Aalto University
School of Electrical
Engineering

Time-Frequency Audio Similarity using Optimal Transport

Linda Fabiani[1], Filip Elvander[1], Sebastian J. Schlecht[2]

[1] *Structured and Stochastic modeling group, Aalto University*
[2] *Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nurnberg*

ASILOMAR Conference on Signals, Systems, and Computers, 29.10.2024

Topics

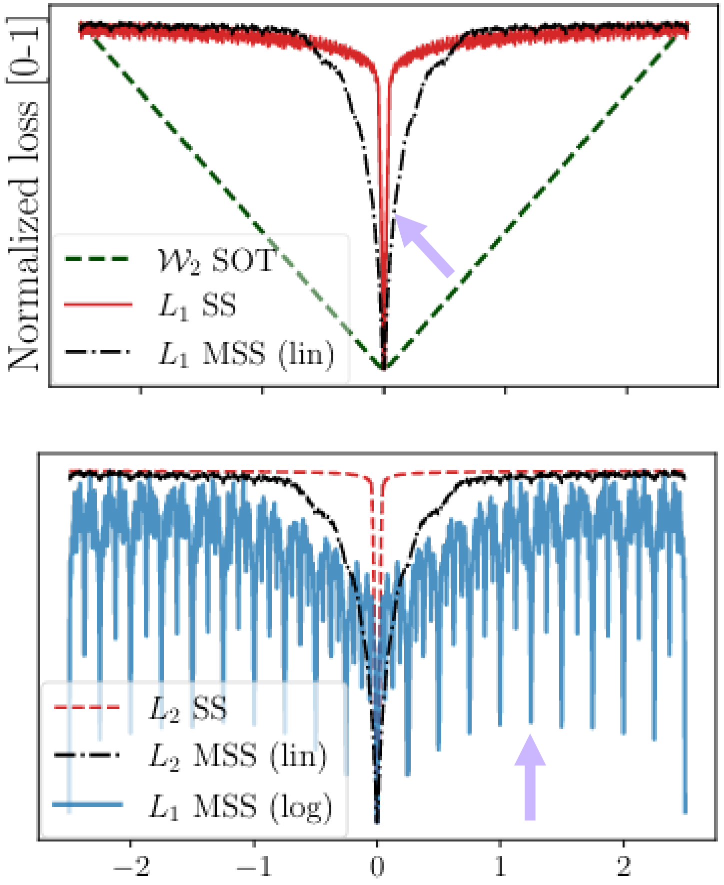
1. Problem motivation
2. Background on optimal transport
3. Optimal transport for audio
4. Applications and results
5. Conclusions and future works

1. Problem motivation

Typical audio-to-audio losses (l_1, l_2, Multi-scale Spectral loss) show limited performances when used in neural audio synthesis tasks

↓

- ✕ **rapid growth** for small shifts and **saturate** as soon as the compared signals no longer have shared time-frequency support
- ✕ Presence of **local minima** that can impact the gradients propagation



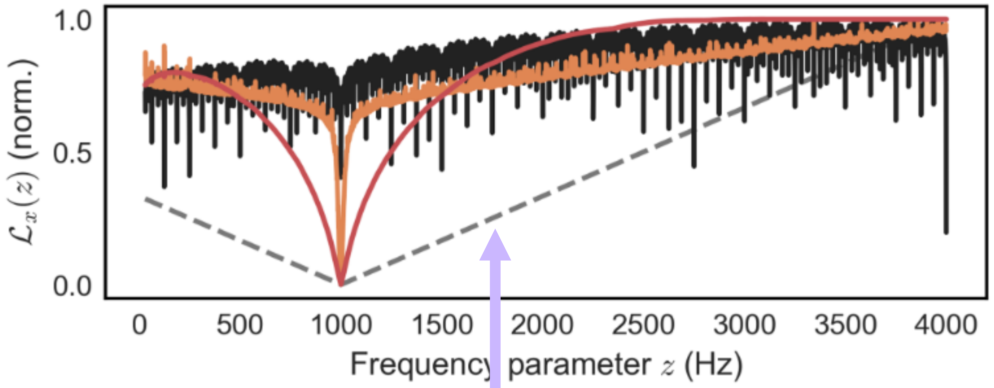
S. Schwar and **M. Muller**, “Multi-scale spectral loss revisited,” *IEEE Signal Processing Letters*, vol. 30, pp. 1712–1716, 2023
J. Turian and **M. Henry**, “I’m sorry for your loss: Spectrally-based audio distances are bad at pitch,” *CoRR*, vol. abs/2012.04572, 2020

1. Problem motivation

Typical audio-to-audio losses (l_1, l_2, Multi-scale Spectral loss) show limited performances when used in neural audio synthesis tasks

- ✓ New loss function for comparing audio signals in **time** and **frequency simultaneously** to produce clean and informative gradients

- ✓ Use optimal transport to ensure robustness with respect to the **geometric space** of the signals' spectral power



Ideal linear behavior

S. Schwar and **M. Muller**, “Multi-scale spectral loss revisited,” *IEEE Signal Processing Letters*, vol. 30, pp. 1712–1716, 2023
J. Turian and **M. Henry**, “I’m sorry for your loss: Spectrally-based audio distances are bad at pitch,” *CoRR*, vol. abs/2012.04572, 2020

2. Background on optimal transport

Problem: find the most efficient way to "move" mass between two probability distributions while minimizing a given cost

$\mathbf{a} \in \mathbb{R}^{\mathbf{N}}$

$\mathbf{b} \in \mathbb{R}^{\mathbf{M}}$

→ Two continuous or discrete distributions (histograms)

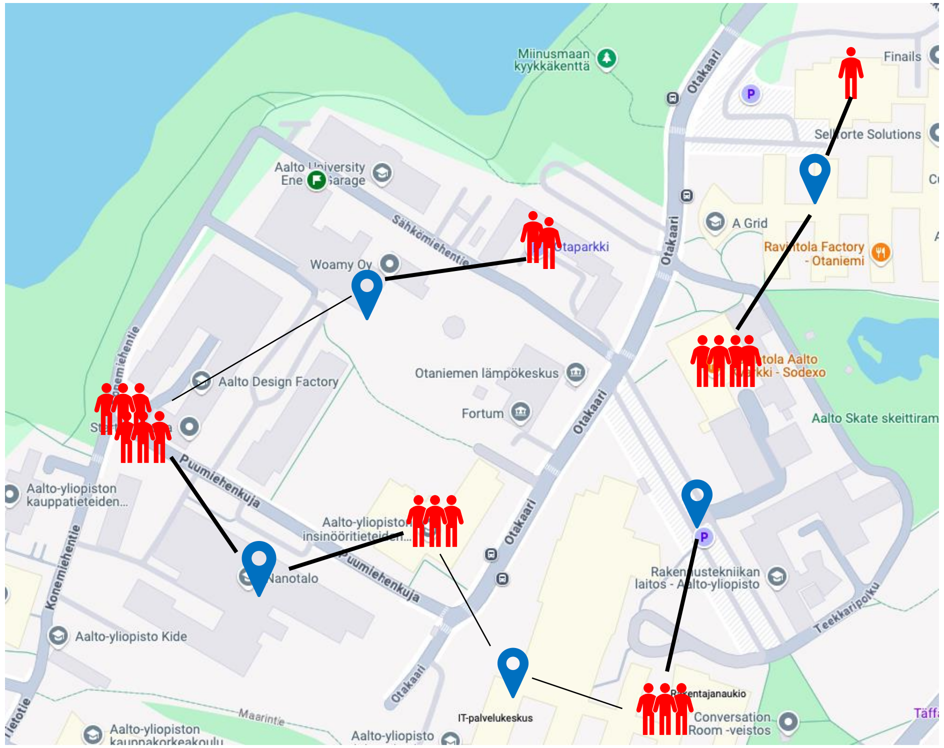
$\{x_1, x_2, \ldots, x_N\}$

$\{y_1, y_2, \ldots, y_M\}$

→ Points in the spaces of distributions **a** and **b**

$$\mathrm{C} \in \mathbb{R}^{N \times M} \longrightarrow c_{ij} = c(x_i, y_j)$$

Cost of transport



: people to allocate

: locations with limited capacity

2. Background on optimal transport

$$\underset{\mathbf{M}\geq 0}{\text{minimize}} \quad \langle \mathbf{C}, \mathbf{M} \rangle$$

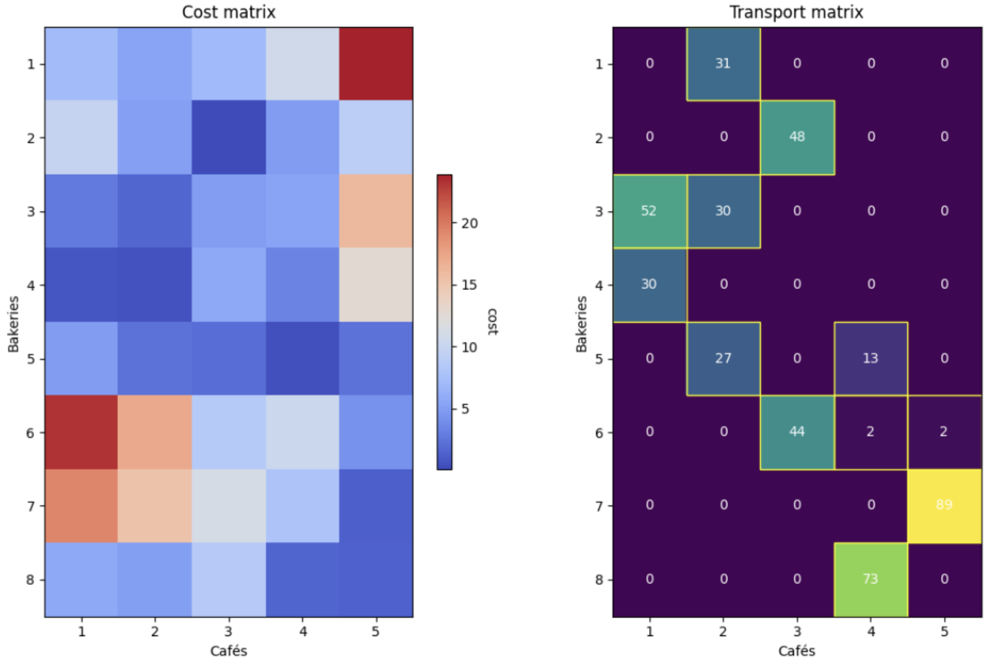
$$\text{s.t.} \quad \underbrace{\mathbf{M}\mathbf{1}_N = \mathbf{a}\,,\ \mathbf{M}^T\mathbf{1}_M = \mathbf{b}.}_{\text{mass conservation constraints}}$$

$$\mathbf{U}(\mathbf{a},\mathbf{b}) = \{\mathbf{M} \in \mathbb{R}_+^{NxN} : \quad \mathbf{M}\mathbf{1}_N = \mathbf{a}\,,\ \mathbf{M}^T\mathbf{1}_N = \mathbf{b}\}$$

$\mathbf{M}^\star$ = optimal transport **plan** out of all possible pairings between **a** and **b**

Convex objective function and constraints → solve as a **linear program** (Sinkhorn, entropy regularized)

$\mathbf{C}$ $\qquad$ $\mathbf{M}$



$$\boxed{\Psi(\mathbf{a},\mathbf{b}) \triangleq \langle \mathbf{C}, \mathbf{M}^\star \rangle}$$

G. Peyre and **M. Cuturi**,“Computational optimal transport,” 2020.
M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” *Advances in neural information processing systems*, vol. 26, pp. 2292–2300, 2013

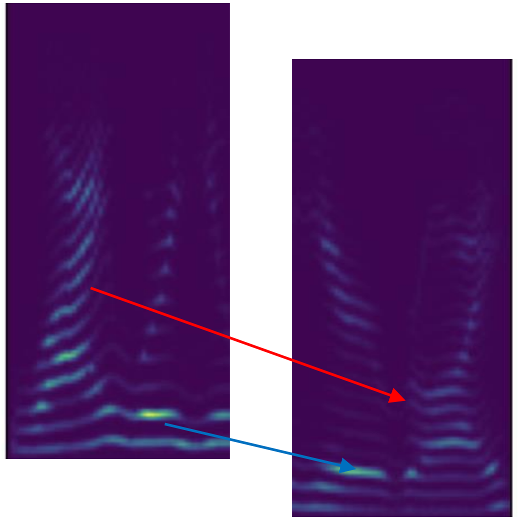
3. Optimal transport for spectrograms:

Sinkhorn-based approach

Mass: Normalized spectral energy

Cost: between pairs of **time** and **frequency bins**

$$c\left((t_1,\omega_1),(t_2,\omega_2)\right) = \underbrace{(t_1-t_2)^2}_{C_\tau} + \underbrace{(\omega_1-\omega_2)^2}_{C_\Omega}$$



$$\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$$

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_{|\Omega|}\}$$

Construct the cost matrix related to transport in **time-frequency domain** using the spectrograms in vectorized form:

$$\boxed{\mathbf{C} = k \cdot \mathbf{C}_{\mathcal{T}} \otimes \mathbf{1}_{|\Omega|\times|\Omega|} + \mathbf{1}_{|\mathcal{T}|\times|\mathcal{T}|} \otimes \mathbf{C}_{\Omega}}$$

The **parameter** $\boldsymbol{k}$ is used to balance the time and frequency costs magnitudes

3. Optimal transport for spectrograms:

Sliced Wasserstein approach

- Helpful to reduce the **dimensionality** of the problem by projecting the distributions onto subspaces of lower dimensions

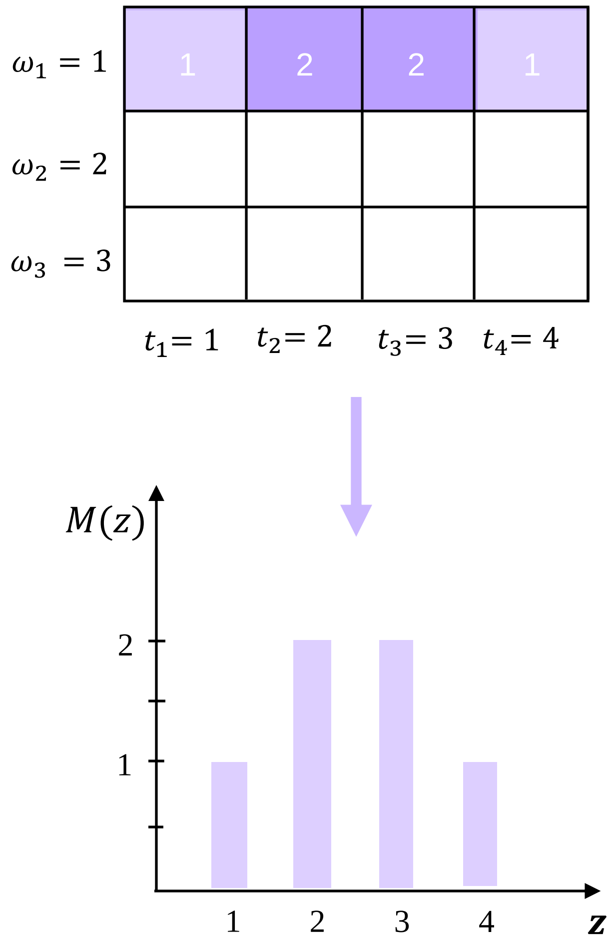
- Solve multiple 1D OT problems and average their distances

New “**position**” vector

$$\boxed{z^{i} = k \cdot t_{\ell} \cdot \theta_i^1 + \omega_m \cdot \theta_i^2} \rightarrow M(z)$$

with $\varphi_i = \frac{2\pi i}{N}$ for $i = 0, 1, 2, \ldots, N-1$

distribution of mass corresponding to the unique values of z



Kolouri, Soheil, Yang Zou, and Gustavo K. Rohde. "Sliced Wasserstein kernels for probability distributions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

4. Experiments – Sinusoids

Data: time-limited sinusoid sampled at 8 kHz, with central frequency 2 kHz and 2 seconds duration

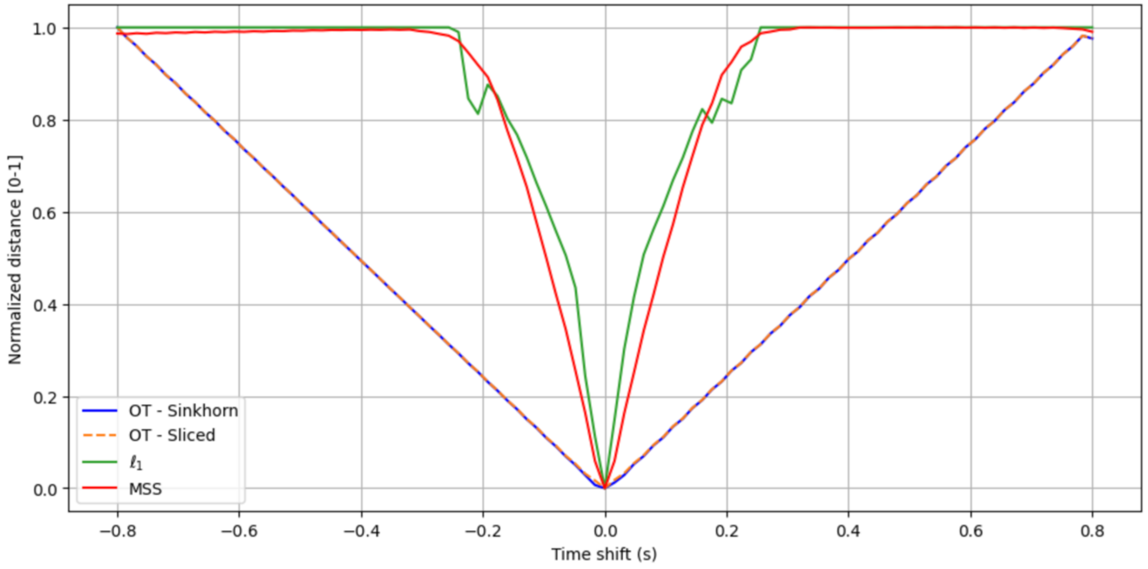
Comparing OT-based losses with $\boldsymbol{l_1}$ **spectral loss**

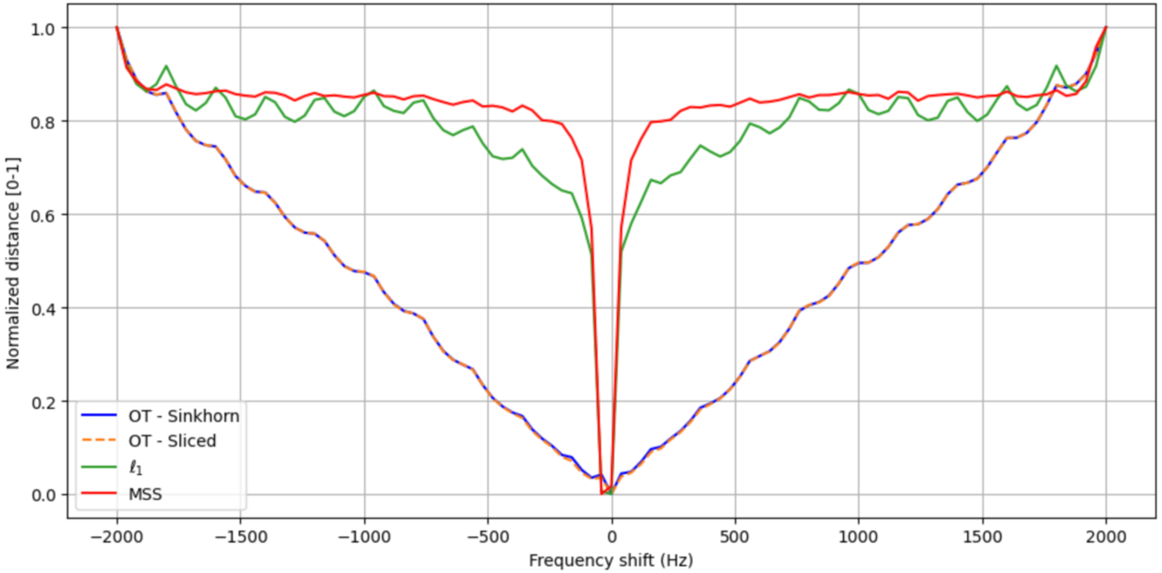
$$\mathcal{L}_1(\mathbf{S}, \hat{\mathbf{S}}) = \|\mathbf{S} - \hat{\mathbf{S}}\|_1$$

and **Multi-Scale Spectral loss**

$$\mathcal{L}_{MSS}(\mathbf{S}_\gamma, \hat{\mathbf{S}}_\gamma) = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \|\mathbf{S}_\gamma - \hat{\mathbf{S}}_\gamma\|_1 + \| \log(\mathbf{S}_\gamma) - \log(\hat{\mathbf{S}}_\gamma)\|$$

Time shifts

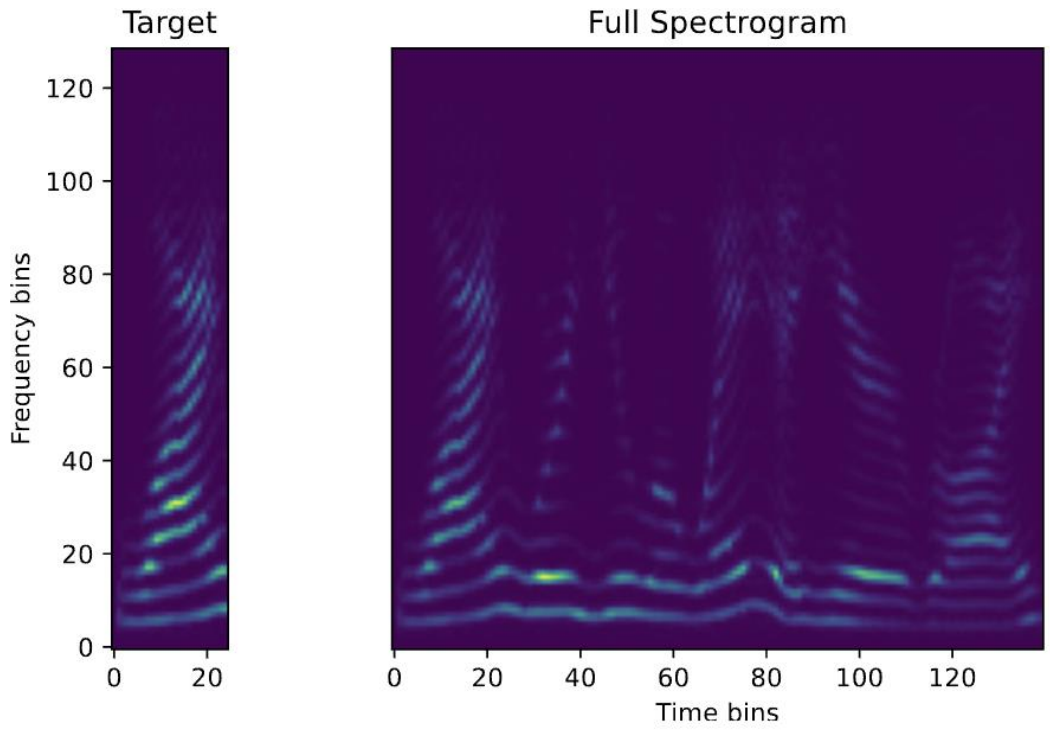




Frequency shifts

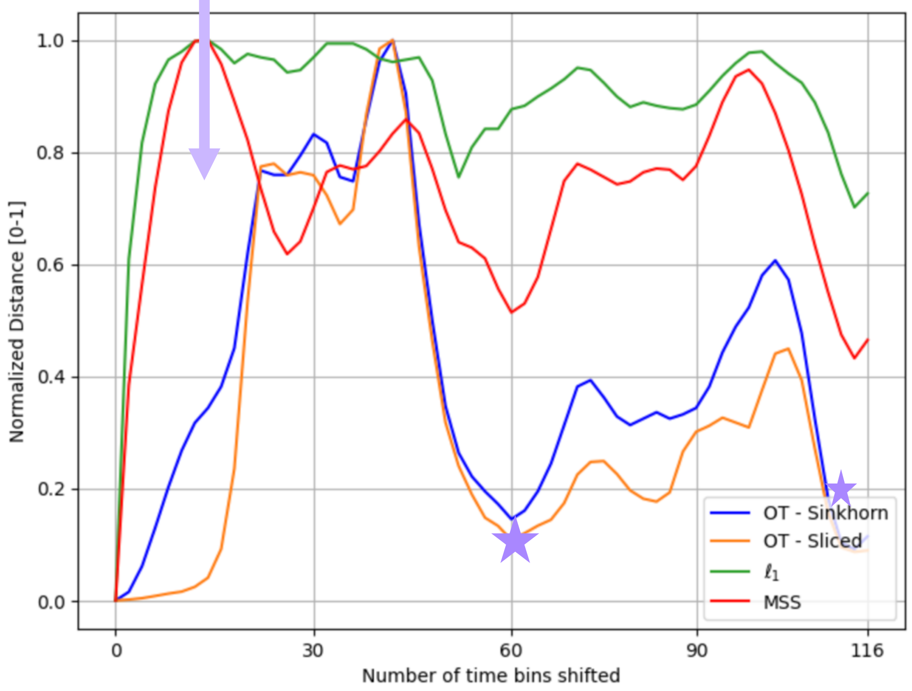
4. Experiments – Speech signal

Data: real speech signal corresponding to the phrase ”Why were you away a year, Roy?” compared through <u>segments</u> of length 24 bins



Gradual increase in distance as spectral overlap decreases

Possible similarity detected between the word ‘**Why**’ in the first segment and the words ‘**away**’ and ‘**Roy**’



Conclusions - Future works

✓ OT methods work with **magnitude** of the spectrograms → Extend them to account for the **phase** of complex valued signals

✓ OT losses show state-of-the-art performances for simple **DDSP reconstruction task** → explore more advanced DDSP tasks (timbre transfer, singing voice synthesis)

- Optimize distance calculation to improve computational expenses
- Include topics from psychoacoustics to the metric and validate with **listening experiments**

A!

A!

Thank you!