Impact of Minimum Hyperspherical Energy Regularization on Time-Frequency Domain Networks for Singing Voice Separation

Neil Shah and Dharmeshkumar M Agrawal
TCS Research, Tata Consultancy Services Pvt. Ltd., Pune, India
E-mail: {neilkumar.shah, dharmesh.agrawal}@tcs.com, Tel: +91-9106176744

***Abstract*—The task of singing voice separation requires the model to maintain a trade-off between signal quality, interference introduced by music accompaniment and algorithmic artifacts. A time domain-based singing voice separation system offers a challenge in designing for low latency and in minimizing computational cost. To overcome this problem, we propose to use Gammatone auditory features for the Time-Frequency (T-F) mask-based singing voice separation task. Minimum Hyperspherical Energy (MHE) regularization in the time-domain network has recently produced the state-of-the-art result in singing voice separation (our baseline). In this work, we apply MHE to the T-F domain networks. The MHE regularized T-F domain network significantly improves the separation performance over the baseline. The MHE regularized Wasserstein Generative Adversarial Network (GAN) achieves 0.21 dB improvement in mean Signal-to-Distortion Ratio (SDR) over the baseline. Our best performing T-F domain un-regularized GAN provides an improvement of 0.75 dB and 0.63 dB in SDR over the baseline and the GAN-MHE, respectively. We experimentally show the failure of MHE regularized T-F domain networks with respect to their un-regularized versions and have shown the need of designing a suitable adversarial objective function. We report that modifying the GAN-MHE's objective function with reconstruction loss and adapting Wasserstein GAN, results in a 0.45 dB improvement in mean SDR over its un-regularized version.**

***Index Terms*—Singing voice separation, Minimum Hyperspherical Energy (MHE), Generative Adversarial Networks (GAN).**

I. Introduction

The ability of human listeners in identifying perceptual differences between musical sources has inspired the recent research in the area of automatic music source separation. Music source separation finds its application in automatic lyrics transcription, artist identification, karaoke generation, editing music, and also enhances pitch estimation and chord recognition results [1], [2]. The separation of singing voice involves the extraction of the main melody from the musical accompaniment [1].

Several studies have been proposed in monaural music separation [1], [3]–[6]. The availability of a single channel in monaural music makes the problem ill-posed in nature, since for an estimated source, there could be multiple possible solutions for other sources [7]. The traditional approach imposes a strong assumption on the data to be low-rank and extracted from a sparse subspace [8]–[10]. However, the samples drawn from a sparse subspace might not always satisfy the low-rank condition. This approximation may not lead to the exact recovery of data and also lacks generalization [11]. The Bayesian method requires the source model to possess statistical signal properties, and therefore the model adaptation of maximum *a posteriori* criterion becomes a necessity [7]. Non-negative matrix factorization, though effective in extracting a perceptually meaningful source, requires an additive transformation for the negative data [5], [12].

Recently, deep learning techniques have surpassed the traditional methods in achieving a state-of-the-art performance in audio source separation tasks [13]–[18]. Huang et al. jointly optimized a soft mask function with the network's objective, for singing voice separation in monaural recordings [2]. The convolutional neural network proposed by Simpson et al. estimates an ideal binary mask (IBM) for automatic removal of vocal from musical mixture for karaoke applications [6]. Nugraha et al. experimentally demonstrated that source spectra estimated by a single DNN can outperform the spectra estimated by other traditional approaches [19]. The U-Net architecture adapted by Jansson et al. claims to reconstruct voice typically found in the commercial pop music [20]. The Wave-U-Net architecture investigated by Stoller et al. separates sources in the time-domain [3]. A recent work using dense neural network explores long-term dependencies in the audio spectrogram [13]. However, the use of higher-dimensional spectrogram features and fine-tuning makes the real-time deployment non-viable and also increases the computational complexity [3], [13], [20].

Most of the studies mentioned above use a mask-based estimation technique since it offers direct access to components in both time and frequency. Moreover, in the Time-Frequency (T-F) domain, the audio sources are weakly overlapping, which yields an optimal solution in the weighted least square sense [21]. Supervised music source separation aims at estimating a T-F mask from the music mixture [20], [22], [23]. Currently, such an approach aims at optimizing the Maximum Likelihood (ML)-based algorithm to estimate the source while inherently predicting a mask. However, the ML-based criterion puts a prior assumption on the data distribution, which may not be globally valid for all the data samples and may force the model to learn non-optimal network parameters [24]. However, the adversarial optimization proposed in [16], [17], interprets the source separation as an inference problem by finding the source estimates from the prior generative model. Moreover, adversarial optimization does not put any parametric assumption on the output data distribution. The study by Sisman et al. has reported a state-of-the-art result obtained using adversarial optimization techniques in the singing voice conversion task [25].

The regularization technique is known to improve the representational power of a network. Improving the effectiveness and generalizability of a model without reducing the network parameters has proven to minimize the redundancy among the neurons. The inability of softmax loss in producing more diverse intra-class samples generates a large number of correlated and redundant neurons, which lead to over-fitting and reduce the network's generalizability. The large-margin softmax loss introduced by Liu et al., preserves the most abundant and discriminating information by making different classification boundary for each class [26]. The recently proposed Minimum Hyperspherical Energy (MHE) regularizer reduces the redundancy by minimizing the hyperspherical energy between the neurons in each layer [27]. The study by Perez-Lapillo et al. explored the use of MHE for the singing voice separation in a Wave-U-Net architecture [1]. The inclusion of MHE in the loss function has provided a state-of-the-art system for the time-domain singing voice separation task.

Inspired by this work, we study the impact of MHE along with the auditory features for the T-F domain-based signing voice separation task. In this study, we have used Generative Adversarial Network (GAN)-based models instead of Wave-U-Net architecture to perform singing voice separation by implicitly learning the T-F mask. Using the above, we outperformed the state-of-the-art results established by the time-domain MHE regularized system (our baseline) on the MUSDB18 dataset [1]. With MHE regularization, the Wasserstein GAN obtains an improvement of 0.21 dB over the baseline. We also provide a comparison between the MHE regularized T-F domain networks and their un-regularized versions. We show how the separation quality can be improved by using a reconstruction loss and the Wasserstein architecture to the MHE regularized adversarial network. We also present the analysis of our best performing un-regularized GAN over the baseline and its MHE regularized version. The paper is organized as follow: section 2 describes Gammatone features and optimizing GAN and their variants with MHE for T-F based singing voice separation, section 3 describes the dataset and network architecture, section 4 discusses evaluation results, followed by section 5 discussing conclusions and future work.

II. Research Methodology

A. Gammatone auditory feature representation

The Gammatone filter is inspired from biologically motivated studies [28] and is used for modeling the human auditory filter response [29]. The bandwidth of a Gammatone filter corresponds to the placement of filters in the basilar membrane of the human auditory system. The impulse response of a Gammatone filter is a multiplication of gamma distribution and a sinusoidal tone centered at a particular frequency [30]. It is mathematically formulated as:

$$g(f,t) = t^{a-1}e^{-2\pi bt}\cos(2\pi ft), t > 0, \tag{1}$$

where a is the filter order, b is the rectangular bandwidth, and f is the center frequency.

B. GAN For T-F based Singing Voice Separation

The GAN learns a mapping between the mixture samples y following a prior distribution $\mathcal{Y}$ and vocal samples x belonging to data distribution $\mathcal{X}$. The Generator (G) aims to learn the data distribution in an adversarial framework. The Discriminator (D) being a binary classifier maximizes the likelihood of vocal samples drawn from $\mathcal{X}$ as real and minimizes the likelihood of the predicted vocal samples drawn from the model distribution $\hat{\mathcal{X}}$ (output of G) as fake. As training optimality $\hat{\mathcal{X}} \rightarrow \mathcal{X}$ is achieved, the G network is supposed to generate realistic vocal samples and the D network is left confused between $\mathcal{X}$ and $\hat{\mathcal{X}}$. This objective function can be defined as [31]:

$$\begin{split}\min_G \max_D V(D,G) = \mathbb{E}_{x\sim\mathcal{X}}[\log D(x)]+ \\ \mathbb{E}_{y\sim\mathcal{Y}}[\log(1-D(G(y))],\end{split} \tag{2}$$

where $\mathbb{E}_{x\sim\mathcal{X}}$ is the expectation taken over the samples x drawn from the data distribution $\mathcal{X}$.

The GAN can be employed for the T-F mask prediction, where the G network is trained to predict vocal T-F representation while learning the mask implicitly (the method is commonly known as task-dependent masking [32]). Also, many recent studies have proven the ability of discriminators in classifying the real from the generated samples [16], [17]. Motivated from [24], [32], [33], we propose to optimize Minimum Mean Square Error (MMSE) between the log T-F representation of vocals and the predicted ones. If the output of the G network is m, input music mixture T-F representation is t, and vocal T-F representation is v, then the optimization function can be written as:

$$MMSE = \frac{1}{2}||\log(t \odot m) - \log(v)||^2, \tag{3}$$

where '$\odot$' represents element-wise multiplication and $MMSE$ is the objective function to be optimized. Thus, the vocal representation can be estimated when the music mixture representation is element-wise multiplied with the T-F mask estimated by the G network. If the sources do not overlap at all, then

$$\begin{split}\prod_i S_i(t,f) = 0, \forall t, f, \\ \hat{S_j}(t,f) = t_j(t,f)\cdot m_j(t,f), \forall t, f,\end{split} \tag{4}$$

where i is the number of sources present in the music mixture, $S_j(t,f)$ denotes a particular T-F unit of a j^{th} source, and $\hat{S_j}(t,f)$ denotes a T-F unit of an estimated j^{th} source. If $S_j(t,f) \neq 0$, then making $m_j(t,f) = 1$ would yield an

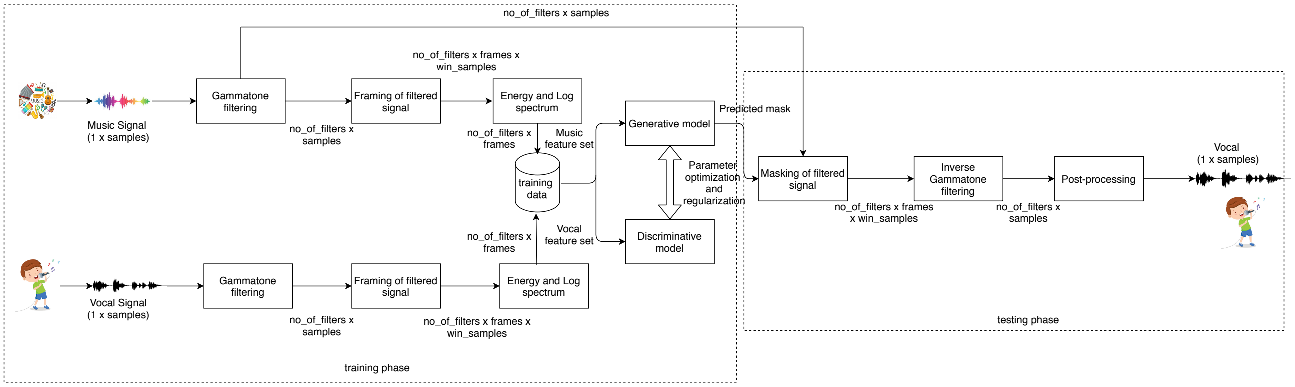

Fig. 1. Block diagram for the prediction of T-F mask-based network. During training the GANs are trained with the extracted Gammatone music signal features as an input and ground-truth vocal features as the target output. During testing the model estimates the T-F mask, generates the Gammatone spectrum and reconstructs the estimated vocal signal.

ideal solution for the j^{th} source. However, some presence of overlapping is always observed in the real-time music mixture. Therefore, restricting the mask to take continuous values between $[0,1]$ approximates it to the Ideal Ratio Mask (IRM) and may yield an optimal solution in a weighted square sense.

Regularizing the G network with $MMSE$ loss prevents the generator from learning representation which may not correspond to the given music mixture at the input, but may belong to the distribution of vocal T-F data representation ($\mathcal{X}$) [24], [33]. The G network in the original GAN objective function (eq: 2) can be $MMSE$ regularized as [24]:

$$\tilde{V}(D,G) = \min_{G}\max_{D} V(D,G) + \\ \frac{1}{2}\min_{G} \mathbb{E}_{x\sim\mathcal{X},y\sim\mathcal{Y}}[\log(x) - \log(G(y)]. \tag{5}$$

C. Optimizing GANs with MHE

There are two alternative configurations of MHE, full-space MHE and half-space MHE. The half-space MHE encourages the neurons in the hidden layer to be less correlated and less redundant [27] than the full-space MHE. However, we prefer to use full-space MHE for our evaluation, as it has shown to obtain better results for singing voice separation [1]. The MHE regularized GANs objective function can be formulated as [27], [31]:

$$J = \tilde{V}(D,G) + \lambda_h \cdot \sum_{j=1}^{L-1} \frac{1}{N_j(N_j-1)} E_{sj}, \tag{6}$$

where λ_h is the weighting parameter, L is the number of hidden layers, N_j is the number of neurons in the j^{th} layer, and E_{sj} is the hyperspherical energy of the neurons in the j^{th} layer and is defined as:

$$E_{sj} = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} f_s(||\hat{w}_i - \hat{w}_j||), \tag{7}$$

where $||.||$ is an Euclidean distance, $\hat{w}_i$ is the weight of an i^{th} neuron projected on the unit hypersphere. The weighting constant of the MHE should be a constant depending on the number of hidden layers [1], [27]. As suggested in [27], we use $f_s = z^{-s}$, a decreasing real-valued Rises s-kernel function, $\lambda_h = \frac{1}{L}$, and $s = 0$.

D. Variants to the GAN's adversarial loss

The vanilla GAN architecture proposed by Goodfellow et al. [31] often suffers from stability issues during training [34], [35]. The discriminator loss quickly converges to zero, and thereby quickly differentiates between the generated and the clean audio samples at its input. Thus, the G network will not be able to generate a perceptually audible clean singing voice. The Wasserstein GAN (WGAN) proposed in [34], uses a continuous distance measure between the distributions and is differentiable almost everywhere. The WGAN updates its G network once, for a certain number of its D network updates. This allows the D network to optimally update its parameters and also provides a good correlation between the G network loss and its generated sample quality.

Adding a reconstruction loss to the Generator's objective function, allows the network to learn a relevant mapping between the generated output and its input. The reconstruction loss compares the generated output with its input mixture and in this process may rectify the mode collapse problem as observed in vanilla GANs [36]. Here we analyze GAN, WGAN, GAN with reconstruction loss (RECON-GAN) and RECON-WGAN for the singing voice separation task as the variants to GAN.

III. Experimental Setup

A. Database

We used the MUSDB18 dataset for model training and evaluation [37][1]. The data consists of 100 song clips for training and 50 for testing, with duration of 6.5 and 3.5 hours,

[1] https://sigsep.github.io/datasets/musdb.html

respectively. The training set is further divided by randomly selecting 20 tracks for validation purpose. The bass, drums, and other sources are mixed to constitute as an accompaniment. We also use CCMixter dataset [38] comprising of 50 more clips for training, as in [1].

B. Network Architecture

The networks are trained in two different configurations. The first configuration consists of an unregularized DNN, GAN, WGAN, RECON-GAN, and RECON-WGAN architectures. The DNN is optimized using *MMSE* criteria between the estimated vocal and the extracted vocal from the musical mixture. The GAN, WGAN, RECON-GAN, and RECON-WGAN have their G network identical to DNN, each with 3 hidden layers and 1000 neurons with Rectified Linear Unit (ReLU) activation. 60 neurons in the output layer predict the T-F mask implicitly with sigmoid activation. The D network also consists of 3 hidden layers with 512 neurons each, along with a tanh activation and a single unit output layer with sigmoid activation.

In the second configuration, all the five networks are regularized with MHE loss function, namely, DNN-MHE, GAN-MHE, WGAN-MHE, RECON-GAN-MHE, and RECON-WGAN-MHE. All the networks are trained for 150 epochs, with a learning rate of 0.0001, using the Adam optimizer [39] with decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and batch size of 1000.

Fig. 1 shows the block diagram of the proposed system. During training, the Gammatone features extracted from the music mixture and vocal are used to learn the network weights for the T-F mask prediction in the testing phase. The 60-channel Gammatone features are computed with 25 ms Hamming window length and overlap of 10 ms between the consecutive frames. A context length of 4 frames (2 left and 2 right) is used for the training purpose. The networks are optimized to predict the log-Gammatone spectral T-F representation, while implicitly learning the T-F mask at the network's output layer.

IV. Evaluation

The statistical and perceptual analysis of the estimated vocal is performed using Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), and Source-to-Artifacts Ratio (SAR), through a BSS-EVAL evaluation toolbox [40]. To maintain similarity in the evaluation as done in [1], we also partitioned the audio tracks into many non-overlapping segments of one-second length and calculate the metrics for each segment. The resulting metric values are then averaged over the songs and the entire test dataset.

Since the evaluation framework evaluates two estimated sources at-a-time, the music accompaniment is estimated by deducting the estimated vocal from the original music mixture, and is used as a second source. The higher values of the metric suggests better separation quality. The SDR reflects the overall separation improvement, SIR represents the amount of suppression of the interfering sources and SAR signifies the number of artifacts introduced due to the suppression

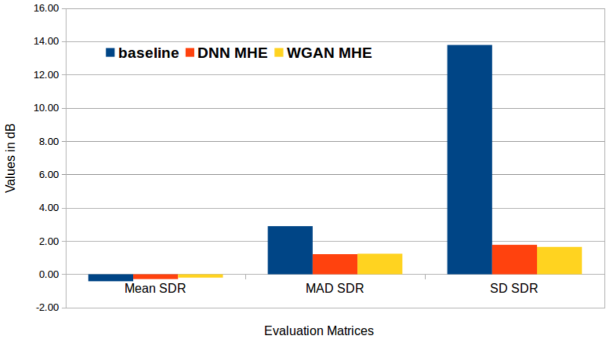

Fig. 2. Impact of MHE on time and T-F domain networks: 1. with baseline time domain MHE system, 2. with MHE regularized T-F domain DNN, 3. with MHE regularized T-F domain WGAN.

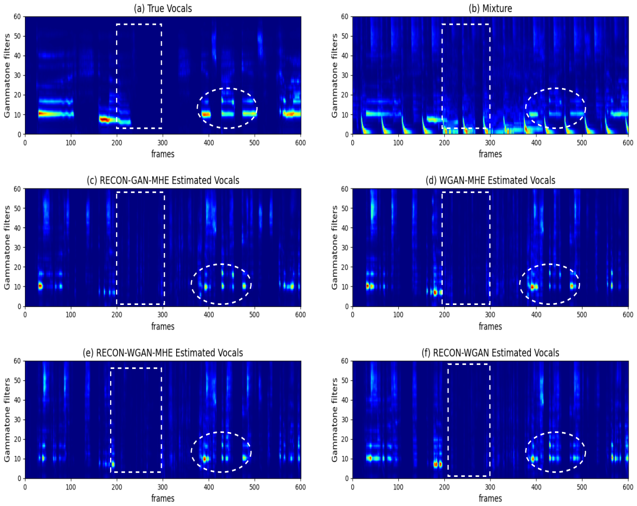

Fig. 3. RECON-WGAN-MHE improves signal quality, reduces interference and removes algorithm artifacts. Gammatone spectrum of: a. true vocals, b: mixture, c: RECON-GAN-MHE, d: WGAN-MHE, e: RECON-WGAN-MHE, and f: RECON-WGAN.

algorithm. The evaluation result weighted over the length of the entire test database is reported in terms of Global SDR (GSDR), and Global SIR (GSIR), Global SAR (GSAR). Since the SDR values are not normally distributed over the entire test database, we also evaluate the Median Absolute Deviation (MAD) of the estimated SDR. The MAD can be calculated by taking the median of absolute deviation from the overall median calculated on the estimated SDR values. We also report the Global Normalized SDR (NSDR) values, to find the improvement of the estimated singing voice over the music mixture. The NSDR is defined as [2]:

$$NSDR(\hat{x}, x, y) = SDR(\hat{x}, x) - SDR(y, x), \qquad (8)$$

where $\hat{x}$ is the estimated singing voice, x is the clean singing voice, and y is the music mixture. The Gammatone spectrum evaluation in Fig. 3, 5, is performed on *Bobby Nobody - Stitch*

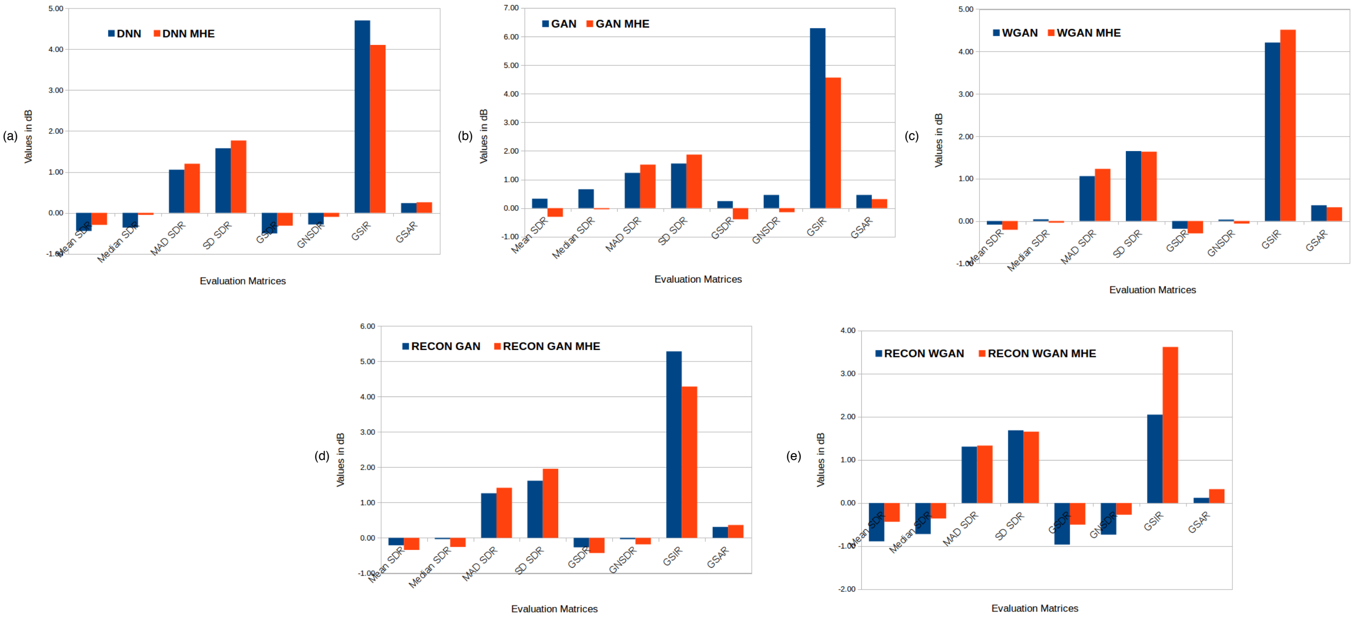

Fig. 4. Impact of MHE on various T-F domain networks: a. DNN and DNN-MHE, b. GAN and GAN-MHE, c. WGAN and WGAN-MHE, d. RECON-GAN and RECON-GAN-MHE, e. RECON-WGAN and RECON-WGAN-MHE. For mean SDR, median SDR, GSIR, GSAR, GSDR, and GNSDR higher is better. For MAD SDR and SD SDR, lower is better.

TABLE I
PERFORMANCE COMPARISONS BETWEEN THE BASELINE, DNN, GAN, WGAN, RECON-GAN, RECON-WGAN, DNN-MHE, GAN-MHE, WGAN-MHE, RECON-GAN-MHE, RECON-WGAN-MHE.

Metrics (in dB)	baseline	DNN	GAN	WGAN	RECON-GAN	RECON-WGAN	DNN-MHE	GAN-MHE	WGAN-MHE	RECON-GAN-MHE	RECON-WGAN-MHE
Mean SDR	-0.42	-0.44	**0.33**	-0.09	-0.21	-0.89	-0.29	-0.30	-0.21	-0.34	-0.44
Median SDR	3.58	-0.36	**0.66**	0.04	-0.04	-0.72	-0.05	-0.04	-0.04	-0.26	-0.36
MAD SDR	2.89	**1.05**	1.23	1.06	1.26	1.31	1.20	1.52	1.23	1.42	1.33
SD SDR	13.78	1.58	**1.56**	1.65	1.62	1.69	1.77	1.87	1.64	1.96	1.65
GSDR	-	-0.50	**0.24**	-0.18	-0.27	-0.97	-0.31	-0.39	-0.29	-0.43	-0.50
GNSDR	-	-0.28	**0.46**	0.03	-0.04	-0.73	-0.10	-0.14	-0.06	-0.18	-0.27
GSIR	-	4.70	**6.29**	4.21	5.29	2.05	4.10	4.56	4.51	4.29	3.62
GSAR	-	0.24	**0.46**	0.37	0.31	0.12	0.26	0.31	0.32	0.36	0.32

Up wave file from the test dataset [37].

A. Impact of MHE on time and T-F domain networks

To evaluate the performance of MHE regularization on T-F domain networks, we consider the objective scores of MHE regularized time-domain Wave-U-Net architecture [1] as our baseline. Fig. 2 shows a comparison of the MHE regularized T-F models that achieved maximum improvement in mean SDR w.r.t. the baseline. Table I shows the estimated scores for the baseline and the T-F domain networks. The baseline model obtained the highest median SDR of 3.58 dB. However, the MAD of 2.89 dB and SD of 13.78 dB for the baseline indicates more variability in the estimated vocal data space. The MAD and SD for the MHE regularized T-F domain networks ranged from 1.52 dB to 1.05 dB and 1.96 dB to 1.56 dB, respectively. Also, the MHE regularized T-F domain network (WGAN-MHE) achieved a maximum improvement of 0.21 dB in mean SDR over the baseline. This shows that regularizing the T-F domain network with MHE is currently the best possible alternative to the MHE regularized time-domain networks.

B. Identifying a suitable objective function for MHE regularized T-F domain networks

Fig. 4 shows the impact of MHE on T-F domain networks against their un-regularized T-F models. Adding MHE regularization to the DNN model improves the overall GSDR and GSAR, but fails to remove the interference. The MHE turns out to be ineffective when used in the adversarial framework such as GANs (Fig. 4 (b)). However, when the MHE is used in advanced adversarial networks such as WGAN (Fig. 4 (c)) and RECON-GAN (Fig. 4 (d)), can reduce the interference and artifacts, respectively. The dotted circle in Fig 3 (d) shows the ability of WGAN-MHE in preserving vocal information by reducing interference. The dotted rectangle in Fig. 3 (c) shows that the RECON-GAN-MHE is able to remove the algorithm artifacts. Based on the above inference, we introduce MHE to the RECON-WGAN, a network combination of reconstruction loss and the Wasserstein architecture. As can be seen in Fig. 4 (e) and Fig. 3 (e), (f), of all the adversarial networks, regularizing the RECON-WGAN with

MHE achieved a maximum performance improvement against its un-regularized version. Table II shows the improvement in the MHE regularized T-F domain adversarial networks to their un-regularized version. The RECON-WGAN-MHE achieved the maximum improvement of 0.45 dB, 1.57 dB, 0.20 dB, and 0.47 dB in terms of mean SDR, GSIR, GSAR, and GSDR over its un-regularized version, respectively. The results confirm the need of designing a suitable adversarial objective function (in our case, reconstruction loss with Wasserstein architecture) for using MHE regularization in singing voice separation.

TABLE II
Improvement in the MHE regularized T-F domain adversarial networks with respect to their un-regularized versions.

MHE-Model	mean SDR(dB)	GSIR(dB)	GSAR(dB)	GSDR(dB)
RECON-WGAN-MHE	**0.45**	**1.57**	**0.20**	**0.47**
RECON-GAN-MHE	-0.13	-1.00	0.05	-0.16
WGAN-MHE	-0.12	0.30	-0.05	-0.11
GAN-MHE	-0.63	-1.73	0.15	-0.63

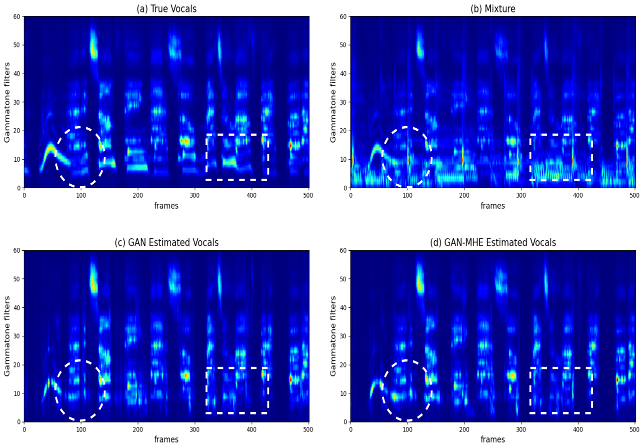

Fig. 5. GAN-MHE fails to properly predict the spectrum: a. true vocals, b. mixture, c. GAN estimated vocals, and d. GAN-MHE estimated vocals: dotted portion shows the area where it fails to reduce artifacts from the music accompaniment.

C. Un-regularized GAN outperforms its MHE regularized version and the baseline

As seen in section IV-B and Table II, only for the RECON-WGAN-MHE configuration, the MHE yields the best result w.r.t. its un-regularized version. However, it can be seen in Table I that the best performing un-regularized GAN achieves the maximum improvement of 0.63 dB in mean SDR over its MHE regularized version. The GAN achieves the highest median SDR of 0.66 dB w.r.t. all other T-F networks. The GSDR, GSIR, and GSAR also suggest that the estimated singing voice using GAN outperforms its MHE version. Fig. 5 shows the spectrum learned by GAN and GAN-MHE T-F domain networks. The dotted portion in Fig. 5 (c) shows the area where GAN is able to reduce the artifacts and interference introduced from the music accompaniment (shown as red coloured vertical patch at 400^{th} frame and 10^{th} filter, in Fig. 5 (b)) as against its MHE version (Fig. 5 (d)). In each of the estimated Gammatone spectrum, the lower frequency region is able to significantly reduce the interference. This observation shows the effectiveness of using Gammatone-based auditory features in generating statistically relevant features, as also observed in [41].

V. Conclusions and Future Work

In the context of singing voice separation from a musical mixture, we have presented a general method of using Minimum Hyperspherical Energy (MHE) regularizer for the Time-Frequency (T-F) domain networks. We have demonstrated that the Gammatone-based auditory features are effective in reducing the interference introduced by the music accompaniment. We have presented a detailed analysis of applying MHE to the time and T-F domain networks. The T-F domain Generative Adversarial Network (GAN) outperforms the time-domain baseline with an improvement of 0.75 dB in mean Signal-to-Distortion Ratio (SDR). We have reported the failure of MHE regularized T-F domain networks over their un-regularized models. To solve this problem, we have shown a need to identify a suitable adversarial objective function for the MHE regularized T-F domain networks. Adding a reconstruction loss to the Wasserstein GAN-MHE (RECON-WGAN-MHE) achieved a 0.45 dB improvement in mean SDR over its un-regularized version. For future work, we plan to use MHE regularization to the other advanced adversarial networks such as cycle-GAN, to evaluate the effect of cycle loss on the singing voice separation task. We also plan to extend this work to speech separation and enhancement.

VI. Acknowledgements

We sincerely thank authorities of Tata Research Development and Design Centre, and TCS Media Entertainment & Advertising Research, Pune for their kind support.

References

[1] J. Perez-Lapillo, O. Galkin, and T. Weyde, "Improving singing voice separation with the wave-u-net using minimum hyperspherical energy," *arXiv preprint arXiv:1910.10071*, 2019.

[2] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks." in *ISMIR*, 2014, pp. 477–482.

[3] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[4] J. Sebastian and H. A. Murthy, "Group delay based music source separation using deep recurrent neural networks," in *2016 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2016, pp. 1–5.

[5] P. Smaragdis, C. Fevotte, G. J. 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] A. J. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 429–436.

[7] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564–1578, 2007.

[8] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries." in *ISMIR*, 2013, pp. 427–432.

[9] I.-Y. Jeong and K. Lee, "Singing voice separation using rpca with weighted l1-norm," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 553–562.

[10] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 10, pp. 2085–2098, 2015.

[11] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 171–184, 2012.

[12] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep nmf for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 66–70.

[13] Y. Huang, "Non-local mmdensenet with cross-band features for audio source separation," in *International Conference on Intelligent Science and Big Data Engineering*. Springer, 2019, pp. 53–64.

[14] W. Yuan, S. Wang, X. Li, M. Unoki, and W. Wang, "A skip attention mechanism for monaural singing voice separation," *IEEE Signal Processing Letters*, vol. 26, no. 10, pp. 1481–1485, 2019.

[15] S. Venkataramani, E. Tzinis, and P. Smaragdis, "A style transfer approach to source separation," *arXiv preprint arXiv:1905.00151*, 2019.

[16] Y. C. Subakan and P. Smaragdis, "Generative adversarial source separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 26–30.

[17] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2391–2395.

[18] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 61–65.

[19] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[20] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," 2017.

[21] G. Mahé, E. Z. Nadalin, R. Suyama, and J. M. Romano, "Perceptually controlled doping for audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 27, 2014.

[22] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 21–25.

[23] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[24] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5039–5043.

[25] B. Sisman, K. Vijayan, M. Dong, and H. Li, "Singan: Singing voice conversion with generative adversarial networks," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 112–118.

[26] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *ICML*, vol. 2, no. 3, 2016, p. 7.

[27] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song, "Learning towards minimum hyperspherical energy," in *Advances in Neural Information Processing Systems*, 2018, pp. 6222–6233.

[28] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *The Journal of the Acoustical Society of America (JASA)*, vol. 87, no. 6, pp. 2592–2605, 1990.

[29] L. H. Carney and T. Yin, "Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model," *Journal of Neurophysiology*, vol. 60, no. 5, pp. 1653–1677, 1988.

[30] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. on Multimedia*, vol. 14, no. 6, pp. 1684–1689, 2012.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems (NIPS)*, Montréal, Canada, 2014, pp. 2672–2680.

[32] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4390–4394.

[33] N. Shah, H. A. Patil, and M. H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1246–1251.

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in *ICML*, Sydney, Australia, 2017, pp. 214–223.

[35] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *NIPS*, California, USA, 2017, pp. 1–10.

[36] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1857–1865.

[37] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-a corpus for music separation," 2017.

[38] A. Liutkus, D. Fitzgerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 76–80.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[41] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel teo-based gammatone features for environmental sound classification," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1809–1813.