Sound Source localization using Diffusion Kernels and Manifold-Based Bayesian Inference

EE602 Term Project

Hrusikesha Pradhan
Vikram Singh
Jyotsana Kumari
Kalpesh Saubhri

Department of Elelctrical Engineering, IIT Kanpur

5th November, 2016

Outline

Introduction

- Sound source localization is of great use in a large variety of practical applications like video conferencing, automatic camera steering and speaker separation. In our work we will be mostly discussing about how the localization done in two research papers.

Supervised Source Localization Using Diffusion Kernels

- In this paper it is assumed that the position of the source is conveyed by a single acoustic impulse response between the source and microphone.
- Prior recordings of signals from various known locations in the room are used for training and calibration.
- Let $h_{\Theta}(n)$ be the acoustic impulse response between the microphone and a source, at relative position $\Theta = \begin{bmatrix} \phi, \theta, \rho \end{bmatrix}$
- For generating the training data, we pick m predefined position of the source $= \{\bar{\theta_1}, \ldots, \bar{\theta_m}\}$.
- From each position an arbitary stationary unknown input signal of finite length is played and recorded after picked up by the microphone.
- The received signal is expressed as $\bar{y}_i(n) = h_{\bar{\theta}_i(n)} * x_i(n)$

- The measurement is repeated from each source location L times taking into account the perturbation of the source position.
- Let $\{x_{i_j}(n), y_{i_j}(n)\}_{j=1}^{L}$ be the input and output signals corresponding to the repeated measurements.
- The goal in this is to recover the source position given the measured signal based on the prior training information.
- Let$\Theta = \{\theta_1, \ldots, \theta_M\}$ denote the unknown M source positions corresponding to the new measurements.
- The signal can be expressed as

$$y_i(n) = h_{\theta_i}(n) * x_i(n) \tag{1}$$

where $x_i(n)$ and $y_i(n)$ are input and output signals of finite length.

- The covariance of the output signal $y_i(n)$ is given by

$$c_{y_i}(\tau) = h_{\theta_i}(\tau) * h_{\theta_i}(-\tau) * c_{x_i}(\tau) \tag{2}$$

  where $c_{x_i}(\tau)$ and $c_{y_i}(\tau)$ denote the covariance functions of $x_i(n)$ and $y_i(n)$.
- Let c_i, $\bar{c}_i$ and c_{i_j} denote the covariance elements of $y_i(n)$, $\bar{y}_i(n)$ and $y_{i_j}(n)$.
- The local covariance matrix of $\bar{c}_i$ is

$$\hat{\Sigma}_i = \frac{1}{L} \sum_{j=1}^{L} c_{i_j} c_{i_j}^T \tag{3}$$

- Affinity matrix W is computed between the m training samples in $\bar{\Theta}$ and the matrix klth element is calculated according to

$$W_{kl} = \frac{\pi}{d_{kl}} exp\left\{ - \frac{(\bar{c}_k - \bar{c}_l)^T \left[\hat{\Sigma}_k + \hat{\Sigma}_l\right]^{-1} (\bar{c}_k - \bar{c}_l}{\varepsilon} \right\} \tag{4}$$

  where ε is the kernel scale and d_{kl} is the normalization factor.

- The distance measure used above approximates the euclidean distance between the parameters,i.e.,

$$\|\bar{\theta}_k - \bar{\theta}_l\|^2 \approx (\bar{c}_k - \bar{c}_l)^T \left[\hat{\Sigma}_k + \hat{\Sigma}_l\right]^{-1} (\bar{c}_k - \bar{c}_l) \tag{5}$$

- This proposed kernel enables to capture the actual variability in terms of the source position based on the measurements.

- Now using a set of M new sequential measurements we compute a matrix A of dimension M by m containing the corresponding covariance elements $\{c_i\}_{i=1}^{M}$ and is given by

$$A_{kl} = exp\left\{ - \frac{(c_k - \bar{c}_l)^T \left[\hat{\Sigma}_l\right]^{-1} (c_k - \bar{c}_l)}{\varepsilon} \right\} \tag{6}$$

- Let $\tilde{A} = AS^{-1/2}$, where $S = diag\{A^T A1\}$ is a diagonal matrix and the normalized matrix satisfies $W = \tilde{A}^T \tilde{A}$
- The eigen vectors of W of length m are the left singular vectors of $\tilde{A}$ and are assumed to describe the m training measurements.
- The right singular vectors of $\tilde{A}$ of length M are given by

$$\psi_j = \frac{1}{\sqrt{\lambda_j}} \tilde{A} \varphi_j \tag{7}$$

- The right singular vectors of $\tilde{A}$ can be viewed as the extension of the spectral representation describing the new M measurements.
- Let Ψ be the embedding of the measurements onto the space spanned by the right singular vectors corresponding to the source position, i.e.,

$$\Psi : c_i \mapsto \left[\psi_1^i, \psi_2^i, \psi_3^i\right] \tag{8}$$

- Let N_i consist of the k-nearest training measurements $\{\bar{c}_j\}$ of c_i in the embedded space.
- Let $\{\gamma_j\}_{j=1}^{k}$ be interpolation coefficients, given by

$$\gamma_j(c_i) = \frac{exp\left(-\|\Psi\left(c_i\right) - \Psi\left(\bar{c}_j\right)\|^2/\varepsilon_{\gamma_j}\right)}{\sum_{\bar{c}_k \in N_i} exp\left(-\|\Psi\left(c_i\right) - \Psi\left(\bar{c}_k\right)\|^2/\varepsilon_{\gamma_j}\right)} \tag{9}$$

- The estimate of the source position is given by the following weighted sum of training locations

$$\hat{\theta}_i = \sum_{\bar{c}_j \in N_i} \gamma_j(c_i)\bar{\theta}_i \tag{10}$$

- The estimation error is now defined by,

$$e(c_i) = \|\theta_i - \hat{\theta}_i\| \tag{11}$$

Manifold-Based Bayesian Inference for Semi-Supervised Source Localization

- The main goal of this paper is to estimate the target function which receives an acoustic sample and returns its corresponding location.
- The target function is estimated in this work using a Bayesian inference framework which involves a likelihood function and a prior probability.
- The source is emitting an unknown signal $s(n)$ which is measured by a pair of microphones.
- The noisy measurements $x(n)$ and $y(n)$ are given by a convolution between the clean source signal and the corresponding AIR, contaminated by stationary noise signals,

$$\begin{aligned} x(n) &= a_1(n,p) * s(n) + u_1(n) \\ y(n) &= a_2(n,p) * s(n) + u_2(n) \end{aligned} \tag{12}$$

- Since the acoustic transfer functions are unavailable, the RTF is estimated as,

$$\hat{H}_{yx}(k,p) = \frac{\hat{S}_{yx}(k,p)}{\hat{S}_{xx}(k,p)} \simeq \frac{A_2(k,p)}{A_1(k,p)} \tag{13}$$

- The feature vector is defined as

$$h(p) = \left[\hat{H}_{yx}(0,p), \ldots, \hat{H}_{yx}(D-1,p)\right]^T \tag{14}$$

- Lets assume we have a training set H_L consisting of l labelled RTF samples and H_u consisiting of u unlabelled RTF samples from unknown locations.
- Our aim is to estimate the locations corresponding to a test set of q pairs of measurement of unknown sources from unknown locations.
- The position of the source is a random variable obtained as an output of the target function that receives the RTF sample as an input.

- The target function can be estimated based on the following posterior probability, given by bayes rule,

$$p(f|P_L, H_L, H_U) \propto p(P_L|f, H_L)P(f|H_L, H_U) \tag{15}$$

- It is assumed that the measured positions $P_L = \{p(h_i)\}_{i=1}^{l}$ follow a noisy observation model, given by:

$$p(h_i) = f(h_i) + \eta_i, i = 1, \ldots, l \tag{16}$$

for $\eta_i \sim \mathbf{N}(0, \sigma^2)$,$i = 1, \ldots, l$ are iid gaussian noises independent of f.

- The prior of the function is assumed to follow Gaussian process:

$$f \sim \mathcal{GP}(v, k) \tag{17}$$

where v is the mean function and k is the covariance function.

- The mean function is taken zero and the function k is a pairwise function that evaluates the covariance of each pair of samples drawn from the process f.
- The covariance between $f(h_i)$ and $f(h_j)$, given by $k(h_i, h_j)$ is,

$$k(h_i, h_j) = exp - \|h_i - h_j\|^2/\varepsilon_k \tag{18}$$

- Instead of deriving a general estimator of the function f we estimate the function value at some specific test point h_t.
- The function at the test point $f(h_t)$ and the concatenation of all labelled training positions p_L are jointly Gaussian.
- This implies that the conditional distribution $p(f(h_t)|P_L, H_L)$ is a multivariate Gaussian and the MAP estimator of $f(h_t)$ which coincides with the MMSE estimator in the gaussian case is given by:

$$\hat{f}(h_t) = \mu_{cond} = \Sigma_{Lt}^{T}(\Sigma_{LL} + \sigma^2 I_l)^{-1} p_L \tag{19}$$

- Now we will use both labelled and unlabelled data to calculate the new prior which is a gaussian process with a modified kernel function.
- We form a discrete representation of the manifold by a graph defined over the entire training set H_D.
- The graph nodes are the training samples and the weights of the edges constituting an affinity matrix W
- Let G denote an abstract collection of random variables that represent the geometry structure of the manifold.
- The likelihood of the geometry variables G is defined by,

$$P(G|f_D) \propto exp\{ -\frac{\gamma_M}{2}(f_D^T M f_D)\} \tag{20}$$

- M is the graph laplacian given by $M = S - W$, where S is a diagonal matrix given by $S_{ii} = \sum_{j=1}^{n} W_{ij}$.
- The likelihood function is a measure of correspondence between the values of the target function f and the structure of the manifold.
- In order for the model to be extendible to additional test data H_T, we make the assumption that given f_D, the geometry variables are independent of the function values in other points,i.e. $p(G|f_H) = p(G|f_D)$.
- $$\tilde{k}(h_i, h_j) = k(h_i, h_j) - \gamma_M \sum_{D_i}^{T} (I_n + \gamma_M M \Sigma_{DD})^{-1} M \Sigma_{Dj} \qquad (21)$$

  this $\tilde{k}$ is termed as manifold based kernel.
- Based on this data driven prior, an alternative estimator for $f(h_t)$ is given by:
$$\hat{f}(h_t) = \tilde{\Sigma_{Lt}^T} (\tilde{\Sigma}_{LL} + \sigma^2 I_l)^{-1} p_L \qquad (22)$$