Prediction and Inverse Problems in Dynamical Systems

Joan Bruna, Pablo Sprechmann

September 30, 2014

1 Introduction

x is the input speech signal. We define $\Phi(x)$ a pooling analysis operator, which can be either the spectrogram, the scattering, or something that we learn. It can have several layers.

Then, given training examples $X = (x_i)_i$, we learn a synthesis dictionary D on $\Phi(X)$.

What is the role of the analysis operator Φ? Its role is to perform a contraction of each speaker's class, so that each source is embedded into a smooth manifold, which is then encoded using sparse coding. It needs to preserve discriminability between sources. We can also specialize the analysis operators for each class (TODO: why? what do we gain by that?).

We will restrict ourselves to analysis operators of the form $\Phi(x) = |Wx|$, where W is overcomplete, and close to a tight frame.

Variations of the model. There are several options for choosing the analysis Φ and several options for choosing the synthesis.

Estimation algorithm. Given $y = x_1 + x_2$, we construct estimates $\hat{x_1}$ and $\hat{x_2}$ by solving

$$\arg\min_{x_i', z_i} \|y - x_1' - x_2'\|^2 + \sum_{i=1,2} \|\Phi(x_i') - D_i z_i\|^2 + \lambda \|z_i\| \ . \tag{1}$$

A priori, this algorithm does not require computing the modulus/phase of the observation y. An initialization of this algorithm, in the case where $\Phi(x) = |Wx|$, assumes that

$$\Phi(y)_k^2 \approx \Phi(x_1)_k^2 + \Phi(x_2)_k^2 \ ,$$

and then estimates $\Phi(x_i)$ independently using each dictionary. Finally, $\hat{x}_i$ are obtained by using the phase of Wy.

In terms of the model, we need to decide which analysis operator to use, and how to train the dictionaries.

Alternative: Use a synthesis+pooling operator combined with analysis. In other words,

Definition of Φ. We have verified empirically that constant Q-transform yields better separation than STFT as the first layer. Now, we have several options:

- $\Phi_1(x) = Sc_1x$ (CQT (1 layer scattering).
- $\Phi_2(x) = F_1x$ (STFT)
- $\Phi_3(x) = Sc_2Sc_1x$ (2 layer scattering)
- $\Phi_4(x) = |A_2|Sc_1x$ (1 layer scattering followed by learnt scattering).
- $\Phi_5(x) = |S_2|Sc_1x$ (1 layer scattering followed by group lasso).

Φ_4 is more stable than Φ_5, but it may produce coefficients harder to separate in the final layer.

One layer vs two layers: A major difference is the temporal context. Two layer operators can extract temporal coherence and exploit temporal dynamics on a larger range than first layer operators.

Definition of the discriminative layer: We will stick to standard Dictionary learning.

There are several levels of supervision:

- Male vs Female: one dictionary for male and another for female.
- Unsupervised: A single "speech" dictionary.
- Speaker vs "rest of the world". We prioritize one speaker and consider the rest to be noise.

Planning:

1. Move to the setting with various speakers. The exhaustive test is 3 levels of supervision and 2 datasets (Timit and Grid). [P, 9/30]
2. In priority, we try: TIMIT denoising and TIMIT speaker separation (Male vs Female). [P, 9/30]
3. Test Φ_1 vs Φ_2 vs Φ_3. [J, 10/1-2]
4. Implement the estimation algorithm (1) [J, 10/1-2].
5. Optimize temporal context. This is governed by Φ. [P, 10/1-2].
6. Implement the learning of analysis operator $A(x) = |Wx|$. Perhaps start with the algorithm from paper ??. What criteria??
7. If we have time, compare also Φ_4.
8. Maths: explain why analysis is a better idea than synthesis for Φ.
9. Writeup.
10. Discriminative fine-tuning (probably out of time).

2 Previous work

3 Spatio-temporal Pooling and Inhibition

4 Unsupervised Learning

We describe first the case with two known speakers. We train two dictionaries $\mathcal{D}_i$, $i = 1, 2$ using the following model:

$$\|x - \mathcal{D}z\|^2 + \phi_{\mathcal{G}}(z) \tag{2}$$

with $\phi_{\mathcal{G}}$ being a spatio-temporal group lasso. Denoting $z^p = \phi_{\mathcal{G}}(z)$, the second layer is

$$\|z^p - \widetilde{\mathcal{D}}\tilde{z}\|^2 + \phi_{\mathcal{G}}(\tilde{z}) \tag{3}$$

5 Discriminative Training with Bi-level

5.1 Inference

Given $x = x_1 + x_2$, we want to infer x_1 and x_2. We use the following scheme:

$$(z_1^*, z_2^*) = arg\min \|x - [\mathcal{D}_1\mathcal{D}_2][z_1; z_2]\|^2 + \phi_{\mathcal{G}}(z_1) + \phi_{\mathcal{G}}(z_2) \tag{4}$$

$$(\tilde{z}_1^*, \tilde{z}_2^*) = arg\min \| - [\mathcal{D}_1\mathcal{D}_2][z_1; z_2]\|^2 + \phi_{\mathcal{G}}(z_1) + \phi_{\mathcal{G}}(z_2) \tag{5}$$

6 Experimental Results