Unsupervised Acoustic Scene Mapping Based on Acoustic Features and Dimensionality Reduction

Idan Cohen   Ofir Lindenbaum   Sharon Gannot

Faculty of Engineering, Bar-Ilan University, Israel

ICASSP 2024, Seoul

Acoustic Scene Mapping

Acoustic Scene Mapping

Applications

- Augmented Reality
- Robot Autonomy

Goal

Allow a visually blind device to reconstruct a mapping of a region of interest inside an enclosure, using only acoustic data

Acoustic Scene Mapping

Applications

- Augmented Reality
- Robot Autonomy

Goal

Allow a visually blind device to reconstruct a mapping of a region of interest inside an enclosure, using only acoustic data

Naïve Solution

Classical Solution for a Simpler Problem

Assume that the **source position is known**:

- Naïve solutions are based on Time difference of arrival (TDOA) estimation and geometrical considerations

Naïve Solution

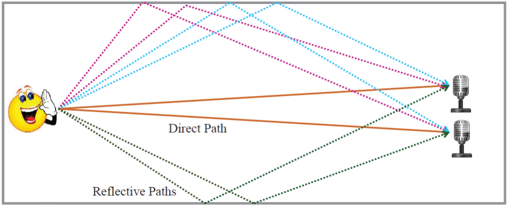
Classical Solution for a Simpler Problem

Assume that the **source position is known**:

- Naïve solutions are based on Time difference of arrival (TDOA) estimation and geometrical considerations

In Reverberated Settings

- Numerous propagation paths (multipath)
- Classical TDOA estimation are inaccurate
- Resultant mappings are unreliable!
- Better use the entire reflection pattern



Our Approach

Acoustic Features Lying on Manifolds

Data-Driven Multi-Microphone Speaker Localization on Manifolds *(Laufer, Talmon and Gannot, 2020)*

Latent Manifold Learning

Local conformal autoencoder for standardized data coordinates (*Peterfound & Lindenbaum et al., 2020*)

The Relative Transfer Function (RTF)

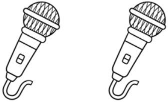
The Relative Transfer Function (RTF)

Two Microphones RTF - Formulation

The measured signals in the two microphones:

$$d_1(n) = \{a_1 * s\}(n) + u_1(n)$$
$$d_2(n) = \{a_2 * s\}(n) + u_2(n)$$

- $s(n)$ - the source signal
- $a_i(n)$, $i = \{1, 2\}$ - the acoustic impulse response relating the source and each of the microphones,
- $u_i(n)$ - noise signals, independent of the source



The Relative Transfer Function (RTF)

RTF Definition

$$H(k) = \frac{A_2(k)}{A_1(k)}$$

where $A_i(k)$ are the transfer functions - the Fourier transform of the acoustic impulse responses $a_i(n)$, $i = \{1, 2\}$

The Relative Transfer Function (RTF)

RTF Definition

$$H(k) = \frac{A_2(k)}{A_1(k)}$$

where $A_i(k)$ are the transfer functions - the Fourier transform of the acoustic impulse responses $a_i(n)$, $i = \{1, 2\}$

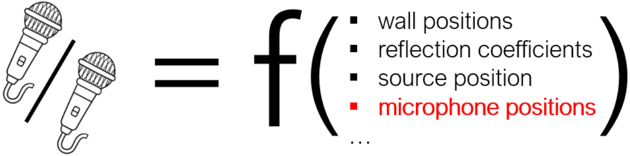
RTF Estimation

Assuming negligible additive noise, we can estimate the RTF using:

$$\hat{H}(k) = \frac{\hat{S}_{d_2 d_1}(k)}{\hat{S}_{d_1 d_1}(k)} \approx H(k)$$

where $\hat{S}_{d_1 d_1}(k)$ and $\hat{S}_{d_2 d_1}(k)$ are the PSD and CPSD respectively

The Relative Transfer Function (RTF) [Laufer-Goldshtein et al., 2015]



Properties of the RTF

- Independent of the source signal
- Function of the acoustic parameters
- Holds the entire reflection pattern
- Changes only as a function of the microphone positions
- **RTFs lie on a low dimensional manifold**

Local Conformal Autoencoder (LOCA)

Local Conformal Autoencoder (LOCA)

Assumptions

- Latent Manifold: $\mathcal{X} \subset \mathbb{R}^d$
- Observable/Measurement space: $\mathcal{Y} \subset \mathbb{R}^D$, where $d \ll D$
- Observed data samples: $\boldsymbol{y}_i = f(\boldsymbol{x}_i)$, for $i = 1, \ldots, N$, where $f : \mathcal{X} \to \mathcal{Y}$ is a smooth, nonlinear and bijective function

Local Conformal Autoencoder (LOCA)

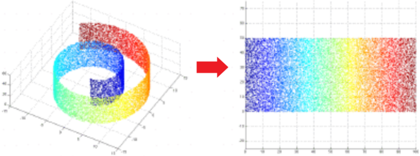
Assumptions

- Latent Manifold: $\mathcal{X} \subset \mathbb{R}^d$
- Observable/Measurement space: $\mathcal{Y} \subset \mathbb{R}^D$, where $d \ll D$
- Observed data samples: $\boldsymbol{y}_i = f(\boldsymbol{x}_i)$, for $i = 1, \ldots, N$, where $f : \mathcal{X} \to \mathcal{Y}$ is a smooth, nonlinear and bijective function

Goal

Finding $f^{-1} : \mathcal{Y} \to \mathcal{X}$ and reconstructing the latent domain using

$$\boldsymbol{x}_i = f^{-1}(\boldsymbol{y}_i)$$



Local Conformal Autoencoder (LOCA)

Problem

Finding $f^{-1} : \mathcal{Y} \rightarrow \mathcal{X}$ is infeasible - we have no access to $\mathcal{X}$!

Local Conformal Autoencoder (LOCA)

Problem

Finding $f^{-1} : \mathcal{Y} \to \mathcal{X}$ is infeasible - we have no access to $\mathcal{X}$!

Solution

Learn an embedding $\rho : \mathbb{R}^D \to \mathbb{R}^d$ that maps the observations $\{\boldsymbol{y}_i\} \in \mathcal{Y}$ such that the distances on the latent manifold are conserved:
$\|\rho(\boldsymbol{y}_i) - \rho(\boldsymbol{y}_j)\|_2 = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$ for any i, j.

Local Conformal Autoencoder (LOCA)

Problem

Finding $f^{-1} : \mathcal{Y} \to \mathcal{X}$ is infeasible - we have no access to $\mathcal{X}$!

Solution

Learn an embedding $\rho : \mathbb{R}^D \to \mathbb{R}^d$ that maps the observations $\{\boldsymbol{y}_i\} \in \mathcal{Y}$ such that the distances on the latent manifold are conserved:
$\|\rho(\boldsymbol{y}_i) - \rho(\boldsymbol{y}_j)\|_2 = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$ for any i, j.

Result - Whitening Requirement

- **Burst**: $\boldsymbol{Y}_i = \{\boldsymbol{y}_i^{(m)}\}_{m=1}^M$ - a set of M samples gained from a local neighbourhood, with $i = 1, \ldots, N$ being the burst index
- The embedding ρ should satisfy:

$$\frac{1}{\sigma^2}\boldsymbol{C}(\rho(\boldsymbol{Y}_i)) = \boldsymbol{I}$$

Local Conformal Autoencoder (LOCA)

Loss Terms

$$L_{\text{white}}(\rho) = \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sigma^2}\hat{\boldsymbol{C}}(\rho(\boldsymbol{Y}_i)) - \boldsymbol{I}_d\right\|_F^2$$

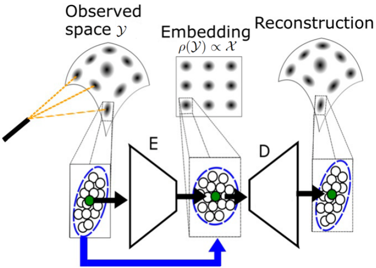
Local Conformal Autoencoder (LOCA)

Loss Terms

$$L_{\text{white}}(\rho) = \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sigma^2}\hat{\boldsymbol{C}}(\rho(\boldsymbol{Y}_i)) - \boldsymbol{I}_d\right\|_F^2$$

$$L_{\text{rec}}(\rho,\gamma) = \frac{1}{N\cdot M}\sum_{i,m=1}^{N,M}\left\|\boldsymbol{y}_i^{(m)} - \gamma(\rho(\boldsymbol{y}_i^{(m)}))\right\|_2^2$$

$$L(\rho,\gamma) = L_{\text{white}}(\rho) + L_{\text{rec}}(\rho,\gamma)$$



- Encoder E learns an embedding ρ that minimizes the whitening loss
- Decoder D learns an inverse embedding γ, to make sure that ρ is invertible
- The learned embedding is the low dimensional manifold of interest

Our Approach

Acoustic Features Lying on Manifolds

Data-Driven Multi-Microphone Speaker Localization on Manifolds (Laufer, Talmon and Gannot, 2020)

Latent Manifold Learning

Local conformal autoencoder for standardized data coordinates (Peterfound & Lindenbaum et al., 2020)

Our Approach

Acoustic Features Lying on Manifolds

Data-Driven Multi-Microphone Speaker Localization on Manifolds (Laufer, Talmon and Gannot, 2020)

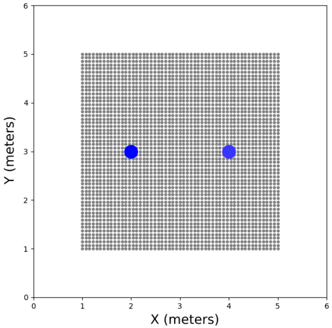
Latent Manifold Learning

Local conformal autoencoder for standardized data coordinates (Peterfound & Lindenbaum et al., 2020)

Synthesis is Clear

- RTFs - points on a low dimensional manifold represented in a high dimensional space
- The manifold is only governed by the position of the microphones
- LOCA enables reconstructing a low dimensional latent manifold sampled in high dimensional space

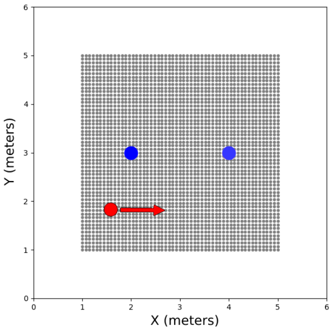
Method Description



Assumptions

- Single/Multiple fixed sound sources
- Sources are not concurrently active

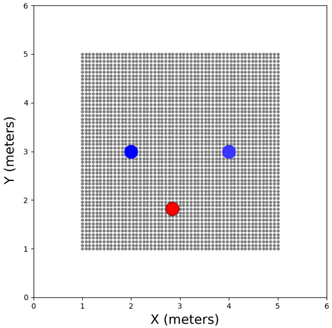
Method Description



Gathering Data

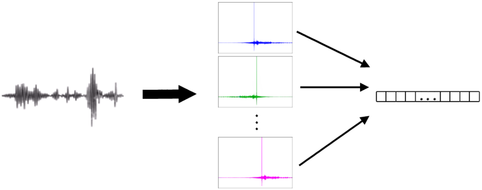
1. A device carrying microphone array travels in the enclosure

Method Description

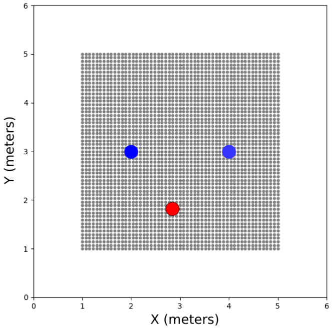


Gathering Data

1. A device carrying microphone array travels in the enclosure
2. The device stops every once in a while
3. The device records the sound signals produced by the sources
4. RTFs are estimated for each of the sources
5. RTFs are concatenated into a single vector to create data samples

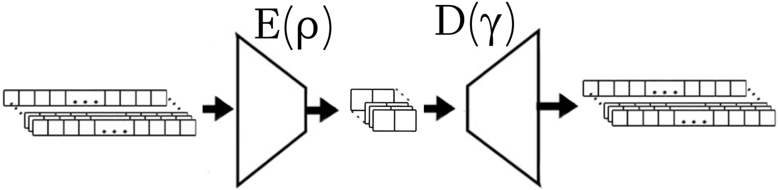


Method Description

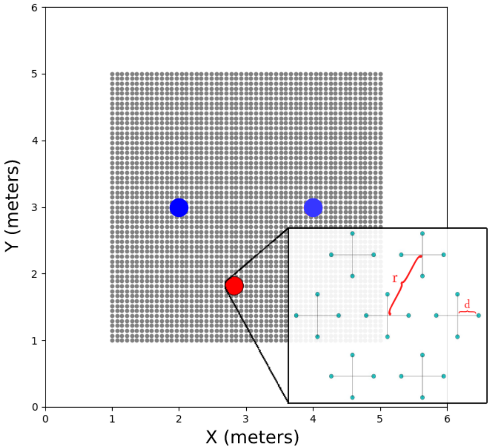


Training

Feed LOCA with bursts to learn a 2D embedding

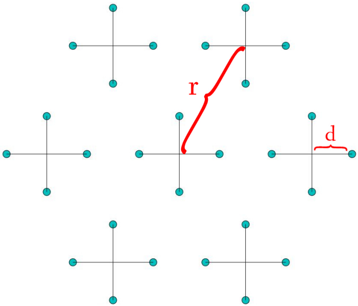


Burst Sampling Strategy



- Gray points– sampling grid of the allowed region

Burst Configuration and Input Feature

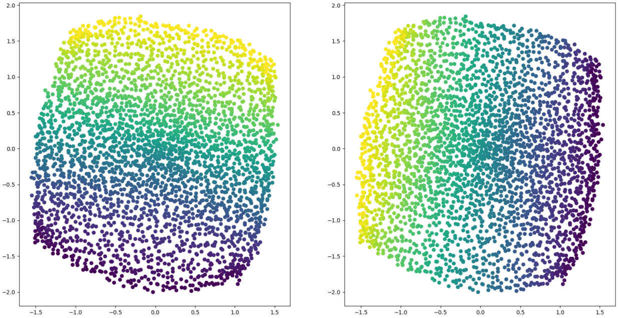


Parameters

- $r = 2$ cm
- $d = 3$ cm
- Frequency bins - 5 to 99 (312.5–6190 Hz)

Source 1				Source 2			
Horizontal Real	Horizontal Imaginary	Vertical Real	Vertical Imaginary	Horizontal Real	Horizontal Imaginary	Vertical Real	Vertical Imaginary

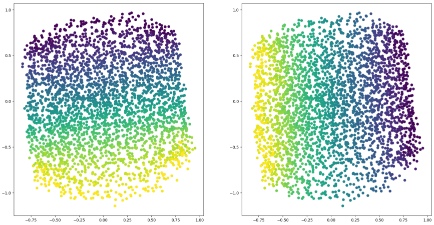
Results - Low Reverberation



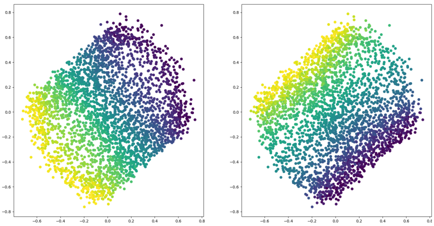
(a) $RT_{60} = 160$ ms

- Same embedding - colored twice to show the correlation with the true positions along x,y axes

Results - Comparing Higher Reverberation Times

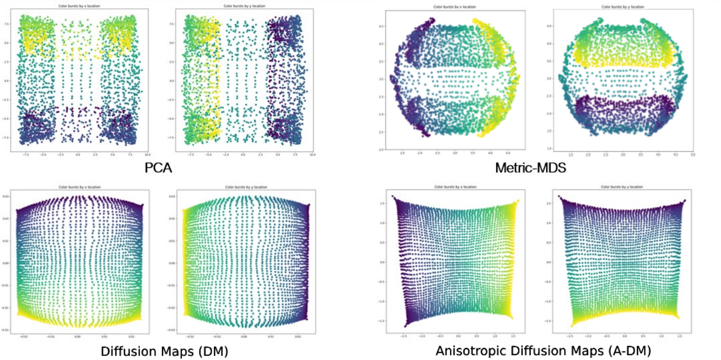


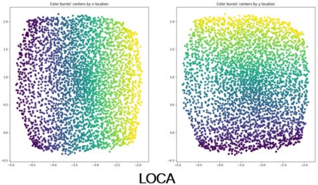
(b) $RT_{60} = 360$ ms (Office)



(c) $RT_{60} = 610$ ms (Lecture Hall)

Comparing Manifold Learning Methods ($RT_{60} = 160$ ms)

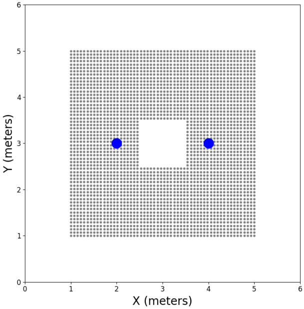




Method	160 ms	360 ms	610 ms
PCA	72.2	72.5	74.3
MMDS	73.4	79.1	82.5
DM	20.4	22.7	65.1
A-DM	17.8	18.3	33.8
LOCA	**11.3**	**13.4**	**18.5**

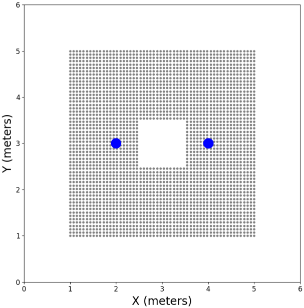
Table 1: MAE (cm) comparison of the embeddings gained by each method.

Generalization to Unseen Samples

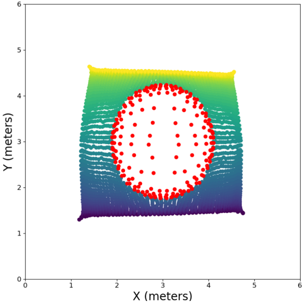


(a) Training region

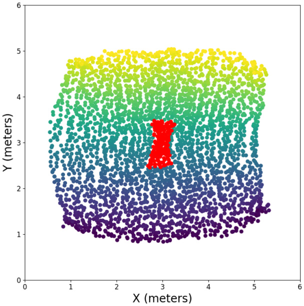
Generalization to Unseen Samples



(a) Training region



(b) Anisotropic Diffusion Maps (A-DM)



(c) LOCA

A-DM	LOCA
16.1 cm	67.4 cm

Table: Reconstruction MAE of the samples from the unfamiliar region.

Conclusions

Summary

- RTFs lie on low dimensional manifold in high dimensional space
- The manifold is controlled by the microphone positions
- LOCA - uncovers the latent manifold
- Handles reverberation better than classical methods

Conclusions

Summary

- RTFs lie on low dimensional manifold in high dimensional space
- The manifold is controlled by the microphone positions
- LOCA - uncovers the latent manifold
- Handles reverberation better than classical methods

Thank you for listening!