Computational Intelligence Laboratory

Lecture 10

Dictionary Learning

Thomas Hofmann

ETH Zurich – `cil.inf.ethz.ch`

May 19, 2017

Section 1

Compressive Sensing

Compressive Sensing

- Why should we gather huge amounts of information if we then compress it anyway and throw away most of it?
- Let's instead compress data while gathering.
- It decreases acquisition time, power consumption and required storage space.

This idea is called **compressive sensing**.

Compressive Sensing

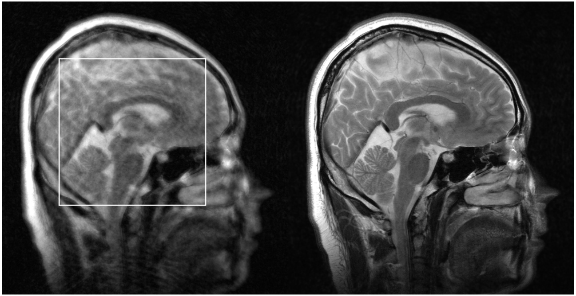
When is it important? Photoshooting in space!

- Saving memory and battery power ...
- ... for a camera which is orbiting Mars – hugely important!
- Fewer images acquired $\Longrightarrow$ less energy consumed
- Storage space could also be an issue

NASA/JPL/Corby Waste

Compressive Sensing for MRI

- Highres MRI: patient has to be perfectly still during scanning
- Standard practice: ask patient to stop respiration
- Scanning time becomes critically important!
- Decreasing number of measurements $\Longrightarrow$ reduced scan time



Xiaojing Ye (2011)

Compressive Sensing: Concept

- Original signal $\mathbf{x} \in \mathbb{R}^D$, K-sparse in orthonormal basis $\mathbf{U}$

$$\mathbf{x} = \mathbf{U}\mathbf{z}, \quad \text{s.t.} \quad \|\mathbf{z}\|_0 = K$$

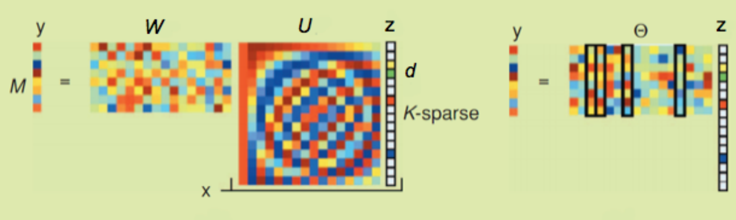
- Main idea: acquire set $\mathbf{y}$ of M linear combinations of signal $\Longrightarrow$ reconstruct signal from these measurements

$$y_k = \langle \mathbf{w}_k, \mathbf{x} \rangle, \quad k = 1, \ldots, M$$

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{U}\mathbf{z} =: \Theta\mathbf{z}, \text{ with } \Theta = \mathbf{W}\mathbf{U} \in \mathbb{R}^{M \times D}$$

  - measurement = linear feature
  - if $M \ll D$: measured signal $\mathbf{y}$ much shorter than $\mathbf{x}$.

Compressive Sensing



$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{U}\mathbf{z} =: \Theta\mathbf{z}, \text{ with } \Theta = \mathbf{W}\mathbf{U} \in \mathbb{R}^{M \times D}$$

- Surprisingly given **any** orthonormal basis $\mathbf{U}$ we can obtain a stable reconstruction for any K-sparse, compressible signal!
- Sufficient conditions:
  1. $\mathbf{W}$ = Gaussian random projection, i.e. $w_{ij} \sim \mathcal{N}(0, \frac{1}{D})$
  2. $M \geq cK \log\left(\frac{D}{K}\right)$, where c is some constant.

Compressive Sensing: Signal Reconstruction

- Recovery of $\mathbf{x} \in \mathbb{R}^D$ from measured signal $\mathbf{y} \in \mathbb{R}^M$
  $\equiv$ need to find sparse representation $\mathbf{z}$:

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{U}\mathbf{z} = \Theta\mathbf{z}, \quad \text{with} \quad \Theta \in \mathbb{R}^{M \times D}$$

  - given $\mathbf{z}$, easily reconstruct $\mathbf{x}$ via $\mathbf{x} = \mathbf{U}\mathbf{z}$
  - finding $\mathbf{z}$ ill-posed: more unknowns than equations ($M \ll D$)

- Optimization problem
  - find sparsest solution s.t. equality holds:

$$\mathbf{z}^* \in \arg\min_{\mathbf{z}} \|\mathbf{z}\|_0, \quad \text{s.t.} \quad \mathbf{y} = \Theta\mathbf{z}$$

  - apply same *reconstruction techniques* as before:
    (1) Convex Optimization or (2) Matching Pursuit

Section 2

Dictionary Learning

Dictionary Learning

Can we work with better and more problem specific dictionaries?

Recap: Dictionary Encoding I

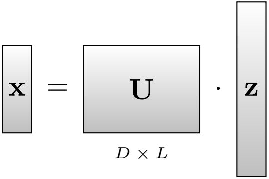
Fixed orthonormal basis:

$$\mathbf{x} = \underset{D \times D}{\mathbf{U}} \cdot \mathbf{z}$$

- Advantage: efficient coding by matrix multiplication $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$
- Disadvantage: only sparse for specific classes of signals
  - strong *a priori* assumptions

Recap: Dictionary Encoding II

Fixed overcomplete basis:



- Advantage: sparse coding for several signal classes
- Disadvantage: finding sparsest code ...
  - may require approximation algorithm (e.g. matching pursuit)
  - problematic if dictionary size L and coherence $m(\mathbf{U})$ are large.

Dictionary Encoding III

Learning the dictionary:

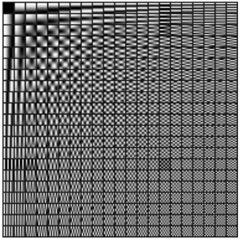
- Advantage: we adapt a dictionary to signal characteristics $\implies$ same approximation error achievable with smaller L
- Challenge: we have to solve a matrix factorization problem

$$\underset{D \times N}{\mathbf{X}} \approx \underset{D \times L}{\mathbf{U}} \cdot \underset{L \times N}{\mathbf{Z}}$$

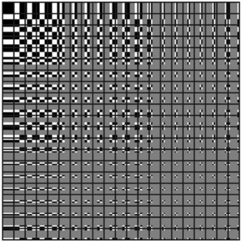
  - subject to sparsity constraint on $\mathbf{Z}$ and
  - subject to atom norm constraint on $\mathbf{U}$.

Dictionary Adaptation

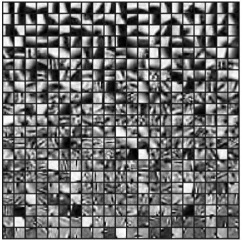
- 8×8 pixel image patches of face images
- 11k examples for training, i.e. $\mathbf{X} \in \mathbb{R}^{64 \times 11000}$
- Dictionary $\mathbf{U} \in \mathbb{R}^{64 \times 441}$ (ca. 7 times overcomplete):



Overcomplete DCT



Overcomplete Haar



Learned dictionary

M. Aharon et al., IEEE Transactions on Signal Processing, 54, 4311-4322, 2006

Inpainting Comparison

Reconstruction:

1. One sparse coding step of observed pixels
2. Predict missing pixels from sparse code

M. Aharon et al., IEEE Transactions on Signal Processing, 54, 4311-4322, 2006

Matrix Factorization

$$(\mathbf{U}^\star, \mathbf{Z}^\star) \in \arg\min_{\mathbf{U},\mathbf{Z}} \|\mathbf{X} - \mathbf{U} \cdot \mathbf{Z}\|_F^2$$

- Frobenius norm: $\|\mathbf{A}\|_F^2 = \sum_{i,j} a_{i,j}^2$
- objective *not* jointly convex in $\mathbf{U}$ and $\mathbf{Z}$
- convex in either $\mathbf{U}$ or $\mathbf{Z}$ (with unique minimum)

Iterative greedy minimization

1. **Coding step**: $\mathbf{Z}^{t+1} \in \arg\min_{\mathbf{Z}} \left\|\mathbf{X} - \mathbf{U}^t\mathbf{Z}\right\|_F^2$,
   subject to $\mathbf{Z}$ being sparse (non-convex) and $\mathbf{U}$ being fixed.
2. **Dictionary update step**: $\mathbf{U}^{t+1} \in \arg\min_{\mathbf{U}} \left\|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\right\|_F^2$,
   subject to $\|\mathbf{u}_l\|_2 = 1$ for all $l = 1, \ldots, L$ and $\mathbf{Z}$ being fixed.

Coding Step

$$\mathbf{Z}^{t+1} \in \arg\min_{\mathbf{Z}} \left\| \mathbf{X} - \mathbf{U}^t \mathbf{Z} \right\|_F^2$$

- Column separable residual: $\|\mathbf{R}\|_F^2 = \sum_{i,j} r_{i,j}^2 = \sum_j \|\mathbf{r}_j\|_2^2$
- N *independent* sparse coding steps: for all $n = 1, \dots N$

$$\begin{aligned} \mathbf{z}_n^{t+1} \quad &\in \quad \arg\min_{\mathbf{z}} \|\mathbf{z}\|_0 \\ \text{s.t.} \quad &\qquad \left\| \mathbf{x}_n - \mathbf{U}^t \mathbf{z} \right\|_2 \leq \sigma \cdot \|\mathbf{x}_n\|_2 \end{aligned}$$

Dictionary Update I

$$\mathbf{U}^{t+1} \in \arg\min_{\mathbf{U}} \left\|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\right\|_F^2$$

- Residual *not separable* in atoms (columns of $\mathbf{U}$)
- Approximation: update one atom at a time ($\forall l$)
  1. Set $\mathbf{U} = [\mathbf{u}_1^t \ \cdots \ \mathbf{u}_l \ \cdots \mathbf{u}_L^t]$, i.e. fix all atoms except $\mathbf{u}_l$.
  2. Isolate $\mathbf{R}_l^t$, the residual that is due to atom $\mathbf{u}_l$.
  3. Find $\mathbf{u}_l^*$ that minimizes $\mathbf{R}_l^t$, subject to $\|\mathbf{u}_l^*\|_2 = 1$.

Dictionary Update II

- Isolate $\mathbf{R}_l^t$: residual due to atom $\mathbf{u}_l$

$$
\begin{aligned}
& \left\| \mathbf{X} - \left[\mathbf{u}_1^t \ \cdots \ \mathbf{u}_l \ \cdots \mathbf{u}_L^t \right] \cdot \mathbf{Z}^{t+1} \right\|_F^2 \\
= & \left\| \mathbf{X} - \left(\sum_{e \neq l} \mathbf{u}_e^t \left(\mathbf{z}_e^{t+1} \right)^\top + \mathbf{u}_l \left(\mathbf{z}_l^{t+1} \right)^\top \right) \right\|_F^2 \\
= & \left\| \mathbf{R}_l^t - \mathbf{u}_l \left(\mathbf{z}_l^{t+1} \right)^\top \right\|_F^2
\end{aligned}
$$

  - $\mathbf{z}_l^\top$ is the l-th row of matrix $\mathbf{Z}$.

Dictionary Update III

How can we find $\mathbf{u}_l^*$?

- $\mathbf{u}_l \left(\mathbf{z}_l^{t+1}\right)^\top$ is an outer product, i.e. a matrix
- Approximating residual with rank 1 matrix

$$\left\|\mathbf{R}_l^t - \mathbf{u}_l \left(\mathbf{z}_l^{t+1}\right)^\top\right\|_F^2$$

- "Approximately" achieved by SVD of $\mathbf{R}_l^t$:

$$\mathbf{R}_l^t = \tilde{\mathbf{U}}\mathbf{\Sigma}\tilde{\mathbf{V}}^\top = \sum_i \sigma_i \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^\top$$

  - $\mathbf{u}_l^* = \tilde{\mathbf{u}}_1$ is first left-singular vector.
  - $\|\mathbf{u}_l^*\|_2 = 1$ naturally satisfied.
  - also update l-th row of $\mathbf{Z}$ (see next slide)

Approximate K-SVD Dictionary Update

Dictionary update by a single power iteration (line 8-9)

1: Input: $\mathbf{X} = \mathbb{R}^{D \times N}$; $\mathbf{U} = \mathbb{R}^{D \times L}$; $\mathbf{Z} = \mathbb{R}^{L \times N}$
2: Output: Updated dictionary $\mathbf{U}$
3: **for** $l \leftarrow 1$ to L **do**
4: $\quad \mathbf{u}_{(:,l)} \leftarrow \mathbf{0}$,
5: $\quad \mathcal{N} \leftarrow \{n | Z_{ln} \neq 0, 1 \leq n \leq N\}$ % active data points
6: $\quad \mathbf{R} \leftarrow \mathbf{X}_{(:,\mathcal{N})} - \mathbf{U}\mathbf{Z}_{(:,\mathcal{N})}$ % residual
7: $\quad \mathbf{g} \leftarrow \mathbf{z}_{(l,\mathcal{N})}^{\top}$
8: $\quad \mathbf{h} \leftarrow \mathbf{R}\mathbf{g} / \|\mathbf{R}\mathbf{g}\|$ % power iteration
9: $\quad \mathbf{g} \leftarrow \mathbf{R}^{\top}\mathbf{h}$
10: $\quad \mathbf{u}_{(:,l)} \leftarrow \mathbf{h}$ % update
11: $\quad \mathbf{z}_{(l,\mathcal{N})} \leftarrow \mathbf{g}^{\top}$
12: **end for**

CD Sigg, T Dikk, JM Buhmann, *Speech Enhancement using Generative Dictionary Learning*, IEEE-TASLP 2012

Initialization

Sensitive to choice of $\mathbf{U}^0$: the initial candidate solution is optimized locally and greedily until no progress possible.

A) Random atoms: Sampling $\left\{\mathbf{u}_l^0\right\}$ on unit sphere

1. Sample with standard normal distribution: $\mathbf{u}_l^0 \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_D\right)$.
2. Scale to unit length: $\mathbf{u}_l^0 \leftarrow \mathbf{u}_l^0 / \left\|\mathbf{u}_l^0\right\|_2$.

B) Samples from $\mathbf{X}$:

1. $\mathbf{u}_l^0 \leftarrow \mathbf{x}_n$, where $n \sim \mathcal{U}\left(1, N\right)$ is sampled uniformly.
2. Scale to unit length: $\mathbf{u}_l^0 \leftarrow \mathbf{u}_l^0 / \left\|\mathbf{u}_l^0\right\|_2$.

C) Fixed overcomplete dictionary, e.g. use overcomplete DCT.

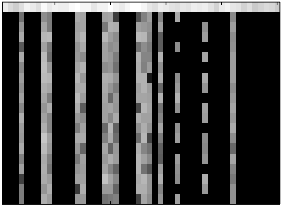
Example

- 8×8 non-overlapping patches
- 20 atoms: 19 initialized randomly, 1 constant atom
- $\sigma = 1/200$
- 40 iterations

Example

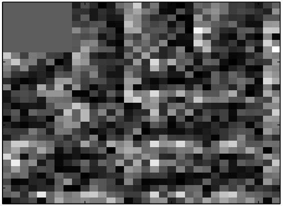
Image

Coding



Approximation

Dictionary

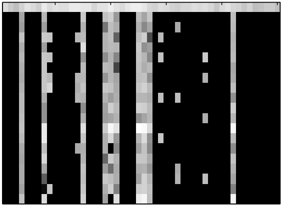


Iteration: $t = 1$

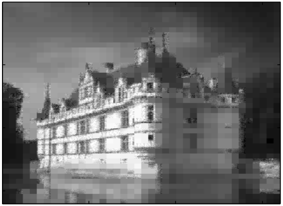
Example

Image

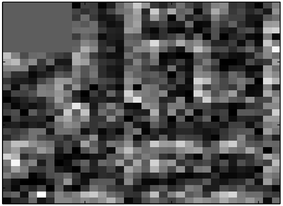
Coding



Approximation



Dictionary

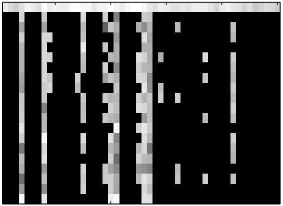


Iteration: $t = 2$

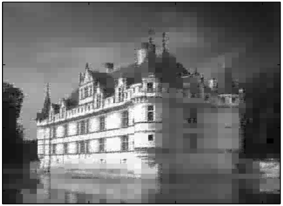
Example

Image

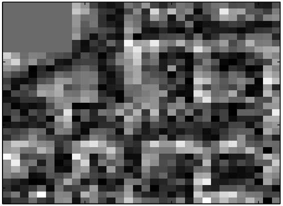
Coding



Approximation



Dictionary

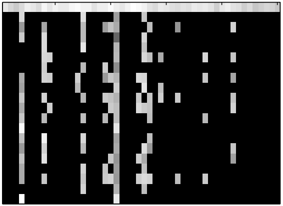


Iteration: $t = 3$

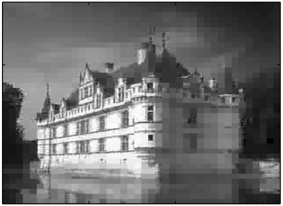
Example

Image

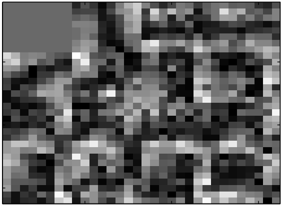
Coding



Approximation



Dictionary

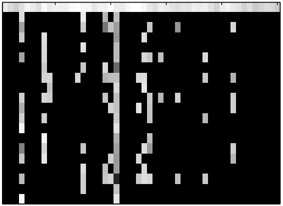


Iteration: $t = 4$

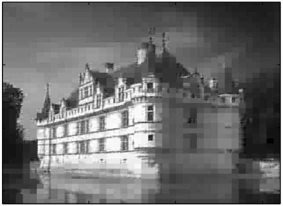
Example

Image

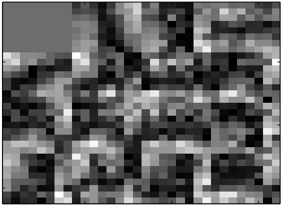
Coding



Approximation



Dictionary

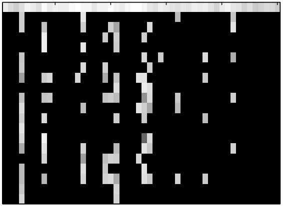


Iteration: $t = 5$

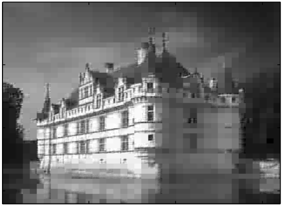
Example

Image

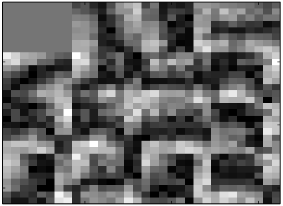
Coding



Approximation



Dictionary

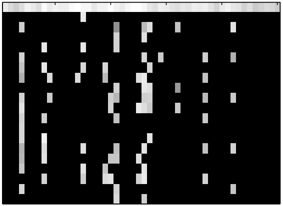


Iteration: $t = 10$

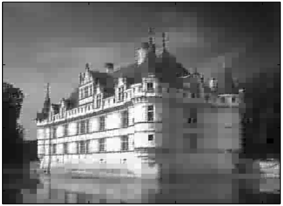
Example

Image

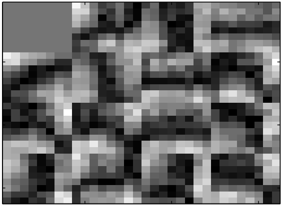
Coding



Approximation



Dictionary

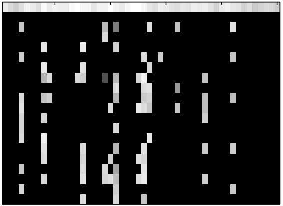


Iteration: $t = 15$

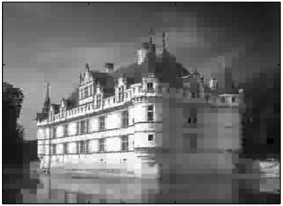
Example

Image

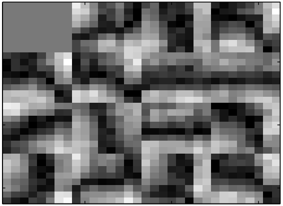
Coding



Approximation

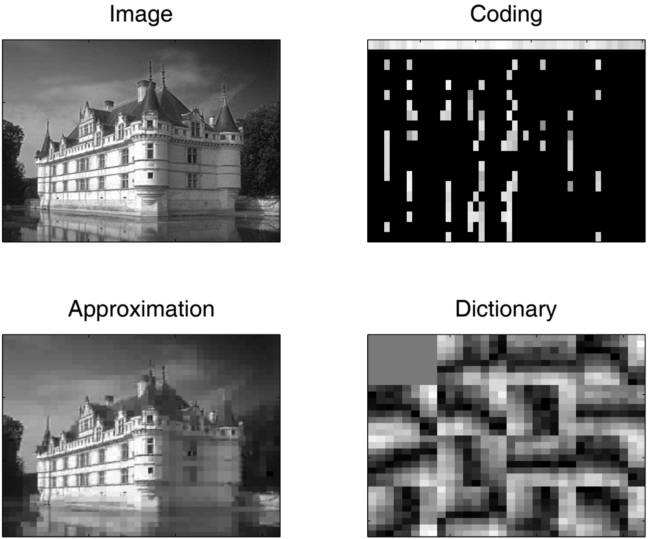


Dictionary



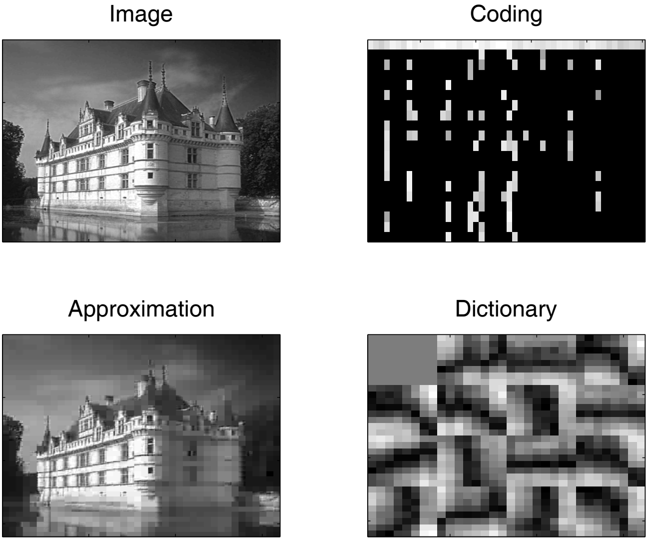
Iteration: $t = 20$

Example



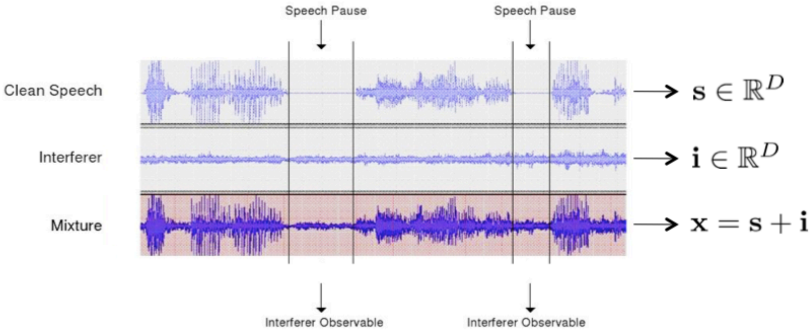
Iteration: $t = 25$

Example



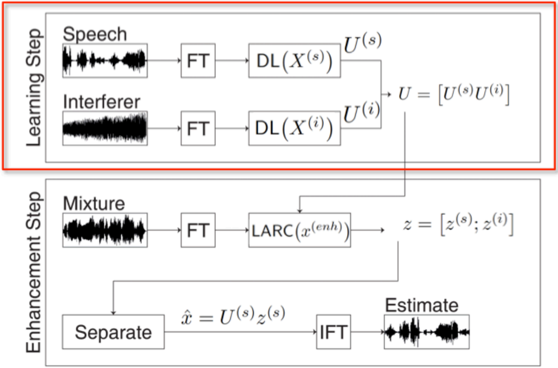
Iteration: $t = 30$

Model Based Speech Enhancement



- *Setting*: Observe additive mixture of speech and interferer signal
- *Target*: Infer clean speech based on the mixed signal
- *Concept*: Exploit speech pause to learn interferer dictionary in an adaptive way

Enhancement Pipeline



- Transform (FT) signal into feature space using short-time Fourier transform (STFT) and modified discrete cosine transform (MDCT)
- Train speech dictionary $\mathbf{U}^{(s)}$ and interferer dictionary $\mathbf{U}^{(i)}$
- Build composite dictionary: $\mathbf{U} = \left[\mathbf{U}^{(s)} \mathbf{U}^{(i)}\right]$

Learning Step

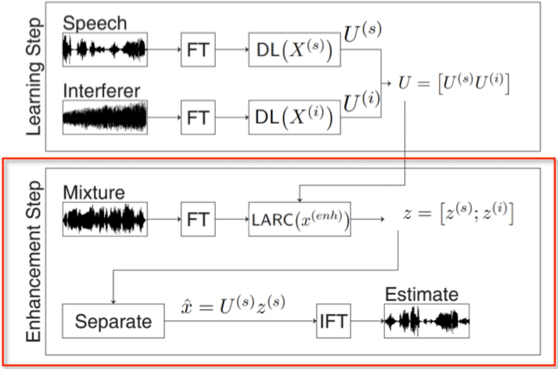
Dictionary learning is performed using the same K-SVD algorithm explained above.

$$\begin{aligned} (\mathbf{U}^\star, \mathbf{Z}^\star) &\in \arg\min_{\mathbf{UZ}} \|\mathbf{X} - \mathbf{U}\cdot\mathbf{Z}\|_F^2 \\ \text{s.t.} \left\|\mathbf{u}^\star_{(:,d)}\right\|_2 &= 1, \quad \text{for all} \quad d = 1, \ldots, L. \\ \|\mathbf{Z}^\star\|_0 &\leq K \end{aligned}$$

Learning of source models

- *Structured speech*: pre-train speech model on corpus
- *Variable interferer*: adapt interferer model in speech pauses

Enhancement Pipeline



- Sparse code mixture in composite dictionary by "*least angle regression with coherence criterion*" (LARC)
- Estimate speech: $\hat{\mathbf{x}} = \mathbf{U}^{(s)} \mathbf{z}^{(s)}$
- Apply inverse transformation (IFT) to map $\hat{\mathbf{x}}$ back to time-domain

Enhancement Step

Sparse coding of mixture $x = s + i$ in composite dictionary:

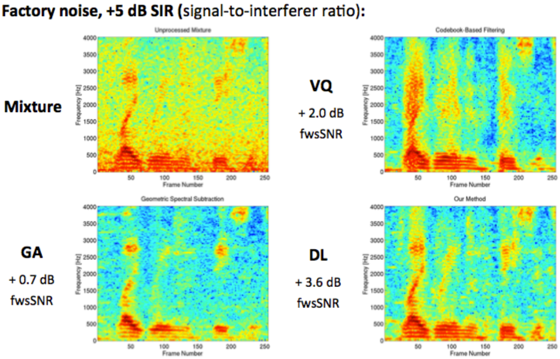
$$\left(\mathbf{z}^\star_{(s)}, \mathbf{z}^\star_{(i)}\right) \quad \in \arg \min_{\mathbf{z}^{(s)}\mathbf{z}^{(i)}} \left\| \mathbf{X} - \left[\mathbf{U}^{(s)}\mathbf{U}^{(i)}\right] \cdot \begin{bmatrix} \mathbf{z}^{(s)} \\ \mathbf{z}^{(i)} \end{bmatrix} \right\|_2$$

$$\text{s.t.} \quad \left\|\mathbf{z}^{(s)}\right\|_0 + \left\|\mathbf{z}^{(i)}\right\|_0 \leq K$$

The enhanced signal is reconstructed using only "*speech*" coefficients and the "*speech*" dictionary:

$$\hat{\mathbf{x}} = \mathbf{U}^\star{}_{(s)}\mathbf{z}^\star_{(s)}$$

Baseline comparison



C. D. Sigg, T. Dikk, JMB, IEEE Transactions Audio, Speech, and Language Processing, 20(6), 1698-1712, 2012

- *Objective measure*: Frequency Weighted Segmental SNR
- Baselines:
  - *GA*: Geometric spectral subtraction
  - *VQ*: Codebook based enhancement

Set-Top Box Application

Enhance sports commentary audio stream: