AN IMPROVED OPTIMAL TRANSPORT KERNEL EMBEDDING METHOD WITH GATING MECHANISM FOR SINGING VOICE SEPARATION AND SPEAKER IDENTIFICATION

[1]*Weitao Yuan**   [1]*Yuren Bian*   [1]*Shengbei Wang*   [2]*Masashi Unoki*†   [3]*Wenwu Wang*

[1] Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tiangong University, China
[2]Japan Advanced Institute of Science and Technology, Japan
[3]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

ABSTRACT

Singing voice separation (SVS) and speaker identification (SI) are two classic problems in speech signal processing. Deep neural networks (DNNs) solve these two problems by extracting effective representations of the target signal from the input mixture. Since essential features of a signal can be well reflected on its latent geometric structure of the feature distribution, a natural way to address SVS/SI is to extract the geometry-aware and distribution-related features of the target signal. To do this, this work introduces the concept of optimal transport (OT) to SVS/SI and proposes an improved optimal transport kernel embedding (iOTKE) to extract the target-distribution-related features. The iOTKE learns an OT from the input signal to the target signal on the basis of a reference set learned from all training data. Thus it can maintain the feature diversity and preserve the latent geometric structure of the distribution for the target signal. To further improve the feature selection ability, we extend the proposed iOTKE to a gated version, i.e., gated iOTKE (G-iOTKE), by incorporating a lightweight gating mechanism. The gating mechanism controls effective information flow and enables the proposed method to select important features for a specific input signal. We evaluated the proposed G-iOTKE on SVS/SI. Experimental results showed that the proposed method provided better results than other models.

Index Terms— Optimal transport, optimal transport kernel embedding, gating mechanism, singing voice separation, speaker identification

1. INTRODUCTION

An important task in deep neural network (DNN) based singing voice separation (SVS) [1–7] and speaker identification (SI) [8–10] is to learn useful representations of the target vocal source/speaker with effective neural layers. Successful representation learning relies heavily on extracting and selecting the discriminative feature set that follows the feature distribution of the target source/speaker. In SVS/SI, the input signal (e.g., a mixture signal for SVS or noisy speech for SI) is usually transformed into a two-dimensional (or multi-channel) representation, which is a time-frequency feature set with all features of the target signal and background interferences mixed together. Therefore, it is very challenging to extract and select the specific feature set that follows the underlying distribution of the target signal.

The feature extracting operations in most SVS/SI works are achieved on the basis of matrix multiplication, such as Linear Unit (LU), Gated LU (GLU) [11], and convolution operator. However, these simple matrix multiplication operations may not effectively preserve the feature distribution of the target source/speaker. In our opinion, the SVS can be considered as a transport problem from the mixture distribution to a mono-component vocal distribution and the SI is a matching problem from the input distribution to that of the target speaker. Therefore, the key to solving SVS/SI is to find a suitable and effective transport that can maintain the feature distribution of the target signal after the transportation.

To address this task, we introduce optimal transport (OT) to SVS/SI. The OT proposed in [12–15] defines an effective geometry-aware Wasserstein distance. The OT method can measure the difference between two distributions by using the Euclidean distance between their optimally transported representations. Based on OT, optimal transport kernel embedding (OTKE), an effective geometric-structure-aware feature aggregation method, was proposed in [16]. The OTKE presents a new weighted pooling operation to embed and aggregate the features with respect to a trainable/learnable reference set. The reference set, which enables an end-to-end training with small computational cost, can learn the latent geometric structure of the target distribution for a specific task. Since OT/OTKE can measure the distance between different distributions, they can be potentially used for SVS/SI to match the latent geometric structure between the input and target feature sets and provide a more suitable transport from the input features to the target features.

This work proposed an improved OTKE (iOTKE) for SVS/SI and extended it to a gated version, i.e., gated iOTKE (G-iOTKE), by applying a lightweight gating mechanism to iOTKE. The proposed method solves the SVS/SI problem by finding a gated OT from the input distribution to the target distribution. This is achieved by first (i) learning and extracting the distribution-related feature set with iOTKE and then (ii) selecting important and relevant features of a specific input with the gating mechanism. Different from existing OTKE [16–18] that aggregates large-size features into fixed and small-size embeddings, the proposed iOTKE keeps the same size for input and output feature sets so that there are enough trainable references to learn the diversity of the target features. For feature selection, we use a lightweight gating mechanism to control the information flow and select relevant features, where the gating mechanism can enhance or weaken a feature in accordance with a specific input. We apply the proposed G-iOTKE to SVS and SI. Experimental results showed that the proposed G-iOTKE can effectively improve the performance of SVS/SI.

2. PROPOSED METHOD

Different sources/speakers in SVS/SI are usually characterized by different features. Even so, the SVS/SI is proven to be quite challenging as the essential features that can be used to distinguish different

*Thanks to the National Natural Science Foundation of China (No. 62176182, No. 61902280, and No. 61771340), the Program for Innovative Research Team in University of Tianjin (No. TD13-5032), and the Scientific Research Project of Tianjin Education Commission (19PTZWHZ00020) for funding. (Corresponding author: Shengbei Wang)

†This work was also supported by a Grant-in-Aid for Scientific Research (B) (21H03463) and Fund for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233).

sources/speakers are hidden in the massive and redundant feature set of the input mixture signal. To the best of our knowledge, essential features of a signal can be well reflected on the geometric structure of its latent distribution. Therefore, this work uses iOTKE to learn the geometric structure of the target signal. To do this, we first learn a geometry-aware and target-distribution-related feature set. This feature set is explicitly used as a "reference set" to realize an optimal feature transportation for a specific input signal.

2.1. Learning target-distribution-related features with iOTKE

Suppose the embedded feature set of the input feature $\mathbf{x}_t$ in SVS/SI is $\mathbf{x}_0 = \phi(\mathbf{x}_t) \in \mathbb{R}^{n \times c}$, where ϕ is the learnable kernel for kernel embedding (KE) in iOTKE, n represents the number of features and c denotes the channel number for each feature. We apply Rectified LU (ReLU) to $\mathbf{x}_0$ to preserve its positive values,

$$\mathbf{x} = \text{ReLU}(\mathbf{x}_0), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}_+^{n \times c}$ ("+" means non-negative value) is the obtained non-negative feature set. This above function ensures that the input for iOTKE is a non-negative and sparse feature set, which is very useful for general modelling of audio signals [19, 20].

The non-negative feature set $\mathbf{x}$ is optimally transported based on a target-distribution-related feature set $\mathbf{z} \in \mathbb{R}_+^{n \times c}$ learned from all training data, where $\mathbf{z}$ is the "reference set" and has the same dimension as $\mathbf{x}$. Specifically, to extract features related to the latent geometric structure/distribution of the target signal, we optimally transport $\mathbf{x}$ to $\mathbf{z}$ in accordance with a predefined cost matrix. Formally, suppose $\mathbf{p} \in \mathbb{R}_+^n$ and $\mathbf{q} \in \mathbb{R}_+^n$ are two discrete distributions of the mass on $\mathbf{x}$ and $\mathbf{z}$, where $\mathbf{p}_i \in \mathbb{R}_+$ and $\mathbf{q}_j \in \mathbb{R}_+$ are the mass at $\mathbf{x}_i \in \mathbb{R}_+^c$ and $\mathbf{z}_j \in \mathbb{R}_+^c$, respectively. There are many different ways to transport the mass from $\mathbf{x}$ to $\mathbf{z}$, here we define a pairwise cost matrix $D = [d_{i,j}] \in \mathbb{R}_+^{n \times n}$ to penalize different movements/transports, where $d_{i,j}$ represents the cost of moving from $\mathbf{x}_i$ to $\mathbf{z}_j$. With these notations, the OT between $\mathbf{x}$ and $\mathbf{z}$ can be defined as the solution for the following optimization problem

$$\min_{X \in \mathbb{R}_+^{n \times n}} \quad \langle D,\, X \rangle := \text{Tr}(D^\mathsf{T} X), \tag{2}$$
$$\text{subject to} \quad X\mathbb{1}_n = \mathbf{p}, X^\mathsf{T}\mathbb{1}_n = \mathbf{q},$$

where $\langle D,\, X \rangle$ denotes the inner product of D and X and Tr denotes the trace of a matrix. The OT problem in Eq. (2) is computationally expensive with cubical complexity, in practice, it can be approximated by the following entropy regularized problem, which has a lower near-quadratic computational complexity [21, 22]:

$$X_\gamma^\star(\mathbf{x}, \mathbf{z}) = \arg \min_{X \in \mathbb{R}_+^{n \times n}} \quad \langle D,\, X \rangle - \gamma E(X), \tag{3}$$
$$\text{subject to} \quad X\mathbb{1}_n = \mathbf{p}, X^\mathsf{T}\mathbb{1}_n = \mathbf{q},$$

where $E(X) = -\sum_i \sum_j h(X_{ij})$, $\gamma \geq 0$, and

$$h(X_{ij}) = \begin{cases} X_{ij} \log X_{ij}, & \text{if } X_{ij} > 0 \\ 0, & \text{otherwise} \end{cases}.$$

When $\gamma = 0$, the entropy regularized problem in Eq. (3) becomes the problem in Eq. (2).

We solve Eq. (3) with the Sinkhorn algorithm [21]. The Sinkhorn algorithm computes the OT from $\mathbf{x}$ to $\mathbf{z}$ using iterative matrix multiplications with the help of $\mathbf{z}$, where $\mathbf{z}$ is unknown beforehand and will be learned during the training process. Therefore, the learned feature set can adapt to the training dataset and match the geometric structure of target features better. The obtained solution of Eq. (3), denoted by $X_\gamma^\star(\mathbf{x}, \mathbf{z})$, is then used to compute the optimal transported feature $\text{OT}_\mathbf{z}^\gamma(\mathbf{x})$,

$$\text{OT}_\mathbf{z}^\gamma(\mathbf{x}) := X_\gamma^\star(\mathbf{x}, \mathbf{z})^\mathsf{T}\mathbf{x}. \tag{4}$$

In this process, the embedded feature set $\mathbf{x}$ is transported to a new feature set $\text{OT}_\mathbf{z}^\gamma(\mathbf{x})$ using $X_\gamma^\star(\mathbf{x}, \mathbf{z})$, which is the solution of the optimization problem in Eq. (3) computed with the learnable reference set $\mathbf{z}$. Since $\mathbf{z}$ is target-distribution-related, the obtained $\text{OT}_\mathbf{z}^\gamma(\mathbf{x})$ is target-distribution-related.

To further stabilize and reduce the training time of iOTKE, we apply LayerNorm [23] to $\text{OT}_\mathbf{z}^\gamma(\mathbf{x})$, i.e.,

$$\mathbf{x}_{\text{OT}} = \text{LayerNorm}(\text{OT}_\mathbf{z}^\gamma(\mathbf{x})), \tag{5}$$

where $\mathbf{x}_{\text{OT}}$ is the obtained features optimally transported from $\mathbf{x}$ in accordance with the target-distribution-related feature set $\mathbf{z}$.

2.2. Feature selection with lightweight gating mechanism

For a specific embedded feature $\mathbf{x}_0$, the features in the transported feature set $\mathbf{x}_{\text{OT}}$ may not all be necessary, since $\mathbf{x}_{\text{OT}}$ is computed based on the reference set $\mathbf{z}$ which is learned from all training data and reflects the whole feature geometry/distribution of the training data. Therefore, we should choose the most important features for $\mathbf{x}_0$ from $\mathbf{x}_{\text{OT}}$.

To do this, we introduce a lightweight gating mechanism to iOTKE. This gating mechanism, as defined in Eq. (6), computes an element-wise Hadamard product between the input feature set $\mathbf{x}_0$ and the activated features from $\mathbf{x}_{\text{OT}}$,

$$\text{G-iOTKE}(\mathbf{x}_0) = \mathbf{x}_0 \otimes \sigma(\mathbf{x}_{\text{OT}}), \tag{6}$$

where $\otimes$ is the Hadamard product and σ is the nonlinear activation function used to activate the features. This gating mechanism, without introducing any trainable parameters, can control effective information flow by enhancing or weakening each feature in accordance with the nonlinear activate function. Intuitively, it enables the proposed G-iOTKE to select features that are important for distinguishing a target signal in SVS/SI. Combining Eq. (1), Eq. (5) and Eq. (6), the proposed G-iOTKE can be formulated as

$$\text{G-iOTKE}(\mathbf{x}_0) = \mathbf{x}_0 \otimes \sigma(\text{LayerNorm}(\text{OT}_\mathbf{z}^\gamma(\text{ReLU}(\mathbf{x}_0)))). \tag{7}$$

2.3. Comparison with existing methods

2.3.1. Comparison with standard OTKE

The iOTKE differs from the standard OTKE [16–18] in two ways.

First, different from the standard OTKE that uses weighted pooling operations to aggregate large features into small ones, the reference feature set and the output feature set in G-iOTKE have the same size as the input feature set. The rich reference features enable G-iOTKE to learn the diversity of the training data which is helpful in maintaining the geometric structure and distribution of the target signal. Note that ensuring the same dimension of input and output feature set is quite important for SVS. This is because in most SVS frameworks, we need to predict and apply a soft or binary mask to the time-frequency representation (e.g., magnitude spectrogram). In this case, the output feature set should be the same size as the input feature set. In particular, since the input and output feature sets are of the same size, the proposed G-iOTKE can be widely applied to many existing DNNs-based audio processing systems as an additional sublayer and it is compatible with many neural structures (e.g., residual skip connection) to improve the performance and accelerate the convergence of DNNs [24].

Second, we design a lightweight gating mechanism for the proposed G-iOTKE. This gating mechanism is effective in improving the performance of G-iOTKE. In addition, we incorporate ReLU and LayerNorm to the G-iOTKE. These two operations are simple but quite effective for feature extraction in SVS/SI. Compared with the standard OTKE, the only increased parameters in G-iOTKE are the bias and gain in LayerNorm, which are quite trivial.

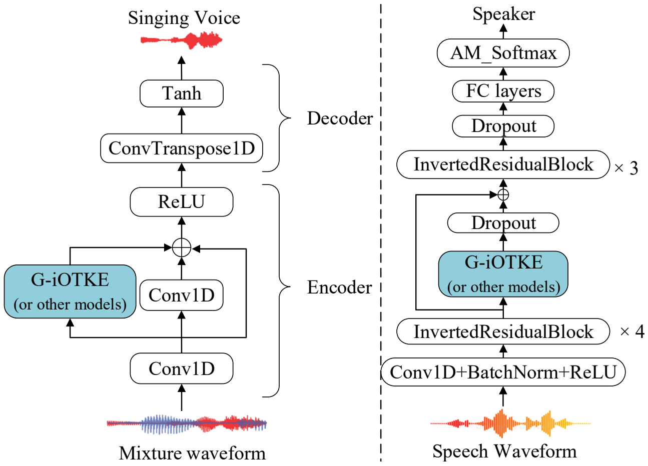

Fig. 1: Frameworks for SVS (left) and SI (right) experiments.

2.3.2. Comparison with GLU

The classic gating mechanism GLU [11] is achieved by two element-wise linear transformations (see Eq. (10)). Compared with GLU, the proposed G-iOTKE realizes better feature extracting (embedding) without using two heavyweight linear projections. Accordingly, it provides better embedding with far fewer parameters.

3. EVALUATIONS

We evaluated the proposed G-iOTKE for two signal processing applications: SVS and SI. To verify the effectiveness of G-iOTKE, we compared it with eight other models, including two OTKE based models, two GLU based models, and four gated models. These compared models are detailed as follows.

The OTKE based models are (i) the standard OTKE (denoted by OTKE) and (ii) the standard OTKE with a ReLU function (OTKE_R)

$$\text{OTKE}(\mathbf{x}_0) = \text{OT}^{\gamma}_{\mathbf{z}}(\mathbf{x}_0), \tag{8}$$

$$\text{OTKE}_R(\mathbf{x}_0) = \text{OT}^{\gamma}_{\mathbf{z}}(\text{ReLU}(\mathbf{x}_0)), \tag{9}$$

where ReLU in OTKE_R ensures the non-negativity of the input distribution. Please note that these two OTKE models have input and output features that are the same size, so that they can be fairly compared with the proposed G-iOTKE.

The GLU based models are (i) the standard GLU (GLU) [11] and (ii) the simplified GLU (GLU_s):

$$\text{GLU}(\mathbf{x}_0) = (\mathbf{x}_0\mathbf{V} + \mathbf{c}) \otimes \sigma(\mathbf{x}_0\mathbf{W} + \mathbf{b}), \tag{10}$$

$$\text{GLU}_s(\mathbf{x}_0) = \mathbf{x}_0 \otimes \sigma(\mathbf{x}_0\mathbf{W} + \mathbf{b}). \tag{11}$$

where the standard GLU has two linear element-wise projections $\mathbf{W}$ and $\mathbf{V}$, and $\mathbf{b}$ and $\mathbf{c}$ are bias vectors. The simplified GLU_s, as defined in Eq. (11), uses only one linear projection.

The four gated models are defined as follows:

$$\text{G}_0(\mathbf{x}_0) = \mathbf{x}_0 \otimes \mathbf{z}, \tag{12}$$

$$\text{G}_1(\mathbf{x}_0) = \mathbf{x}_0 \otimes \sigma(\text{OT}^{\gamma}_{\mathbf{z}}(\text{ReLU}(\mathbf{x}_0))), \tag{13}$$

$$\text{G}_2(\mathbf{x}_0) = \mathbf{x}_0 \otimes \sigma(\text{LayerNorm}(\text{OT}^{\gamma}_{\mathbf{z}}(\mathbf{x}_0))), \tag{14}$$

$$\text{G}_3(\mathbf{x}_0) = \mathbf{x}_0 \otimes \sigma(\text{OT}^{\gamma}_{\mathbf{z}}(\mathbf{x}_0)), \tag{15}$$

where $\mathbf{z}$ in Eq. (12) is a simple mask (filter) learned from the training data. Thus the process of G_0 can be considered as applying an adaptive filtering operation to $\mathbf{x}_0$. The models defined in Eqs. (13)-(15), as compared with Eq. (7), are variations of the proposed G-iOTKE without using LayerNorm, or ReLU, or any of them.

We implemented the proposed G-iOTKE and other models in two representative SVS [20] and SI [10] frameworks. As shown in Fig. 1, the G-iOTKE (or other compared models) is used as additional sublayers in the baseline SVS framework (left) and the SI framework (right). All models were compared under the same experimental setting to enable a fair comparison.

Table 1: The SVS performance of all models on MUSDB18, where c is the channel numbers for each feature (i.e., frequency resolution).

Model	**SI-SDR-BM (dB)**		
	c=400	c=800	c=1600
Baseline (B)	5.93	6.28	6.68
B+OTKE	1.91	2.41	2.71
B+OTKE_R	4.45	4.94	5.53
B+GLU	1.28	0.77	0.92
B+GLU_s	0.28	0.49	0.50
B+G_0	4.24	4.61	5.22
B+G_1	6.14	6.81	**7.41**
B+G_2	1.80	2.21	2.66
B+G_3	1.70	2.33	2.50
B+G-iOTKE	**6.15**	**6.85**	7.31

3.1. Unsupervised representation learning for SVS

The baseline SVS framework we used was an unsupervised auto-encoder model proposed in [20]. The encoder in this framework consists of two one-dimensional strided convolutions with appropriate zero-padding. We added the proposed G-iOTKE and other models to the encoder of the baseline framework to verify their effectiveness (see the left part in Fig. 1). These models are termed B (Baseline)+G-iOTKE, B+OTKE/OTKE_R, B+GLU/GLU_s, and B+$\text{G}_{0/1/2/3}$.

The database we used was the MUSDB18 [25], which is made of 150 two-channel multi-tracks signals sampled at 44.1 kHz (100 for training and 50 for testing). Each multi-track includes the vocal and accompaniment sources. In accordance with [20], all models worked in three frequency resolutions: $c = 400/800/1600$, where c is the frequency resolution. The performances of all models were measured by SI-SDR-BM, which computes the reconstruction error of vocal upon a learned representation using informed binary masking (BM). More details about this measurement can be found in [20].

3.1.1. Comparison of SVS performance

The SVS performances of all models are listed in Table 1. We can see that the two OTKE based models (B+OTKE and B+OTKE_R) do not improve the baseline model but degrade its performance. This result suggests that simply introducing the OTKE to the baseline framework is not useful to improve the SVS performance. It is also noticed that the OTKE_R provided better performance than the OTKE, which means that ensuring the non-negativity of the input distribution is useful for SVS. When comparing the above two OTKE based models with the proposed G-iOTKE, we can find that the proposed G-iOTKE significantly improved the SVS performance, which means that the gating mechanism is quite helpful in improving the SVS performance.

In addition, we can see that the GLU/GLU_s and $\text{G}_{2/3}$ had much lower performance than the baseline model, that is, these models do not work well in the baseline framework. As seen in Eqs. (14) and (15), the G2 and G3 models did not use ReLU before OT, i.e., they do not ensure the non-negativity of representations. Since the baseline SVS framework tries to learn non-negative representations, the lack of ReLU in G2 and G3 leads to performance degradation in these two models.

In particular, we can see that the G_1 model provided slightly lower performance than the proposed G-iOTKE for $c = 400/800$, and it had much better performance than $\text{G}_{0/2/3}$, which indicates that the ReLU and gating mechanism play essential roles in the proposed G-iOTKE to improve the SVS performance. It is also noticed that the G-iOTKE was slightly worse than G_1 for c=1600, that is, the

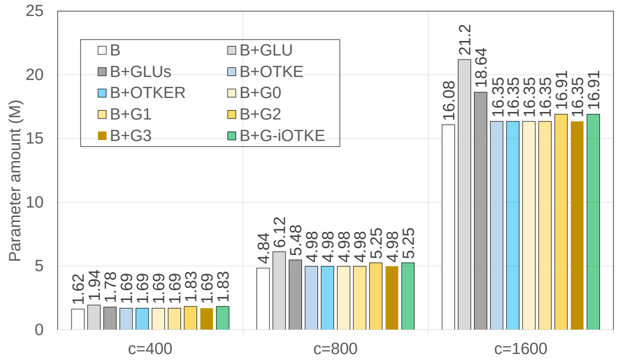

Fig. 2: The trainable parameter amount of different models in SVS.

LayerNorm in G-iOTKE did not provide improvement for c=1600. The reason might be that the learnable parameters in LayerNorm (including the bias and gain) increase the risk of over-fitting for the high resolution case of c=1600.

3.1.2. Comparison of parameter amount

We compared the parameter amount of all models. As shown in Fig. 2, the GLU based models (B+GLU and B+GLU$_s$) had a much larger parameter amount than other models, especially for $c = 1600$. This is because they used two linear projections. The OTKE based models (B+OTKE and B+OTKE$_R$) and the G$_{0/1/3}$ based models (B+G$_{0/1/3}$) had the same parameter amount. The G$_2$ based model (B+G$_2$) had the same parameter amount as G-iOTKE (B+G-iOTKE), as they both used LayerNorm. However, since G$_2$ did not use ReLU, its performance was much lower than that of G-iOTKE. When comparing the OTKE based models (B+OTKE and B+OTKE$_R$) with the proposed G-iOTKE (B+G-iOTKE), we can see that the proposed G-iOTKE only slightly increased the parameter amount.

3.2. Speaker identification (SI) on mobile devices

Speaker identification is a fundamental technology for many mobile-device applications, such as automation, authentication, and security [10]. These application scenarios require lightweight DNNs to reduce the storage size. To address this issue, Nunes et al. proposed a portable model for SI, called Additive Margin MobileNet1D (AM-MobileNet1D) [10]. We adopted this model as the baseline SI framework to evaluate different models (see the right part in Fig. 1).

The AM-MobileNet1D combines MobileNet1D [10, 26, 27] and additive margin softmax (AM-Softmax) loss function [9, 28]. We added the proposed G-iOTKE and other models to MobileNet1D and kept all other components intact. The dataset we used was the well-known TIMIT [29]. We followed the protocol and setting in [10] to train all models. The performance of SI was measured by the Frame Error Rate (FER) and the Classification Error Rate (CER) [10].

The FER and CER results of all models are listed in Table 2, which also lists the results of some representative SI methods in [8, 16, 30–32] for comparison (see the first seven rows). The baseline AM-MobileNet1D provided lower FER than the existing SI methods [8, 16, 30–32]. The OTKE based AM-MobileNet1D slightly degraded the baseline AM-MobileNet1D for FER, whereas the OTKE$_R$ slightly improved the FER of the baseline AM-MobileNet1D. In contrast, both GLU/GLU$_s$ could improve the baseline AM-MobileNet1D. The G$_2$ and G$_3$ based AM-MobileNet1D also slightly improved the FER compared with the baseline AM-MobileNet1D, whereas G$_0$ and G$_1$ had degraded performance. By comparing all FER results, we can see that the G-iOTKE based AM-MobileNet1D achieved the best result for FER. For CER, none of the compared models, including OTKE/OTKE$_R$, GLU/GLU$_s$, and G$_0$/G$_1$/G$_2$/G$_3$ could improve the baseline AM-MobileNet1D. The best CER result (0.28 dB) was obtained by AF-SincNet [31]. The proposed G-iOTKE achieved quite similar performance to this result. These results verified the effectiveness of the proposed G-iOTKE.

Table 2: Speaker identification results on TIMIT.

Model	**FER(%)**	**CER(%)**
SincNet [8]	47.38	1.08
AM-SincNet [30]	28.09	0.36
AF-SincNet [31]	26.90	**0.28**
Ensemble-SincNet [31]	35.98	0.79
ALL-SincNet [31]	36.08	0.72
CL-SincNet [32]	37.36	1.08
MobileNet1D [16]	26.50	0.57
AM-MobileNet1D [16]	21.30	0.43
OTKE-AM-MobileNet1D	22.41	0.94
OTKE$_R$-AM-MobileNet1D	20.52	0.65
GLU-AM-MobileNet1D	20.05	0.58
GLU$_s$-AM-MobileNet1D	20.97	0.58
G$_0$-AM-MobileNet1D	22.61	0.65
G$_1$-AM-MobileNet1D	21.64	0.79
G$_2$-AM-MobileNet1D	20.59	0.65
G$_3$-AM-MobileNet1D	21.03	0.58
G-iOTKE-AM-MobileNet1D	**19.74**	0.29

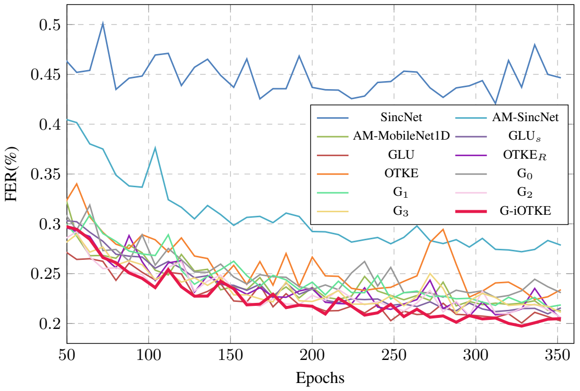

Fig. 3: The evolution of FER for all models over the training epoch.

Finally, we showed the FER evolution of all models over the training epoch. As shown in Fig. 3, the proposed G-iOTKE had lower FERs than other comparison models for most epochs.

4. CONCLUSIONS

This work introduced optimal transport (OT) to singing voice separation (SVS) and speaker identification (SI) and proposed the gated improved optimal transport kernel embedding (G-iOTKE) to learn effective representations of the target signal. The proposed G-iOTKE includes two steps: (1) learning target-distribution-related features with iOTKE and (2) selecting features with a lightweight gating mechanism. The iOTKE extracts latent geometric-aware and distribution-related features of a target signal with optimal transport on the basis of a reference set learned from all training data. Therefore, the output features of iOTKE can maintain the diversity of the input data and preserve the original geometric structure of the target signal. For a specific input signal, we use the gating mechanism to control the information flow and select important features relevant to it. Experimental results showed that the proposed G-iOTKE provided better results in both SVS and SI experiments than other models. In the future, we will compare the proposed G-iOTKE with more state of the art SVS/SI methods.

5. REFERENCES

[1] Y. Li and D. Wang, “Separation of singing voice from music accompaniment for monaural recordings,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, 2007.

[2] Z. Rafii, A. Liutkus, F. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, “An overview of lead and accompaniment separation in music,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 8, pp. 1307–1335, 2018.

[3] Y. Zhang, Y. Liu, and D. Wang, “Complex ratio masking for singing voice separation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 41–45.

[4] S. I. Mimilakis, K. Drossos, E. Cano, and G. Schuller, “Examining the mapping functions of denoising autoencoders in singing voice separation,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 266–278, 2020.

[5] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, “Phoneme level lyrics alignment and text-informed singing voice separation,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2382–2395, 2021.

[6] X. Ni and J. Ren, “Fc-u^2-net: A novel deep neural network for singing voice separation,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 489–494, 2022.

[7] S. Yuan, Z. Wang, U. Isik, R. Giri, J. Valin, M. M. Goodwin, and A. Krishnaswamy, “Improved singing voice separation with chromagram-based pitch-aware remixing,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 111–115.

[8] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with sincnet,” in *Spoken Lang. Tech. Workshop*, 2018, pp. 1021–1028.

[9] J. A. C. Nunes, D. Macêdo, and C. Zanchettin, “Additive margin sincnet for speaker recognition,” in *Int. Joint Conf. on Neural Netw.*, 2019, pp. 1–5.

[10] J. A. C. Nunes, D. Macedo, and C. Zanchettin, “Am-mobilenet1d: A portable model for speaker recognition,” in *Int. Joint Conf. on Neural Netw.*, 2020, pp. 1–8.

[11] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *Proc. 34th Int. Conf. Machine Learning*, 2017, pp. 933–941.

[12] G. Peyré and M. Cuturi, “Computational optimal transport,” *Found. Trends Mach. Learn.*, vol. 11, no. 5-6, pp. 355–607, 2019.

[13] A. Rolet and V. Seguy, “Fast optimal transport regularized projection and application to coefficient shrinkage and filtering,” *Vis. Comput.*, pp. 1–15, 2021.

[14] M. A. Schmitz, M. Heitz, N. Bonneel, F. M. N. Mboula, D. Coeurjolly, M. Cuturi, G. Peyré, and J. Starck, “Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning,” *SIAM J. Imaging Sci.*, vol. 11, no. 1, pp. 643–678, 2018.

[15] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. J. Guibas, “Convolutional wasserstein distances: efficient optimal transportation on geometric domains,” *ACM Trans. Graph.*, vol. 34, no. 4, pp. 66:1–66:11, 2015.

[16] G. Mialon, D. Chen, A. d’Aspremont, and J. Mairal, “A trainable optimal transport embedding for feature aggregation and its relationship to attention,” in *Proc. 9th Int. Conf. Learn. Represent.*, 2021.

[17] X. Wei, Y. Gong, F. Wang, X. Sun, and J. Sun, “Learning canonical view representation for 3d shape recognition with arbitrary views,” in *Int. Conf. Comput. Vis.*, 2021, pp. 397–406.

[18] Y. Tian, J. Li, and T. Lee, “Transport-oriented feature aggregation for speaker embedding learning,” in *Interspeech*, 2022, pp. 316–320.

[19] P. Smaragdis and S. Venkataramani, “A neural network alternative to non-negative audio models,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 86–90.

[20] S. I. Mimilakis, K. Drossos, and G. Schuller, “Unsupervised interpretable representation learning for singing voice separation,” in *Proc. 28th Eur. Signal Process. Conf.*, 2020, pp. 1412–1416.

[21] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Proc. Adv. Neural Inf. Process. Sys.*, 2013, pp. 2292–2300.

[22] J. M. Altschuler, J. Weed, and P. Rigollet, “Near-linear time approximation algorithms for optimal transport via sinkhorn iteration,” in *Proc. Adv. Neural Inf. Process. Sys.*, 2017, pp. 1964–1974.

[23] L. J. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *CoRR*, vol. abs/1607.06450, 2016.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[25] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “The MUSDB18 corpus for music separation,” Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[26] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *CoRR*, vol. abs/1704.04861, 2017.

[27] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *IEEE Conf. Compu. Vis. Pattern Recog.*, 2018, pp. 4510–4520.

[28] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification,” *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, 2018.

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” p. 27403, 1993. [Online]. Available: https://ui.adsabs.harvard.edu/abs/1993STIN...9327403G

[30] J. A. Chagas Nunes, D. Macêdo, and C. Zanchettin, “Additive margin sincnet for speaker recognition,” in *Int. Joint Conf. Neural Netw.*, 2019, pp. 1–5.

[31] L. Chowdhury, H. Zunair, and N. Mohammed, “Robust deep speaker recognition: Learning latent representation with joint angular margin loss,” *Appl. Sci.*, vol. 10, no. 21, 2020.

[32] L. Chowdhury, M. Kamal, N. Hasan, and N. Mohammed, “Curricular sincnet: Towards robust deep speaker recognition by emphasizing hard samples in latent space,” in *Int. Conf. Biometrics Special Interest Group*, 2021, pp. 1–4.