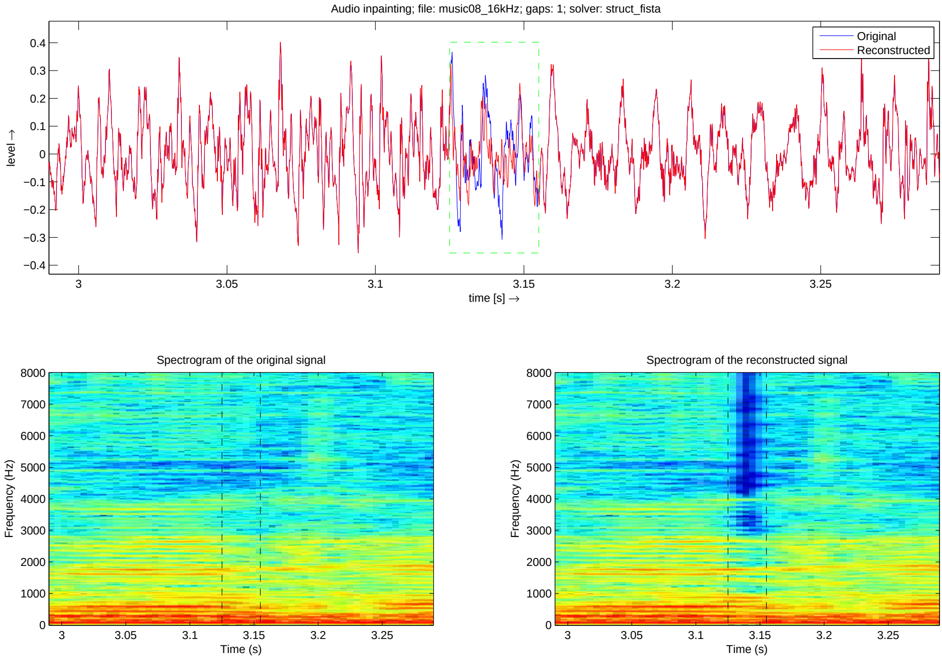
- Signal: pop music, sampled at 16 kHz
- Gap size: 30 ms (480 samples)
- Dictionary: tight Gabor frame,
  Hann window 64 ms, time shift 16 ms, 1024 channels

Experiment — Single gap inpainting

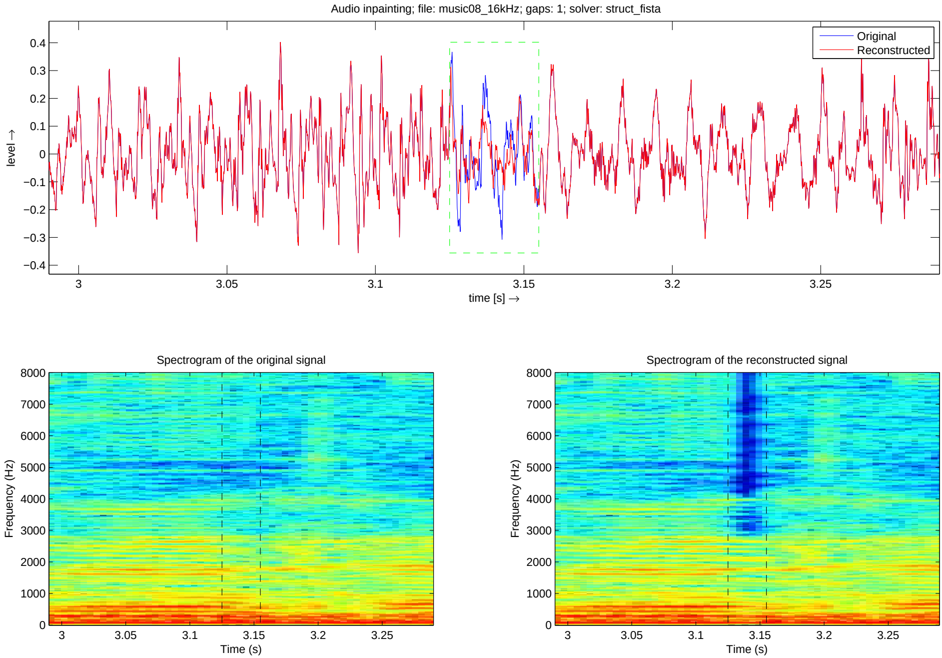
- Signal: pop music, sampled at 16 kHz
- Gap size: 30 ms (480 samples)
- Dictionary: tight Gabor frame,
  Hann window 64 ms, time shift 16 ms, 1024 channels
- Structure used: Horizontal groups, unweighted,
  size from single coefficient (64 ms) to 35 coefficients (624 ms)
- structured FISTA reconstruction with same λ
- evaluation in terms of SNR



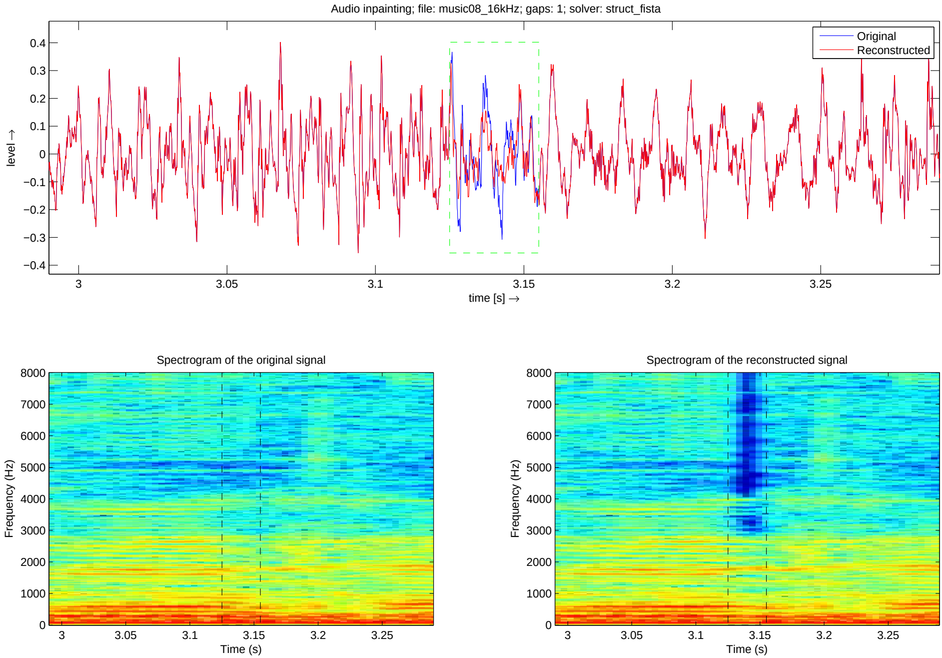
neighbourhood = 1



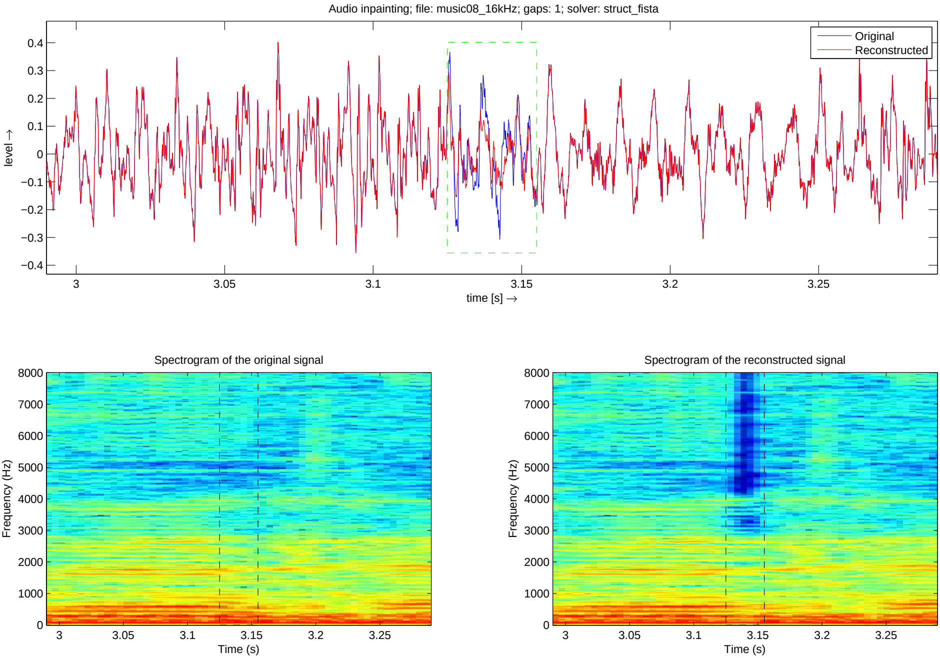
neighbourhood = 3



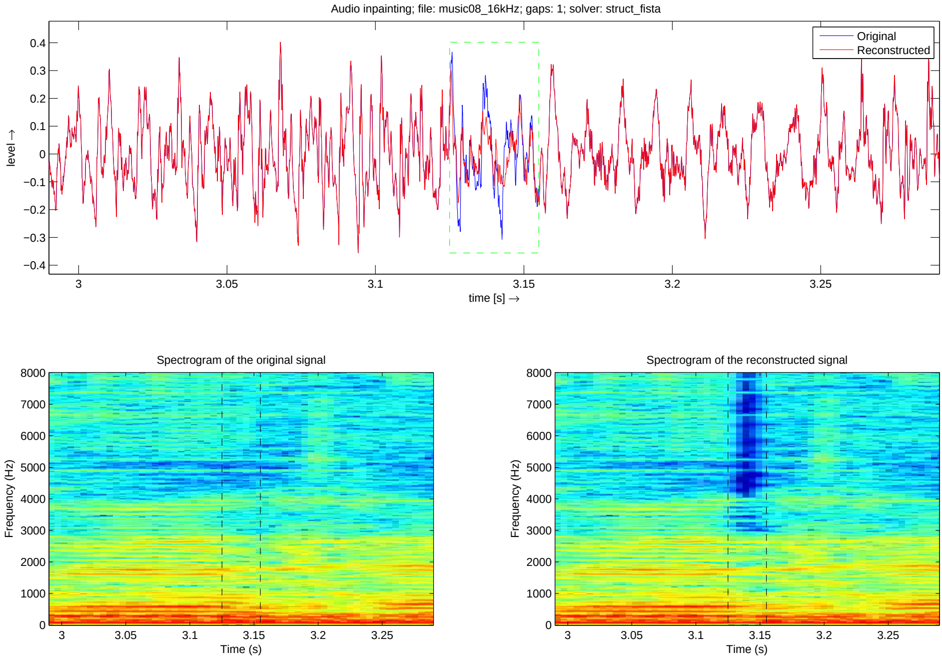
neighbourhood $= 5$



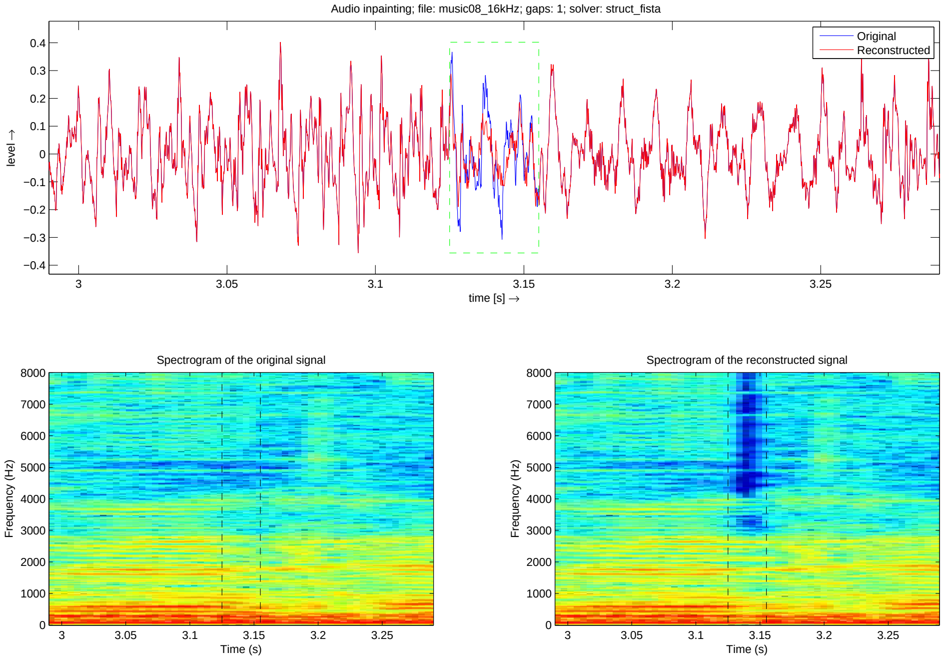
neighbourhood = 7



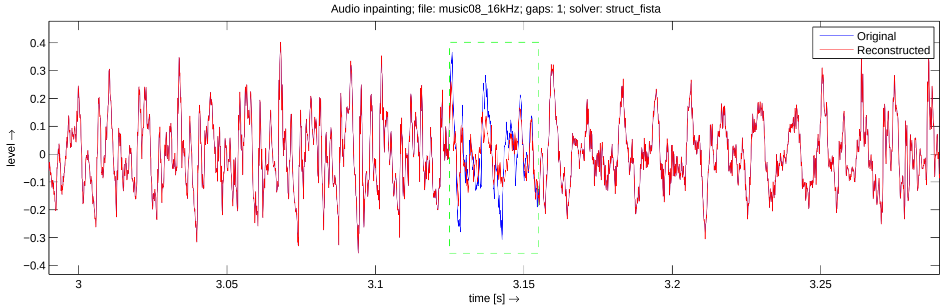
neighbourhood $= 9$

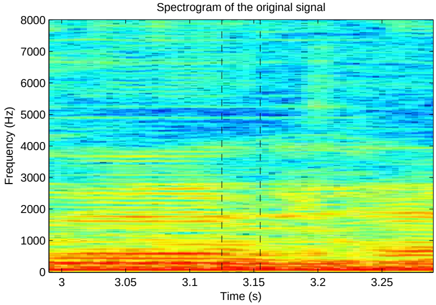


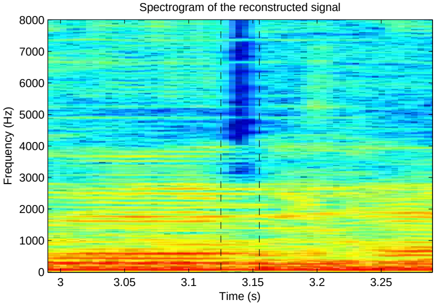
neighbourhood = 11



neighbourhood = 13

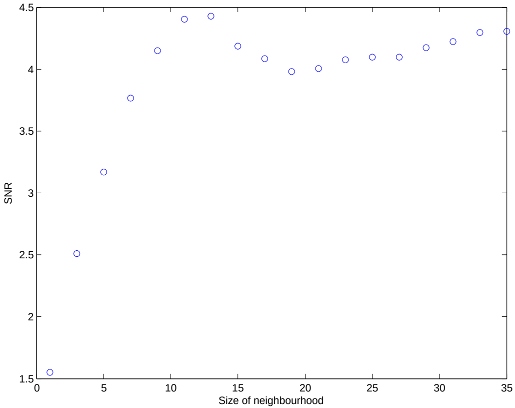






neighbourhood $= 15$

Experiment — Single gap inpainting



- best SNR achieved by groups of 11 or 13 neighbours (250 ms)
- groupless LASSO performs worse

Remarks & questions

- Subjective or close-to-subjective evaluation must take place; PEAQ, PEMO-Q
- Our toolbox relies on LTFAT (time-frequency frames) and UnlocBox (convex optimization algorithms)

Dictionary selection

- Static
  - DCT, MDCT for segmented signal (SMALLbox, *[Adler 2012]*)
  - Gabor transform
  - non-uniform filter banks, like:
    - Constant-Q *[Velasco 2011]*
    - ERBlet (adjusted to human sound perception) *[Necciari 2013]*

Dictionary selection

- Static
  - DCT, MDCT for segmented signal (SMALLbox, *[Adler 2012]*)
  - Gabor transform
  - non-uniform filter banks, like:
    - Constant-Q *[Velasco 2011]*
    - ERBlet (adjusted to human sound perception) *[Necciari 2013]*
- Adaptive
  - with dictionary learning (locally around the gap)
  - K-SVD, INK-SVD *[Aharon 2006]*
  - can improve SNR by a few dB when applied correctly
  - gender-dependent *[Mach 2013]*

Remarks & questions — reweighting the dictionary?

- audio inpainting can be regarded as *inverse problem* where we observe only a projection of the original $\mathbf{y}$

Remarks & questions — reweighting the dictionary?

- audio inpainting can be regarded as *inverse problem* where we observe only a projection of the original $\mathbf{y}$
- recall our data term

$$\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2$$

- here, $\mathbf{D}^{\mathrm{r}}$ can be considered a dictionary for $\mathbf{y}^{\mathrm{r}}$
- but $\mathbf{D}^{\mathrm{r}}$ can have different properties compared to $\mathbf{D}$
- in particular, equal ℓ_2-norms are lost
  (if we assume atoms in $\mathbf{D}$ have the same ℓ_2-norm)

Remarks & questions — reweighting the dictionary?

- audio inpainting can be regarded as *inverse problem* where we observe only a projection of the original $\mathbf{y}$
- recall our data term
$$\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2$$
- here, $\mathbf{D}^{\mathrm{r}}$ can be considered a dictionary for $\mathbf{y}^{\mathrm{r}}$
- but $\mathbf{D}^{\mathrm{r}}$ can have different properties compared to $\mathbf{D}$
- in particular, equal ℓ_2-norms are lost
(if we assume atoms in $\mathbf{D}$ have the same ℓ_2-norm)
- Should dictionary $\mathbf{D}^{\mathrm{r}}$ be ℓ_2-reweighted for data fitting step?

Remarks & questions — reweighting the dictionary?

- audio inpainting can be regarded as *inverse problem* where we observe only a projection of the original $\mathbf{y}$
- recall our data term
$$\|\mathbf{y}^{\mathrm{r}} - \mathbf{D}^{\mathrm{r}}\mathbf{x}\|_2^2$$
- here, $\mathbf{D}^{\mathrm{r}}$ can be considered a dictionary for $\mathbf{y}^{\mathrm{r}}$
- but $\mathbf{D}^{\mathrm{r}}$ can have different properties compared to $\mathbf{D}$
- in particular, equal ℓ_2-norms are lost
(if we assume atoms in $\mathbf{D}$ have the same ℓ_2-norm)
- Should dictionary $\mathbf{D}^{\mathrm{r}}$ be ℓ_2-reweighted for data fitting step?
- Circumvention: analysis formulation

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z}^{\mathrm{r}} - \mathbf{y}^{\mathrm{r}}\|_2^2 + \lambda\|\tilde{\mathbf{D}}\mathbf{z}\|_1$$

Conclusion

- Sparse representations useful in audio inpainting problem
- *Structured* sparsity improves results
- Structure designed to the signal
- Must be combined with other methods when the signal is more complicated

Conclusion

- Sparse representations useful in audio inpainting problem
- *Structured* sparsity improves results
- Structure designed to the signal
- Must be combined with other methods when the signal is more complicated
- Future directions
  - Utilizing the harmonic structure of partials
  - Perceptually motivated non-uniform filterbanks
  - Inpaint also residual noise

Thank you for your attention!