



HAL
open science

Audio signal interpolation using optimal transportation of spectrograms

David Valdivia, Marien Renaud, Elsa Cazelles, Cédric Févotte

► **To cite this version:**

David Valdivia, Marien Renaud, Elsa Cazelles, Cédric Févotte. Audio signal interpolation using optimal transportation of spectrograms. 2025. hal-04960810

HAL Id: hal-04960810

<https://hal.science/hal-04960810v1>

Preprint submitted on 21 Feb 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Audio signal interpolation using optimal transportation of spectrograms

David Valdivia
IRIT, Université de Toulouse

Marien Renaud
IMB, Université de Bordeaux

Elsa Cazelles, Cédric Févotte
IRIT, CNRS, Université de Toulouse

Abstract—We present a novel approach for generating an artificial audio signal that interpolates between given source and target sounds. Our approach relies on the computation of Wasserstein barycenters of the source and target spectrograms, followed by phase reconstruction and inversion. In contrast with previous works, our new method considers the spectrograms globally and does not operate on a temporal frame-to-frame basis. An other contribution is to endow the transportation cost matrix with a specific structure that prohibits remote displacements of energy along the time axis, and for which optimal transport is made possible by leveraging the unbalanced transport framework. The proposed cost matrix makes sense from the audio perspective and also allows to reduce the computation load. Results with synthetic musical notes and real environmental sounds illustrate the potential of our novel approach.

Index Terms—Optimal transport, time-frequency analysis, audio interpolation.

I. INTRODUCTION

We consider the problem of generating artificial sounds that interpolate between given source and target signals. Though we consider audio signals in this paper (for creative and editing purposes), the methodology proposed in this paper is not limited to this class of signals. Given two temporal signals $\mathbf{y}^s \in \mathbb{R}^L$ (source) and $\mathbf{y}^t \in \mathbb{R}^L$ (target) we wish to generate an interpolant $\mathbf{y}^\alpha \in \mathbb{R}^L$ (for any $\alpha \in [0, 1]$) such that $\mathbf{y}^0 = \mathbf{y}^s$ and $\mathbf{y}^1 = \mathbf{y}^t$. To achieve this task, a baseline option is to compute the Euclidean barycenter of \mathbf{y}^s and \mathbf{y}^t with weights $(1 - \alpha, \alpha)$ given by

$$\mathbf{y}^\alpha = \underset{\mathbf{y}}{\operatorname{argmin}} (1 - \alpha) \|\mathbf{y} - \mathbf{y}^s\|_2^2 + \alpha \|\mathbf{y} - \mathbf{y}^t\|_2^2 \quad (1)$$

$$= (1 - \alpha) \mathbf{y}^s + \alpha \mathbf{y}^t. \quad (2)$$

As shown by (2), the Euclidean barycenter is just a weighted average of the source and target signals. The objective of this paper is to explore an alternative option based on *optimal transport* (OT), that can produce truly hybrid sounds. In the proposed setting, the normalized time-frequency (t-f) distributions of the source and target signals are interpreted as discrete probability distributions supported by the t-f grid. We then propose to compute a Wasserstein barycenter [1], also called an interpolant, of these t-f distributions. The resulting t-f barycenter may then be inverted using suitable t-f grid reassignment and phase reconstruction to obtain a temporal

interpolant. We also propose to endow the transportation cost matrix with a specific structure that prohibits remote displacements of energy along the time axis, and for which optimal transport is made possible by leveraging the unbalanced transport framework. The proposed cost matrix makes sense from the audio perspective and also allows to reduce the computation load.

To the best of our knowledge, [2] is the first prior work to consider optimal transport of spectra for audio interpolation. The setting in [2] is however more restricted than ours because the authors consider the frames of the spectrogram individually (frame-to-frame OT). Furthermore, the source and target t-f distributions result from a specific pitch-based processing of the spectrograms, which exploits the harmonic nature of tonal musical signals. Peak frequencies are identified in the source and target spectra using t-f reassignment [3]. Those peak frequencies and their energy (integrated over a selected neighborhood) lay ground to the OT procedure. This approach is well suited to pitched sounds. An other prior work in audio interpolation with OT is [4]. In this work, a nonnegative matrix factorization (NMF) of the source and target spectrograms is first computed [5]. Simplifying a little, an interpolated spectrogram is then constructed by multiplying the temporal activations of the target spectrogram with a dictionary of interpolated spectra. The latter are obtained by interpolating the source and target spectral dictionaries obtained by NMF, using the OT procedure of [2]. In contrast with the approach of [2], [4], we consider the spectrogram *globally*, meaning that transportation of energy is possible over neighboring time frames. Furthermore, we do not resort to any pre-processing and work directly on the raw spectrograms. This allows to work with a wider range of sounds, and in particular wide-band signals. Other technical differences related to the support of the interpolated distribution and to phase reconstruction will be discussed next.

While OT has been considered in many image processing problems, it has been less popularized in audio signal processing settings and this paper is also a step in that direction. Besides [2], [4] that consider audio interpolation, other applicative examples of OT in audio signal processing are music transcription [6], [7] and audio classification [8], [9]. The methods derived in this paper are non-supervised (the source and target signals are the only two inputs). Supervised learning-based approaches have also been considered for audio interpolation, see, e.g., [10]. These approaches can

This work is supported by ANITI under the French ANR Cluster IA programme. Part of this work was conducted while Marien Renaud was a MSc intern student at IRIT.

produce spectacular results but rely on the intensive training of generative networks.¹

The paper is organized as follows. Section II introduces notations and elements of OT. We then present a first interpolation method based on Wasserstein barycenters in Section III. Some limitations of this baseline method are remedied by our second method, that relies on unbalanced OT and a structured cost matrix, as described in Section IV. Finally, we report experiments with synthetic musical notes and real environmental sounds in Section V, before concluding in Section VI.

II. PRELIMINARIES

A. Notations

We denote by $\mathbf{X}^s, \mathbf{X}^t \in \mathbb{R}_+^{M \times N}$ the spectrograms of the source and target signals. They are equal to the magnitude of a short-time Fourier transform (STFT) of \mathbf{y}^s and \mathbf{y}^t . The values of M (number of frequency bins) and N (number of time frames) are dictated by the length of the chosen STFT analysis window and hop size. Dropping the superscripts s and t for conciseness, the coefficients of the spectrogram \mathbf{X} are denoted by $X_{mn} \in \mathbb{R}_+$, with $(m, n) \in \{1, \dots, M\} \times \{1, \dots, N\}$. The t-f couples (m, n) index t-f points of the form $\omega_{mn} = (f_m, t_n)$ with $f_m = \frac{m-1}{M} \frac{f_s}{2}$ and $t_n = (n-1) \frac{H}{f_s}$, where f_s and H denote the sampling frequency (Hz) and hop size (in samples).

Let us define by \mathbf{x}^s and \mathbf{x}^t the vectorized versions of \mathbf{X}^s and \mathbf{X}^t . The vectorization underlies an arbitrary one-to-one mapping between the t-f grid $(m, n) \in \{1, \dots, M\} \times \{1, \dots, N\}$ and a point $i \in \{1, \dots, I\}$, with $I = MN$, such that $x_i^s = X_{mn}^s$ and $x_i^t = X_{mn}^t$. We will sometimes use the shorthand $\omega_i = \omega_{mn}$. We assume that the spectrograms have been normalized (globally), so that $\sum_i x_i^s = \sum_i x_i^t = 1$.

B. Optimal transport and Wasserstein barycenters

Equipped with these notations and conventions, we may view the source and target spectrograms \mathbf{X}^s and \mathbf{X}^t as discrete probability distributions μ^s and μ^t supported by the t-f points $\omega_i \in \mathbb{R}^2$, i.e.,

$$\mu = \sum_{m=1}^M \sum_{n=1}^N X_{mn} \delta_{\omega_{mn}} = \sum_{i=1}^I x_i \delta_{\omega_i} \quad (3)$$

where we dropped the superscripts s or t , and where δ is the Dirac delta function. Using OT terminology, the coefficients X_{mn} or x_i will sometimes be referred to as ‘‘mass’’.

Definition 1: The optimal transport distance, or 2-Wasserstein distance, between distributions μ^s and μ^t is given by $\sqrt{\text{OT}(\mu^s, \mu^t)}$, with

$$\text{OT}(\mu^s, \mu^t) = \min_{\mathbf{P} \in \Pi(\mathbf{x}^s, \mathbf{x}^t)} \langle \mathbf{C}, \mathbf{P} \rangle, \quad (4)$$

where $\Pi(\mathbf{x}^s, \mathbf{x}^t) := \{\mathbf{P} \in \mathbb{R}_+^{I \times I} | \mathbf{P}\mathbf{1}_I = \mathbf{x}^s, \mathbf{P}^\top \mathbf{1}_I = \mathbf{x}^t\}$ is the set of transport plans \mathbf{P} with marginals \mathbf{x}^s and \mathbf{x}^t ($\mathbf{1}_I$ denotes the vector in \mathbb{R}^I with all entries equal to one), and \mathbf{C}

¹Audio interpolation may also refer to the tasks of filling missing gaps in a corrupted signal [11], [12] or synthesizing virtual 3D sounds [13]. Such tasks are not considered here but could benefit from our results in future work.

is the transportation cost matrix defined entry-wise by $C_{ii'} = \|\omega_i - \omega_{i'}\|^2$.

Definition 2: The Wasserstein barycenter of coefficient $\alpha \in [0, 1]$ between the two distributions μ^s and μ^t is defined by

$$\mu^\alpha \in \underset{\nu}{\text{argmin}} (1 - \alpha) \text{OT}(\mu^s, \nu)^2 + \alpha \text{OT}(\mu^t, \nu)^2. \quad (5)$$

It can easily be shown with the triangular inequality that μ^α is explicit in terms of an optimal plan $\mathbf{P}^* \in \Pi(\mathbf{x}^s, \mathbf{x}^t)$ in (4) [14], and is given by

$$\mu^\alpha = \sum_{i, i'=1}^I P_{ii'}^* \delta_{(1-\alpha)\omega_i + \alpha\omega_{i'}}. \quad (6)$$

Note that \mathbf{P}^* is independent of α .

III. SIGNAL INTERPOLATION USING TIME-FREQUENCY WASSERSTEIN BARYCENTERS

When μ^s and μ^t are the t-f distributions (3) of the source and target signals, the distribution μ^α given by (6) defines a t-f barycenter that one might wish to invert back into the time domain in order to reconstruct an audio interpolant \mathbf{y}^α of \mathbf{y}^s and \mathbf{y}^t . This simple and yet novel idea raises practical issues regarding the t-f support of μ^α and phase reconstruction, as discussed next.

A. Time-frequency reassignment

As indicated by (6), the support of μ^α is the set of t-f points $\omega_{ii'}^\alpha := (1 - \alpha)\omega_i + \alpha\omega_{i'}$, $(i, i') \in \{1, \dots, I\}^2$. Equivalently, given $\alpha \in [0, 1]$, the support of μ^α is given by the $(MN)^2$ t-f points $\{((1 - \alpha)f_m + \alpha f_{m'}, (1 - \alpha)t_n + \alpha t_{n'})\}_{mm'nn'}$. As such, μ^α is not supported by the same t-f grid than μ^s and μ^t , it instead is spread over many more t-f points which form an irregular grid. In this paper, we choose to reconstruct a native-grid spectrogram $\mathbf{X}^\alpha \in \mathbb{R}_+^{M \times N}$ from μ^α by reassigning the mass $P_{ii'}^*$ supported by $\omega_{ii'}^\alpha$ to the closest native t-f point ω_{j^*} , such that $j^* = \text{argmin}_j \|\omega_j - \omega_{ii'}^\alpha\|$.

B. Phase reconstruction

Given the barycentric spectrogram \mathbf{X}^α reconstructed in the previous step, we use the classical Griffin and Lim phase reconstruction algorithm [15] and overlap-add inverse STFT to reconstruct a temporal signal \mathbf{y}^α .

C. Discussion

The baseline approach presented at the beginning of this section presents some limitations that we now discuss. Firstly, the computation of \mathbf{P}^* in (4) might be too computationally intensive even for small-size problems corresponding to few seconds of audio. A standard remedy is to consider entropy-regularized OT, which can be solved more efficiently using the Sinkhorn algorithm [16]. Unfortunately, we found out that this solution is not desirable for spectrogram interpolation, as the entropy term tends to spread the mass of the optimal plan too smoothly, cutting out time-frequency details. Secondly, even for short audio signals, we found preferable to prevent the mass from moving to anywhere on the t-f support. As such, we

propose in the next section a structured cost matrix that favors mass displacement along the frequency axis while controlling the permissible displacement over neighboring time frames. We then leverage the unbalanced OT framework to efficiently compute the corresponding optimal transport plan.

IV. SIGNAL INTERPOLATION USING UNBALANCED OPTIMAL TRANSPORT AND A STRUCTURED COST MATRIX

A. Structured cost matrix

In order to produce audio results that sound more natural, we propose to prevent remote time displacements of mass by introducing the cost matrix $\bar{\mathbf{C}} \in \mathbb{R}_+^{I \times I}$ with coefficients

$$\bar{C}_{ii'} = \begin{cases} (f_m - f_{m'})^2 + (t_n - t_{n'})^2 & \text{if } |n - n'| \leq p, \\ +\infty & \text{otherwise,} \end{cases} \quad (7)$$

where (f_m, t_n) (resp. $(f_{m'}, t_{n'})$) is the t-f point mapped with i (resp. i'). In contrast with the standard ℓ_2^2 cost matrix \mathbf{C} introduced in Section II-B, $\bar{\mathbf{C}}$ now explicitly forbids mass displacements of time range larger than p frames. As our experiments will show, the user-defined parameter p plays a significant role.² Note that when $p = 0$, the cost matrix $\bar{\mathbf{C}}$ only allows frequency displacements in matching frames $n = n'$ (mass cannot spill over adjacent frames). This is close to the frame-to-frame setting of [2], with the exception that in our case the spectrograms are still considered as a whole distribution (and normalized globally) as opposed to every frame being normalized like in [2].

B. Unbalanced OT and spectrogram interpolants

As the proposed structured cost matrix (7) proscribes certain displacement of mass, the OT problem (4) using $\bar{\mathbf{C}}$ might not have a solution verifying the marginals constraint [14]. We therefore consider unbalanced OT [17], [18] which relaxes the conservation of mass constraints in problem (4):

$$\text{UOT}_\beta(\mu^s, \mu^t) = \min_{\mathbf{P} \in \mathbb{R}_+^{I \times I}} \langle \mathbf{C}, \mathbf{P} \rangle + \beta \text{KL}(\mathbf{P} \mathbf{1}_I, \mathbf{x}^s) + \beta \text{KL}(\mathbf{P}^T \mathbf{1}_I, \mathbf{x}^t), \quad (8)$$

where KL is the Kullback-Leibler divergence between nonnegative numbers, and $\beta \in \mathbb{R}_+$ is a parameter that controls the trade-off between mass transportation and mass creation/loss. For a large β , one recovers the OT problem (4), whereas a small β will encourage displacements of mass between points that are close in the t-f grid, promoting a local consistency of frequency content between the source and target spectrograms.

Following (6), we reconstruct a t-f interpolant from an optimal solution $\mathbf{P}^\beta \in \mathbb{R}_+^{I \times I}$ of (8):

$$\mu_\beta^\alpha = \sum_{i, i'=1}^I P_{ii'}^\beta \delta_{(1-\alpha)\omega_i + \alpha\omega_{i'}}. \quad (9)$$

An audio interpolant \mathbf{y}^α is then reconstructed from μ_β^α as in Section III.

²In practice, we use dimensionless time and frequency variables and simply set $C_{ii'} = (m - m')^2 + (n - n')^2$, and likewise for $\bar{C}_{ii'}$.

A. Implementation details and experimental setting

The interpolant (6) requires to solve the OT problem (4), for which we used the POT toolbox [19]. The interpolant (9) requires to solve the UOT problem (8), for which we use the majorization-minimization (MM) algorithm introduced in [20]. We implemented a dedicated version of the MM algorithm that leverages the particular structure of $\bar{\mathbf{C}}$ in (7). Indeed, we may only consider indices ii' with finite entries $\bar{C}_{ii'}$, because $P_{ii'}^\beta = 0$ whenever $\bar{C}_{ii'} = +\infty$. This allows for an efficient implementation that can run with audio signals of up to a few seconds with standard sampling frequencies.

In the following, we presents results with synthetic musical notes and real environmental sounds. For comparison with OT and readability of the spectrograms, we use short audio signals of duration 1 second sampled at $f_s = 16\text{kHz}$. The spectrograms are computed with a Hann window of duration 40ms with overlap 50%. Audio samples and code are available online.³

B. Interpolating musical notes

Our first experiment consists in interpolating two versions of the same sequence C3-G3 played either with a piano (source) or a guitar (target). The two input signals were generated with the music production software Ableton Live. We set the interpolation parameter to $\alpha = 0.5$. For the UOT-based method, we set the hyperparameter to $\beta = 1$ and the time-limiting parameter to $p = 0$. Results are shown in Fig. 1.

We observe that the structure of $\bar{\mathbf{C}}$ has a decisive influence on the nature of the interpolants. Since OT is *constrained* to move all the mass from the source to the target, some mass is spread along the time axis. This produces a continuous audio signal in which the temporal structure of the source and target signals is lost, see Fig. 1(c). On the other hand, by limiting the mass displacement on the time axis, UOT is able to produce an interpolant that preserves the dynamics of the input signals as seen in Fig. 1(d). This results in the same sequence C3-G3 played by a hybrid instrument (in between a guitar and a piano), as can be gathered from the online audio samples.

C. Interpolating environmental sounds

We now consider environmental sounds that have a more *textured* structure as opposed to mostly harmonic. The source is the chirp of a cicada and the target is a sample of flowing water. The cicada's chirp is characterized by strong high-frequency energy and by a clear rhythmic pattern, whereas the water flow is continuous and essentially wide-band.

We set $\alpha = 0.5$ and $\beta = 1$ like in previous section, but we now explore various values of the time-limiting parameter p , see Fig. 2. Fig. 2(c) and Fig. 2(d) show that OT and UOT with $p = 0$ capture essential characteristics of the input signals: the temporal structure induced by the cicada's chirp and wide-band energy content. However, Fig. 2(c) shows that the temporal structure of the OT interpolant is not as sharp, again because of the conservation of mass constraint. Increasing the

³<https://davidvaldiviad.github.io/audio-signal-interpolation-ot/>

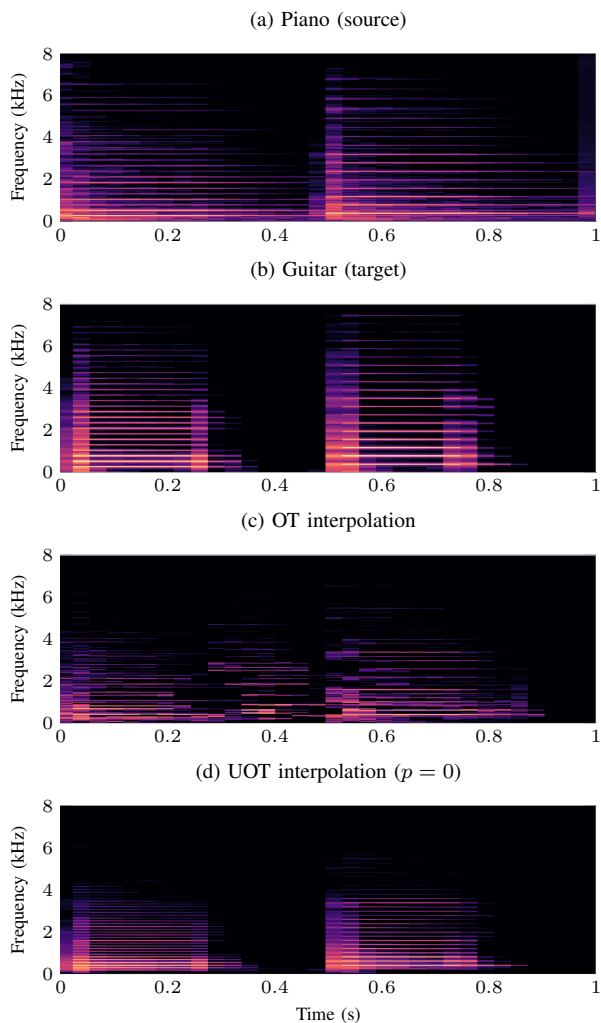


Fig. 1. Interpolation between two versions of the same sequence of notes C3-G3 played on a piano and a guitar, using $\alpha = 0.5$.

value of p for the UOT interpolant has a significant impact, see Figs. 2(e-f): larger values of p allow for increasing temporal transportation (i.e., along the temporal axis). The perceptual impact of the value of p can be assessed by listening to the audio samples available online.

VI. DISCUSSION

In our opinion, our work opens promising research directions regarding the optimal transportation of spectrograms for signal processing applications. In this paper we have resorted to a t-f grid reassignment followed by standard phase retrieval to proceed to the temporal reconstruction of the interpolants. An exciting research direction would be to design inversion methods that circumvent the reassignment and work directly in the irregular grid. Another direction would be to design t-f interpolants that are constrained to be supported by the native grid. A third direction would be to leverage the phase of the source and target signals to produce better estimates of the phase of the interpolant. This latter direction in particular may further improve audio quality.

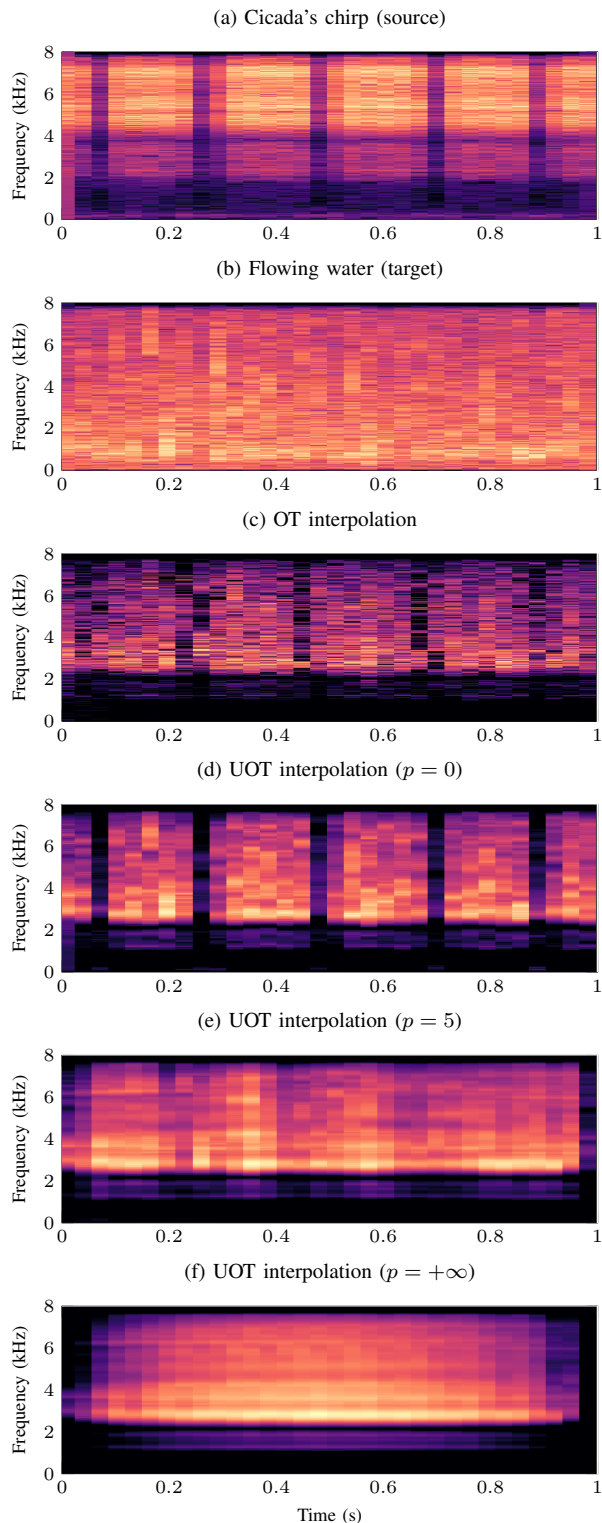


Fig. 2. Interpolation between a cicada's chirp and flowing water, using $\alpha = 0.5$. Various values of the time-limiting parameter p are considered. In subplot (f) we use $p = +\infty$ as a shorthand for using $\tilde{\mathbf{C}} = \mathbf{C}$ (the time limit is lifted).

REFERENCES

- [1] M. Agueh and G. Carlier, “Barycenters in the Wasserstein space,” *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.
- [2] T. Henderson and J. Solomon, “Audio transport: a generalized portamento via optimal transport,” in *Proc. International Conference on Digital Audio Effects (DAFx)*, 2019.
- [3] F. Auger, P. Flandrin, Y.-T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, and H.-T. Wu, “Time-frequency reassignment and synchrosqueezing: An overview,” *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 32–41, 2013.
- [4] G. Roma, O. Green, and P. A. Tremblay, “Audio morphing using matrix decomposition and optimal transport,” in *Proc. International Conference on Digital Audio Effects (DaFx)*, 2020.
- [5] P. Smaragdīs, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman, “Static and dynamic source separation using nonnegative factorizations: A unified view,” *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.
- [6] R. Flamary, C. Févotte, N. Courty, and V. Emiya, “Optimal spectral transportation with application to music transcription,” in *Advances in Neural Information Processing Systems (NIPS)*, 2016.
- [7] F. Elvander, S. I. Adalbjörnsson, J. Karlsson, and A. Jakobsson, “Using optimal transport for estimating inharmonic pitch signals,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 331–335.
- [8] A. Dessein, N. Papadakis, and J.-L. Rouas, “Regularized optimal transport and the rot mover’s distance,” *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 590–642, 2018.
- [9] E. Cazelles, A. Robert, and F. Tobar, “The Wasserstein-fourier distance for stationary time series,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 709–721, 2021.
- [10] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, “Gansynth: Adversarial neural audio synthesis,” in *Proc. International Conference on Learning Representations (ICLR)*, 2019.
- [11] S. J. Godsill and P. J. Rayner, *Digital audio restoration*. Springer, 1998.
- [12] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, “Audio inpainting,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.
- [13] R. J. Radke and S. Rickard, “Audio interpolation for virtual audio synthesis,” *Journal of the Audio Engineering Society*, 2002.
- [14] G. Peyré and M. Cuturi, “Computational optimal transport: With applications to data science,” *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.
- [15] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 8, 1983, pp. 804–807.
- [16] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Advances in Neural Information Processing Systems*, vol. 26, 2013.
- [17] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, “Unbalanced optimal transport: Dynamic and kantorovich formulations,” *Journal of Functional Analysis*, vol. 274, no. 11, pp. 3090–3123, 2018.
- [18] M. Liero, A. Mielke, and G. Savaré, “Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures,” *Inventiones mathematicae*, vol. 211, no. 3, pp. 969–1117, 2018.
- [19] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, “POT: Python optimal transport,” *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021.
- [20] L. Chapel, R. Flamary, H. Wu, C. Févotte, and G. Gasso, “Unbalanced optimal transport through non-negative penalized linear regression,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 23 270–23 282, 2021.