

Assignment 03

CMPE 58P, Machine Listening

Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey

Instructor: A. T. Cemgil

Due: 15 May 2009, Mon, 13:00.

1. (10) (Transform domain Source separation)

Download three clips (piano, guitar, speech) from the course website.

- Compute the MDCT (Modified Discrete Cosine Transform) of each clip and display the transform coefficients (In matlab, you can use the `imagesc` command of matlab).
- Plot the histogram of the transform coefficients. Compare it with the histogram of the time samples.
- Fit the following models to the transform coefficients s_k where $k = (\nu, \tau)$:

- Model 1

$$s_k \sim \mathcal{N}(s_k; 0, v)$$

- Model 2

$$s_k \sim \mathcal{N}(s_k; 0, v_k)$$

$$v_k \sim \mathcal{IG}(v_k; a, b)$$

Here IG is an inverse gamma distribution. Show in model 2 that the marginal $p(s_k)$ is a T-distribution. Plot both marginals on top of the histogram and comment about the quality of the fit.

- For each clip, sort the coefficients in decreasing magnitude and retain only the top 20 percent. Reconstruct the audio by setting other coefficients to zero using IMDCT. What happens perceptually when you keep less and less coefficients?
- Construct a 2×2 mixing matrix

$$A = \begin{pmatrix} 1 & 1 \\ \tan(\theta_1) & \tan(\theta_2) \end{pmatrix}$$

Assuming that each column is a basis vector, plot the subspaces spanned by each vector. Prepare a two channel mixture via

$$\tilde{X} = A\tilde{S}$$

where \tilde{S} is a 2×2^{16} matrix with time samples of each clip in each row $S(m, :)$. Compute the MDCT of $\tilde{X}(m, :)$. Call it X . Plot $X(:, k)$ for all k on the 2-D plot. Change the mixing directions θ and experiment with different θ .

- (f) Now assume A is unknown. Devise an algorithm to estimate A . How would you modify your algorithm that uses only the top 20 percent (in magnitude) samples of X .
- (g) Consider now a new mixing matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \tan(\theta_1) & \tan(\theta_2) & \tan(\theta_3) \end{pmatrix}$$

2. (10) (Single Channel Source separation)

The goal of this exercise is to investigate the principles behind single channel source separation. Consider the following model, where $k = (\nu, \tau)$ denotes the time-frequency index of the transform coefficients and $i = 1 \dots I$ denotes the source index.

$$\begin{aligned} s_{k,i} &\sim \mathcal{N}(s_k; 0, v_{k,i}) \\ x_k &= \epsilon_k + \sum_i s_{k,i} \\ \epsilon_k &\sim \mathcal{N}(0, R) \end{aligned}$$

- (a) Derive $p(s_{k,i}|x_k, v_{k,1:I})$. What happens when $R \rightarrow 0$?
- (b) For the clips $i = 1 \dots 2$ on the webpage, compute the transform coefficients $s_{k,i}$ via MDCT and compute ML estimates of $v_{k,i}$. Then, try to reconstruct the sources using the expected value of $\langle s_{k,i} \rangle$ under $p(s_{k,i}|x_k, v_{k,1:I})$. Compute the Signal to Noise (SNR) ratio defined as

$$\text{SNR}(s, r) = 20 \log_{10} \frac{\|s\|}{\|s - r\|}$$

where s is the original signal and r is the reconstruction. Also, listen to the reconstructed sources and comment.

- (c) In the above exercise, we assumed that the variances were estimated with perfect knowledge. Now, assume that the variances are tied for each frequency band ν , i.e., for each source, they obey the following model:

$$v_{(\nu,\tau),i} = v_{\tau,i}$$

Estimate these variances from the isolated clips and compute the SNR for this case. How does it degrade and what are the artifacts?

- (d) (Optional but recommended) Is there a better way of choosing the $v_{k,i}$ than the ML estimate given the sources ? (Better in the sense of obtaining a higher SNR reconstruction.) One way to investigate this would be to find the optimum variances at each frequency atom and compare it to the ML estimate. Then, one could try to find a link between the ML estimate and the optimum variances.