Enhancing Synthetic Speech Naturalness Through Optimal Transport

RIT | Rochester Institute of Technology

Mona Anil Udasi | Advisor: Dr. Anton Selitskiy | Golisano College of Computing and Information Sciences | Computer Science

Motivation

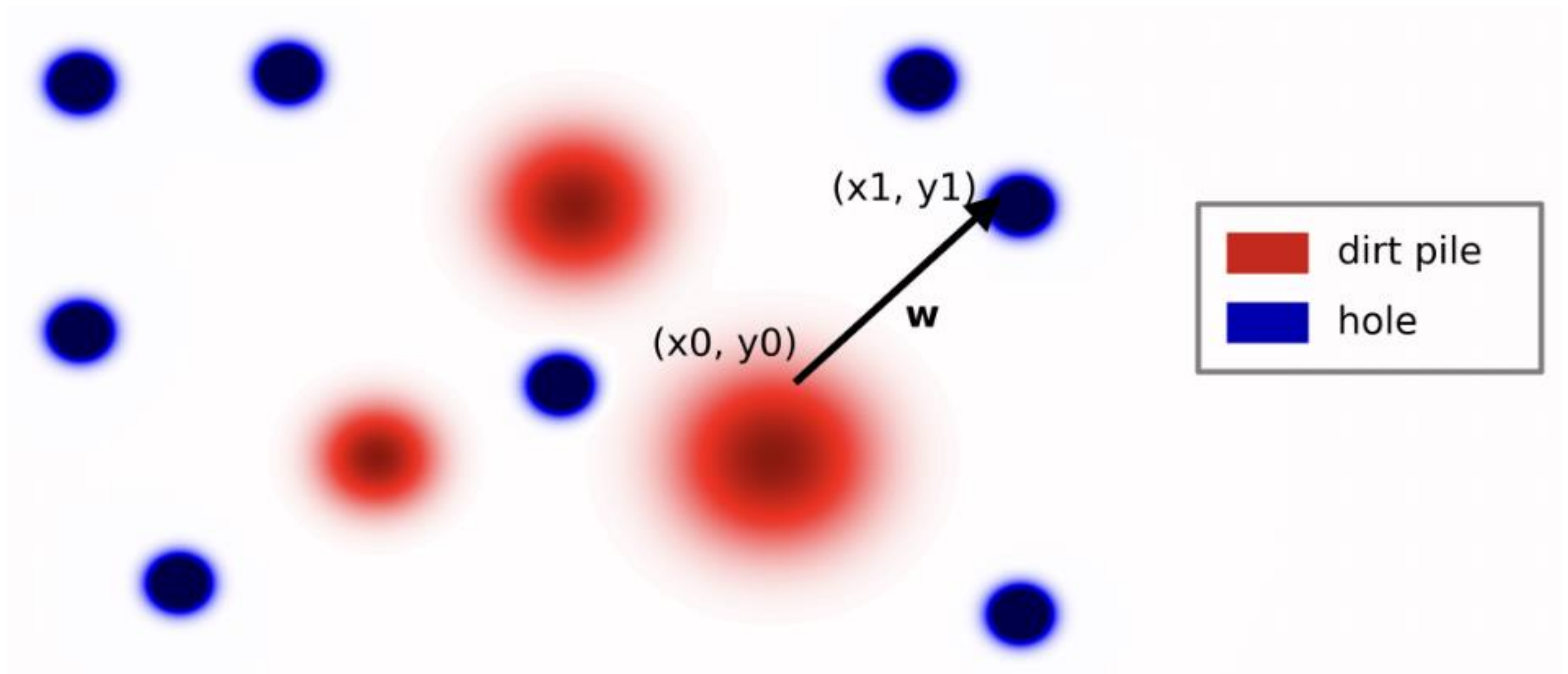
- Enhancing the naturalness of synthetic speech remains a major challenge.
- Synthetic speech often suffers from **unnatural prosody/rhythm, articulation, and pronunciation.**
- With the growing use of AI voice agents, improving speech quality is critical for user experience.

Project Overview

- **Optimal Transport (OT)**: i. Align two distributions via mapping
  ii. Minimize effort to transport mass
- Represent natural and synthetic speech as probability distributions.
- Treat synthetic speech as the source domain and natural speech as the target domain, and apply OT to align them efficiently.
- Evaluate improvements using a speech deepfake detection model.

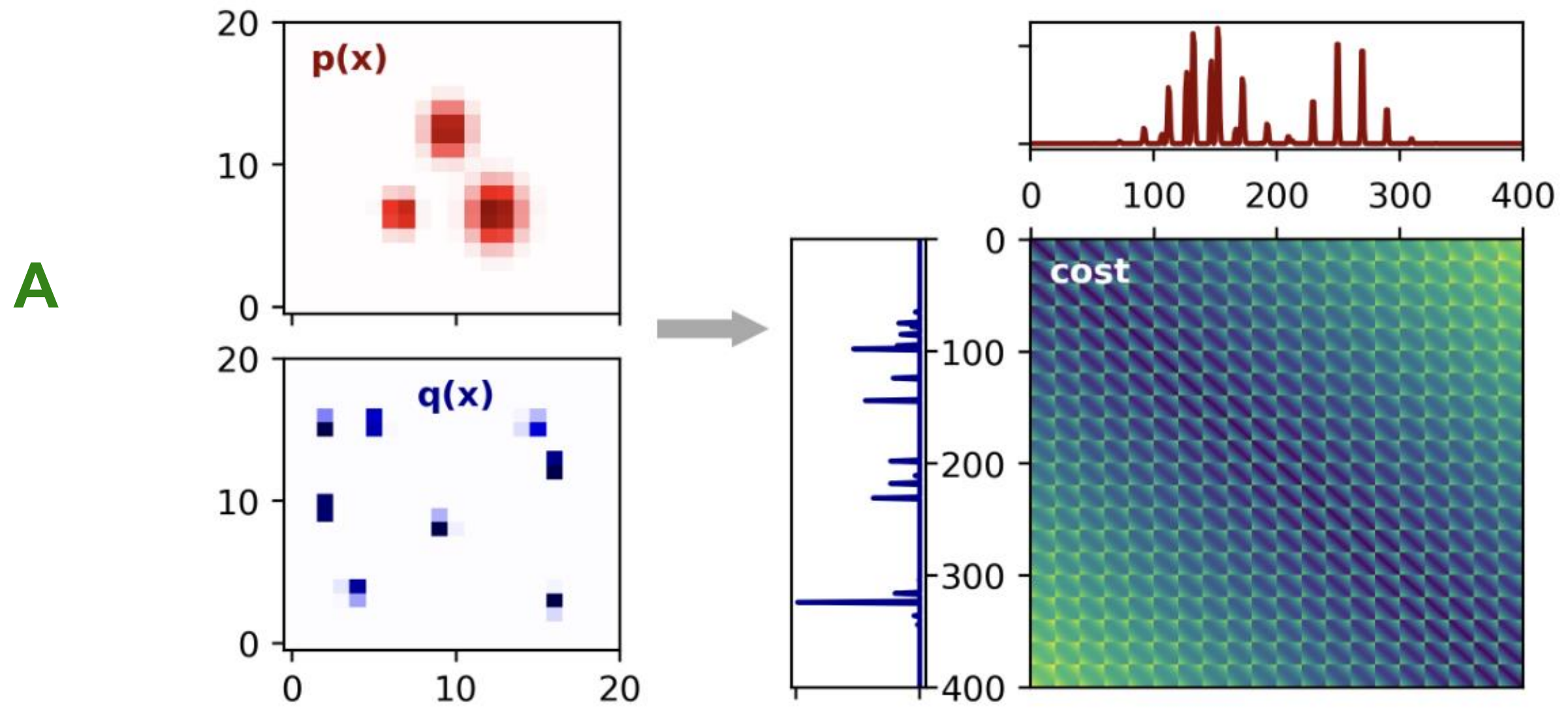
Optimal Transport

A single transport path



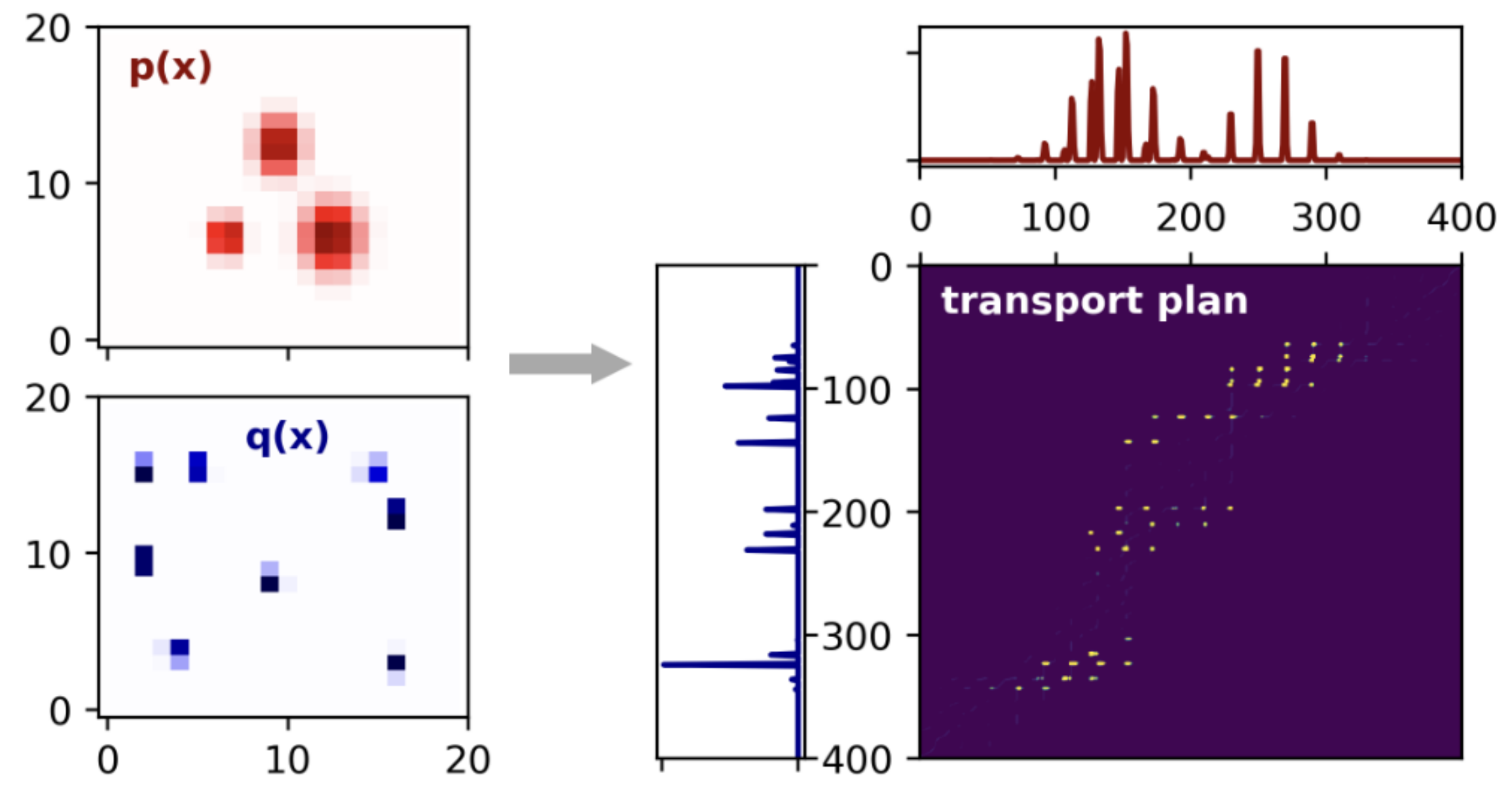
- **The arrow shows the transport of mass (w) from (x_0, y_0) to (x_1, y_1) [2].**

Calculating the cost matrix



- ***Left***, **the heat maps *p(x)* and *q(x)* correspond to the three dirt piles and scattered holes, respectively. *Right*, the cost matrix, C [2].**

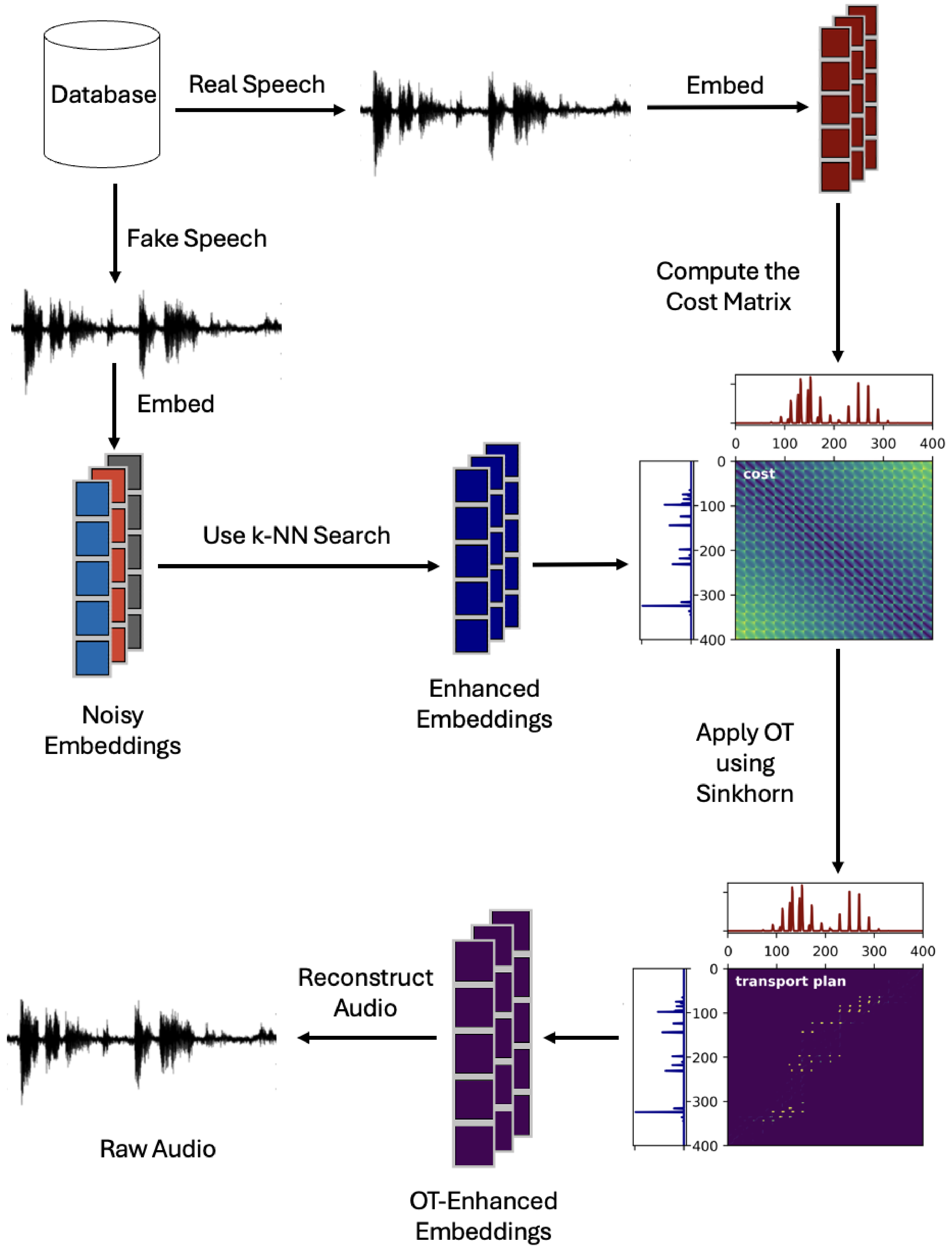
A complete transport plan



- ***Left*, the heat maps *p(x)* and *q(x),* as seen in Fig. A. *Right*, we calculate the transport plan matrix, T [2].**

System Pipeline

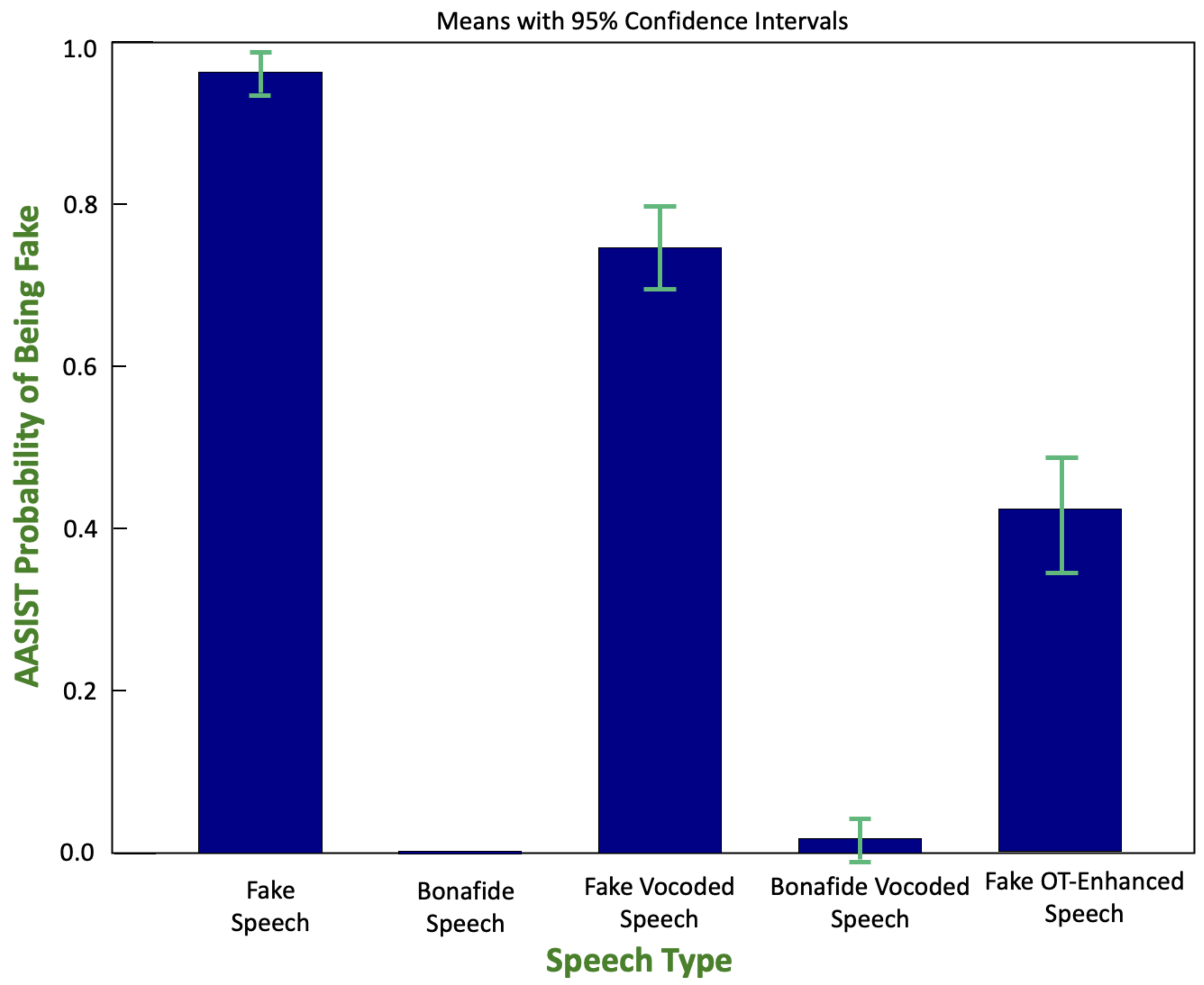
Real and fake speech samples from the ASVspoof 2019 dataset are embedded using the neural model WavLM [2] and reconstructed into audio with the neural vocoder HiFi-GAN.



Results

Human evaluations on selected speech samples showed slight improvements in prosody and phonetic quality after applying OT. In contrast, AASIST, which outputs a probability score for how likely speech is fake, revealed the following results:

- Fake speech was initially detected with an extremely high score of 0.97.
- Applying OT significantly reduced the detection score to 0.43.
- Vocoding alone slightly reduced the score to 0.72 but was not sufficient.
- This confirms that OT-driven alignment, not just vocoding, contributed meaningfully to reducing detectability.



Conclusion

- Optimal Transport helped shift synthetic speech embeddings closer to natural speech distributions.
- AASIST evaluations confirmed that OT alignment significantly reduced the detectability of fake speech.
- Working with spectrograms may improve results by preserving more speech details.

Acknowledgement

- [1] S.Chen et al., “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”
- [2] Alex Williams. A Short Introduction to Optimal Transport and Wasserstein Distance. 2020.