Proceedings Paper:
Mai, Y. and Goetze, S. orcid.org/0000-0003-1044-7343 (2025) MetricGAN+KAN: Kolmogorov-Arnold networks in metric-driven speech enhancement systems. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Proceedings. ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 06-11 Apr 2025, Hyderabad, India. Institute of Electrical and Electronics Engineers (IEEE) ISBN 9798350368758

https://doi.org/10.1109/ICASSP49660.2025.10890542

MetricGAN+KAN: Kolmogorov-Arnold Networks in Metric-Driven Speech Enhancement Systems

Yemin Mai[1] and Stefan Goetze[1,2]
[1]Speech and Hearing (SPandH) group, School of Computer Science, The University of Sheffield, Sheffield, United Kingdom
[2]South Westphalia University of Applied Sciences, Iserlohn, Germany
{ymai5, s.goetze}@sheffield.ac.uk, goetze.stefan@fh-swf.de

***Abstract*—Neural-network-based speech enhancement (SE) approaches have shown to be particularly powerful in combination with perceptually motivated metrics to produce high-quality enhanced speech signals. Among these deep learning (DL)-based SE models, MetricGAN and its extension can generate output signals directly optimising quality metrics. The recently proposed Kolmogorov-Arnold networks (KANs) with learnable activation functions have shown great success in replacing multi-layer perceptrons (MLPs). This work proposes the use of KANs in a MetricGAN framework and analyses their performance in replacing different types of network layers. The best-performing proposed MetricGAN+KAN model uses approximately 80% fewer parameters and achieves 13.2% higher SE performance (measured by PESQ) on the Voicebank-DEMAND dataset, compared to the MetricGAN+ baseline.**

***Index Terms*—Speech enhancement, quality metrics, Kolmogorov-Arnold network (KAN), Generative adversarial network (GAN), MetricGAN**

I. Introduction

Single-channel speech enhancement (SE) has been a popular research field for some decades [1], focusing on improving the quality [2]–[5] or intelligibility [6]–[8] of speech signals in noisy, reverberant environments [9]. Machine learning (ML)-based approaches have led to significant performance gains in recent years, and become the first choice of modelling for SE [10]–[13]. Generative adversarial networks (GANs) [14] which consist of two sub-models, a generator and a discriminator, have proven to be effective in SE. The MetricGAN [15] approach and its extensions [16]–[24] have achieved the-state-of-the-art results on the Voicebank-DEMAND [25] dataset. However, only limited research exists for optimising the model structure of the MetricGAN framework even though this was already suggested by the authors of [16]. Furthermore, it is time-consuming to train the model with replay buffer, which is necessary for addressing catastrophic forgetting [26].

Recently, KANs [27] have been proposed, integrating learnable activation functions parameterised by B-spline curves into neurons. Authors of KANs have also mentioned that KANs can overcome catastrophic forgetting [27]. Hence, this work analyses the use KANs in the MetricGAN+ framework. The proposed KAN-based SE model is therefore denoted as MetricGAN+KAN, and this work aims at validating some advantages of KANs in a MetricGAN setting and analysing, which network layers can be changed to KAN-based structures. To this end, different model structures, i.e. positions to replace model layers with KAN-based layers, are compared in terms of performance as well as model parameters on the Voicebank-DEMAND [25] task.

II. Review of the MetricGAN+ Framework

MetricGAN+ [16] is a spectro-temporal masking-based SE approach. For this, the noisy input signal is first converted to a magnitude spectrogram $X_{f,\tau}$ and a phase spectrogram $\gamma_{f,\tau}$ by the short-time Fourier transform (STFT), where f is the frequency index and τ is the frame index. For the enhancement process, a spectral mask $M_{f,\tau}$ is computed and multiplied with the magnitude spectrogram of the noisy signal to obtain an estimate of the clean magnitude spectrogram

$$\hat{S}_{f,\tau} = M_{f,\tau} \cdot X_{f,\tau}. \tag{1}$$

Then, an estimate of the clean signal is re-synthesised using $\hat{S}_{f,\tau}$ and $\gamma_{f,\tau}$, i.e. by applying the inverse STFT to $\hat{S}_{f,\tau} e^{j\gamma_{f,\tau}}$.

MetricGAN+ [16] consists of two neural networks (NNs), a generator $\mathcal{G}$ aiming to estimate the mask $M_{f,\tau}$ and a discriminator $\mathcal{D}$ assessing the quality of the masking-based SE by metric prediction.

The generator $\mathcal{G}$ takes a noisy spectrogram $X_{f,\tau}$ as the input and outputs the mask $M_{f,\tau}$. Figure 1 visualises the generator $\mathcal{G}_0$ of the MetricGAN+ baseline [16], which can be split into a part containing recursive layers, i.e. a bidirectional long short-term memory (LSTM) [28] and a part containing non-recursive layers, i.e. linear layers for MetricGAN+ with a leaky rectified linear unit (ReLU) activation function. A learnable sigmoid outputs the mask $M_{f,\tau}$.

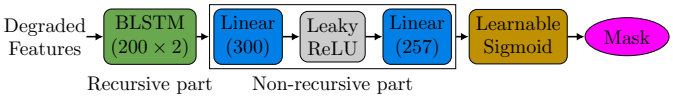

Figure 1: Generator model $\mathcal{G}_0$ (MetricGAN+ baseline).

Figure 2 visualises the discriminator model structure of MetricGAN+, denoted as $\mathcal{D}_0$. Subscripts $_0$ in Figures 1 and 2 indicate the MetricGAN+ baseline [16] model in contrast to model variants introduced later in this work. After a batch normalisation (BN) layer, it can be split into a convolutional part and a non-convolutional part.

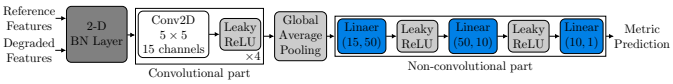

Figure 2: Discriminator model $\mathcal{D}_0$ (MetricGAN+ baseline).

The discriminator $\mathcal{D}$ predicts a metric score $Q'(\cdot)$ normalised between 0 and 1 (often a normalised version of the perceptual evaluation of speech quality (PESQ) metric) given the noisy (or enhanced) magnitude spectrogram and the corresponding clean spectrogram $S_{f,\tau}$. The discriminator is thus a differentiable surrogate function which imitates non-differentiable intrusive metrics of audio quality such as PESQ [29], subjective mean opinion score (MOS) [30], or DNSMOS [31] since such perceptually motivated metrics often correlate better with human perception [2], [32] than traditional loss functions such as the simple mean squared error (MSE) and hence can address discriminator evaluation mismatch (DEM) [15].

In terms of training, the loss function for the discriminator is given by

$$L_{\mathcal{D}} = \mathbb{E}_{\mathbf{X},\mathbf{S}} \left[(\mathcal{D}(\mathbf{S},\mathbf{S}) - Q'(\mathbf{S},\mathbf{S}))^2 + (\mathcal{D}(\mathcal{G}(\mathbf{X}),\mathbf{S}) - Q'(\mathcal{G}(\mathbf{X}),\mathbf{S}))^2 + (\mathcal{D}(\mathbf{X},\mathbf{S}) - Q'(\mathbf{X},\mathbf{S}))^2 \right], \quad (2)$$

where $\mathbb{E}$ is the expectation operator, $\mathbf{X}$ is the magnitude spectrogram matrix of the noisy signal containing $X_{f,\tau} \forall f, \tau$ and $\mathbf{S}$ is the respective clean spectrogram matrix. The loss function for the generator is given by

$$L_{\mathcal{G}} = \mathbb{E}_{\mathbf{X}} \left[(\mathcal{D}(\mathcal{G}(\mathbf{X}),\mathbf{S}) - w)^2 \right], \quad (3)$$

with w being the desired metric score that the discriminator assigned to the enhanced speech, which is set to 1 in [16] maximising the enhancement or varied in [22]. In each epoch, MetricGAN+ is trained using the following procedure:

1) Train the generator $\mathcal{G}$ using back-propagation (BP) [33].
2) Store current enhanced signals $\hat{\mathbf{S}} = \mathcal{G}(\mathbf{X})$ and the corresponding scores Q' into the so-called replay buffer.
3) Train the discriminator $\mathcal{D}$ using clean signals $\mathbf{S}$, current enhanced speech signals $\hat{\mathbf{S}} = \mathcal{G}(\mathbf{X})$ and noisy signals $\mathbf{X}$.
4) Repeat 3), but use a part of the previously enhanced signals $\hat{\mathbf{S}} = \mathcal{G}(\mathbf{X})$ from the replay buffer, controlled by the hyperparameter `history_portion` [16], [22].

As mentioned above, the replay buffer is used to address catastrophic forgetting in the discriminator. It can greatly improve the performance of the discriminator, and subsequently improve the quality of signals enhanced by the generator. However, training with the replay buffer also increases training time.

Authors of MetricGAN+ [16] already mention that the structure of the discriminator can be improved which will be analysed in this work by using KANs in the recursive and non-recursive parts of the generator (cf. Figure 1) as well as in the convolutional and non-convolutional part of the discriminator (cf. Figure 2).

III. Review of Kolmogorov-Arnold networks (KANs)

KANs [27] are a recently proposed type of NN architecture having gained considerable attention on GitHub[1]. The novelty of KANs is that the activation function is placed within the neuron, and is learnable. It is inspired by the Kolmogorov-Arnold representation theorem [34]

$$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right), \quad (4)$$

where $f : [0,1]^n \to \mathbb{R}$ is smooth, $\phi_{q,p} : [0,1] \to \mathbb{R}$, and $\Phi_q : \mathbb{R} \to \mathbb{R}$. Based on (4), KANs replace the weight matrix in a traditional NN layer with a matrix of functions, denoted by $\mathbf{\Phi} = \{\phi_{q,p}\}$ where $p = 1, 2, \ldots, n_{\text{in}}$ and $q = 1, 2, \ldots, n_{\text{out}}$. $\phi(\cdot)$ is a learnable activation function formulated as the scaled sum of a base activation function $b(x)$ and a learnable curve $g(x)$,

$$\phi(x) = w_1 b(x) + w_2 g(x), \quad (5)$$

where w_1 and w_2 are scaling factors which can be learnable. It is noteworthy that in the original paper [27], only one scaling factor is used for $\phi(\cdot)$. Liu *et al.* used sigmoid linear units (SiLUs) [35]

$$b(x) = \frac{x}{1 + e^{-x}} \quad (6)$$

as the base activation functions for KANs. Several implementations have been reviewed in [36]. For the learnable curves, B-splines, which require basis functions and controlling points were proposed initially with B-spline basis functions of order k defined as [37], [38]

$$B_{i,0}(x) = \begin{cases} 1 & \text{if } x_i \le x \le x_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

$$B_{i,k}(x) = \frac{x - x_i}{x_{i+k} - x_i} B_{i,k-1}(x) + \frac{x_{i+k+1} - x}{x_{i+k+1} - x_{i+k}} B_{i+1,k-1}(x), \quad (8)$$

with x_i for $i = -k, -k+1, \ldots, G+k$ being the predefined boundary, and G the grid size. The spline curve is given by

$$g(x) = \sum_{i=0}^{G+k-1} c_i B_{i,k}(x), \quad (9)$$

where c_i for $i = 0, 1, \ldots, G+k-1$ is the trainable controlling point, and $g(x)$ is defined on $[x_0, x_G]$.

KANs have also been integrated into convolutional neural networks (CNNs) and recurrent neural networks (RNNs), called convolutional KANs (CKANs) [39] and recurring KANs (RKANs) [40], respectively. In CKANs, the kernel becomes a group of learnable activation functions. The result of a 2-D convolution (with stride 1) is given by

$$a_{x,y}^{(l)} = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi_{i,j}^{(l)} \left(a_{x+i,y+j}^{(l-1)} \right), \quad (10)$$

where $a_{x,y}^{(l)}$ is the feature map at layer l. In RKANs, the prediction at time t is given by

$$\hat{\mathbf{y}}_t = \mathbf{\Phi} \mathbf{h}_t, \quad (11)$$

where $\mathbf{h}_t$ is the hidden state at time t.

As mentioned in [27], advantages of KANs are that KANs can overcome catastrophic forgetting, because the update of c_i only changes part of the spline curve, and that KANs have a better scaling law than traditional NNs, i.e., using fewer parameters to achieve similar (or higher) performance compared to traditional NNs. Bodner *et al.* [39] have shown that these two advantages also hold for CKANs.

IV. Experiments

A. Dataset

The Voicebank-DEMAND dataset [25] is used for the following experiments, created from the Voicebank dataset [41] mixed with various types of noises (cafeteria, car interior, a kitchen, meeting, metro station, restaurant, train station and heavy traffic) from the DEMAND dataset [42], recorded indoor and outdoor, and two others (babble noise and speech-shaped noise).

The training set of Voicebank-DEMAND consists of 11572 noisy speech signals at 4 signal-to-noise ratios (SNRs) of 0, 5, 10, and 15 dB paired with the respective clean speech reference signals from 28 different speakers (14 male, 14 female), with English or Scottish accents. The testset contains 824 utterances, mixed at SNRs of 2.5, 7.5, 12.5 and 17.5 dB, with five different noises which do not appear in the training set (bus, cafe, office, public square and living room) and contains speech from two (one male, one female) speakers who do not appear in the training set.

B. Implementation

Models are trained using the SpeechBrain framework [43]. For the implementation of KANs, `efficient-kan` [44] is used, and `torch-conv-kan` [36] for CKANs (with parametric ReLU [45] activation at the output). For RKANs, a gated recurrent unit (GRU)

[1] https://github.com/KindXiaoming/pykan.

[46] version of RKANs, namely `GRU-KAN`, is implemented. GRU-KAN uses the same formulae as GRU, but uses (11) for computing the prediction.

C. Experiment Setup

In the following, layers in the MetricGAN+ structure are successively replaced by KAN layers to analyse where these are advantageous later in Section V. Figures 3 and 4 illustrate this for layers in the generator $\mathcal{G}$ from Figure 1 being replaced step-by-step, i.e. first in Figure 3 one linear layer from Figure 1 is replaced by a KAN layer to result in the MetricGAN+KAN generator structure $\mathcal{G}_1$ and then, in Figure 4, the KAN layer replaces both linear layers from the MetricGAN+ baseline in Figure 1, resulting in MetricGAN+KAN generator structure $\mathcal{G}_2$.

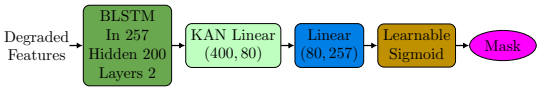

Figure 3: MetricGAN+KAN generator structure $\mathcal{G}_1$, replacing one linear layer from Figure 1 by a KAN layer.

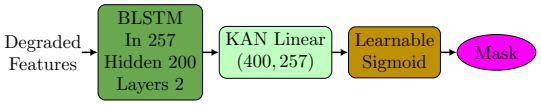

Figure 4: MetricGAN+KAN generator structure $\mathcal{G}_2$, replacing both linear layers from Figure 1 by a KAN layer.

To save space for visualisations of all generator structures under test, Table I summarises the recursive and non-recursive parts as defined in Figure 1 leading to 6 different generator structures $\mathcal{G}_1$ to $\mathcal{G}_6$ under test (in addition to the MetricGAN+ baseline).

Table I: Generator structures. Recursive part specifies the type of RNN and non-recursive part specifies the type of feed-forward layers as visualised in Figure 1. Number in the parenthesis specifies the hidden size (or the number of neurons), and $\times$ specifies the number of layers (default value is 1). MGK abbreviates MetricGAN+KAN.

Structure	Recursive part	Non-recursive part
MetricGAN+ [16]	BLSTM (200×2)	Linear $(300, 257)$
MGK-$\mathcal{G}_1$	BLSTM (200×2)	KAN (80), Linear (257)
MGK-$\mathcal{G}_2$	BLSTM (200×2)	KAN (257)
MGK-$\mathcal{G}_3$	BLSTM (40)	KAN (257)
MGK-$\mathcal{G}_4$	BGRU (40)	KAN (257)
MGK-$\mathcal{G}_5$	BGRU (100)	Linear $(300, 257)$
MGK-$\mathcal{G}_6$	BGRU-KAN (40×2)	KAN (257)

Accordingly, Table II defines the model structure of 5 different KAN-based discriminator structures $\mathcal{D}_1$ to $\mathcal{D}_5$ and Table III shows the number of parameters of each of the generator and discriminator models.

All models are trained for 400 epochs. For KAN hyper-parameters, the range of splines is $[-1, 1]$, the grid size is 5, and the spline order is set to 3. Other hyper-parameters used in MetricGAN+KAN are the same as MetricGAN+.

The code to reproduce experiments can be found on GitHub[2].

[2]https://github.com/Unuse1ess/MetricGAN-KAN

Table II: Discriminator structures. Convolutional part specifies the type of CNN and non-convolutional part specifies the type of feed-forward layers (cf. Figure 2). The number in the parenthesis specifies the number of output channels. The shape of all the convolutional kernels is 5×5.

Structure	Convolutional part	Non-convolutional part
MetricGAN+ [16]	Conv2d (15×4)	Linear $(50, 10, 1)$
MGK-$\mathcal{D}_1$	Conv2d (15×4)	Linear (50), KAN (1)
MGK-$\mathcal{D}_2$	Conv2d (15×4)	KAN (1)
MGK-$\mathcal{D}_3$	CKAN2d (15×2)	KAN (1)
MGK-$\mathcal{D}_4$	CKAN2d (15×3)	KAN (1)
MGK-$\mathcal{D}_5$	CKAN2d (20)	KAN (1)

Table III: The number of parameters of all (a) generator and (b) discriminator variants.

(a) Generator parameters.

$\mathcal{G}_0$	1 895 514
$\mathcal{G}_1$	2 038 674
$\mathcal{G}_2$	2 725 857
$\mathcal{G}_3$	301 537
$\mathcal{G}_4$	277 617
$\mathcal{G}_5$	353 314
$\mathcal{D}_6$	361 267

(b) Discriminator parameters.

$\mathcal{D}_0$	19 010
$\mathcal{D}_1$	18 989
$\mathcal{D}_2$	17 839
$\mathcal{D}_3$	57 531
$\mathcal{D}_4$	108 157
$\mathcal{D}_5$	9 205

V. Results

Table IV shows the performance of different generator structures as specified in Table I in terms of PESQ and the composite metrics for signal, background and overall quality, respectively. The discriminator structure is kept constant being the baseline [16] discriminator $\mathcal{D}_0$.

Table IV: Results for different generators. NHP indicates that no history portion was used during training, i.e. `history_portion` is set to 0.

	PESQ	**CSIG**	**CBAK**	**COVL**
Noisy	1.97	3.35	2.44	2.63
MetricGAN+ [16]	2.89	3.78	2.92	3.31
MGK-$\mathcal{G}_1$-$\mathcal{D}_0$	2.85	3.73	2.90	3.26
MGK-$\mathcal{G}_1$-$\mathcal{D}_0$ (NHP)	2.68	3.93	2.73	3.30
MGK-$\mathcal{G}_2$-$\mathcal{D}_0$	2.82	3.80	**2.93**	3.28
MGK-$\mathcal{G}_3$-$\mathcal{D}_0$	2.85	3.69	2.83	3.23
MGK-$\mathcal{G}_4$-$\mathcal{D}_0$	**2.94**	3.82	2.88	3.35
MGK-$\mathcal{G}_5$-$\mathcal{D}_0$	2.88	**3.94**	2.80	**3.38**
MGK-$\mathcal{D}_6$-$\mathcal{D}_0$	**2.93**	3.84	2.93	3.36

Table V shows the results for the different discriminator structures $\mathcal{D}$ for fixed generator $\mathcal{G}_0$. Model architecture details are according to Table II.

Table V: Results of only changing discriminators.

	PESQ	**CSIG**	**CBAK**	**COVL**
Noisy	1.97	3.35	2.44	2.63
MetricGAN+ [16]	2.89	3.78	2.92	3.31
MGK-$\mathcal{G}_0$-$\mathcal{D}_1$	2.94	4.00	2.91	3.45
MGK-$\mathcal{G}_0$-$\mathcal{D}_2$	2.93	3.97	3.01	3.44
MGK-$\mathcal{G}_0$-$\mathcal{D}_3$	3.02	4.03	3.02	3.50
MGK-$\mathcal{G}_0$-$\mathcal{D}_4$	**3.30**	4.02	**3.04**	**3.63**
MGK-$\mathcal{G}_0$-$\mathcal{D}_4$ (NHP)	2.72	3.96	2.75	3.32
MGK-$\mathcal{G}_0$-$\mathcal{D}_5$	2.96	**4.15**	3.19	3.55

Discriminators $\mathcal{D}_1$ and $\mathcal{D}_2$ which integrate KANs improve the performance of SE marginally with slightly fewer parameters, com-

pared to MetricGAN+. In terms of discriminators integrated with CKANs, discriminators $\mathcal{D}_3$ and $\mathcal{D}_4$ show significant improvement, however with higher parameter count, and discriminator $\mathcal{D}_5$ shows slight improvement with significantly fewer parameters.

Analysing the performance of different generator structures, generators $\mathcal{G}_1$ and $\mathcal{G}_2$ show slightly degraded performance even with significantly higher parameter count. Generator $\mathcal{G}_3$ uses less hidden states and a smaller number of layers leading to significantly fewer parameters, and still achieves similar performance. Generator $\mathcal{G}_5$ shows that the GRU architecture leads to better performance than LSTMs, and generator $\mathcal{G}_4$ shows that KANs further improve the performance of generator $\mathcal{G}_5$ with significantly fewer parameters. Generator $\mathcal{D}_6$ shows that GRU-KAN also works well compared to MetricGAN+. In summary, the integration of KANs leads to only smaler improvements, while CKANs are able to significantly improve SE performance. Both KANs and CKANs can have similar performance compared to traditional NNs using fewer parameters, and CKANs outperforms CNNs.

Among the tested variants of MetricGAN+KAN, MGK-$\mathcal{G}_4$-$\mathcal{D}_4$ achieves the best results in terms of PESQ and overall quality COVL taking model complexity into account, i.e. while using 79.9% fewer parameters (85.4% fewer for the generator, 468.9% more for the discriminator), achieving 13.2% higher PESQ scores compared to the MetricGAN+ baseline [16] (cf. Table VI).

Table VI: Results of further generator/discriminator combinations compared to the baseline [16].

	PESQ	**CSIG**	**CBAK**	**COVL**
Noisy	1.97	3.35	2.44	2.63
MetricGAN+ [16]	2.89	3.78	2.92	3.31
MGK-$\mathcal{G}_4$-$\mathcal{D}_3$	3.00	3.98	2.95	3.46
MGK-$\mathcal{G}_4$-$\mathcal{D}_3$ (NHP)	2.66	3.85	2.88	3.24
MGK-$\mathcal{G}_4$-$\mathcal{D}_4$	**3.27**	3.97	2.97	**3.59**
MGK-$\mathcal{G}_5$-$\mathcal{D}_3$	3.07	**4.10**	3.00	3.57
MGK-$\mathcal{G}_5$-$\mathcal{D}_4$	3.08	4.08	2.99	3.56
MGK-$\mathcal{D}_6$-$\mathcal{D}_3$	2.99	4.03	**3.04**	3.49
MGK-$\mathcal{D}_6$-$\mathcal{D}_4$	3.12	3.90	2.95	3.48

Figure 5 compares the (a) clean, (b) noisy, and (c)-(d) enhanced signals by (c) MetricGAN+ and (d) MGK-$\mathcal{G}_4$-$\mathcal{D}_4$ in terms of their magnitude spectrograms.

Results in experiments without replay buffer indicated by NHP in Tables IV, V and VI also show that MetricGAN+KAN still suffers from catastrophic forgetting, and the use of replay buffer is still necessary. Models trained without replay buffer still show significantly lower performance. A possible reason may be that the locality of splines only exists in (9). In (5), the spline curve is scaled, which may reduce the effectiveness of the locality. In other words, in some neurons, the base activation function has a higher influence than the spline curve (cf. Figure 6). However, further experiments might be necessary regarding this matter.

VI. Possible Future Work

In this work, MetricGAN+KAN is tested on the commonly used Voicebank-DEMAND dataset to be comparable to most recent SE literature. Since it is known that Voicebank-DEMAND is a relatively simple task [22], [47], the model needs further experiments on different datasets, such as e.g. [47]. Furthermore, reasons for the still existing catastrophic forgetting in MetricGAN+KAN (or KANs) need further investigation. Additionally, [36] presented several implementations of the learnable curve, which may be a possible direction for further research.

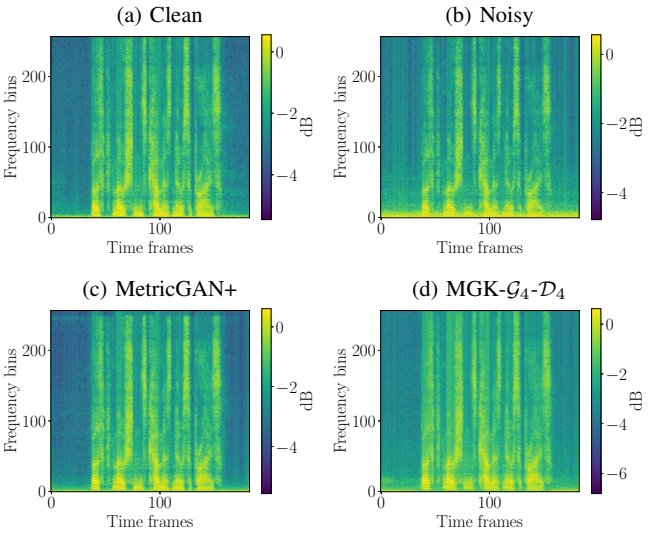

Figure 5: Comparison of magnitude spectrograms.

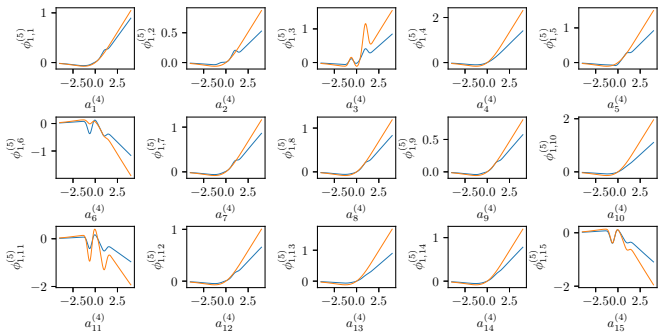

Figure 6: Changes of Eq. (5) in MGK-$\mathcal{G}_0$-$\mathcal{D}_4$, taken from epoch 100 (in blue) and epoch 400 (in red). $\phi_{i,j}^{(\ell)}$ represents the learnable activation function of input dimension i, output dimension j at layer ℓ, with the corresponding input $a_j^{(\ell-1)}$.

VII. Conclusion

This work analysed the use of KANs for a MetricGAN+ SE system. The integration of KANs, CKANs and RKANs can improve the SE performance of MetricGAN+. The proposed model, MGK-$\mathcal{G}_4$-$\mathcal{D}_4$, achieves 13.2% higher PESQ scores with 79.85% fewer parameters compared to the MetricGAN+ baseline [16], indicating a better scaling law. Experimental results also show that the use of KANs in MetricGAN+KAN cannot mitigate catastrophic forgetting, and the replay buffer is still necessary.

References

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons, Ltd, 2006.

[2] S. Goetze, A. Warzybok, I. Kodrasi, J. Jungmann, B. Cauchi, J. Rennies, E. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A Study on Speech Quality and Speech Intelligibility Measures for Quality Assessment of Single-Channel Dereverberation Algorithms," in *Proc. IWAENC'14*, Sep. 2014.

[3] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, 2020.

[4] B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, S. Doclo, and S. Goetze, "Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1151–1163, July 2019.

[5] G. Close, W. Ravenscroft, T. Hain, and S. Goetze, "Perceive and predict: self-supervised speech representation based loss functions for speech enhancement," in *Proc. ICASSP 2023*, 2023.

[6] C. Völker, A. Warzybok, and S. Ernst, "Comparing Binaural Preprocessing Strategies III: Speech Intelligibility of Normal-Hearing and Hearing-Impaired Listeners," *Trends in Hearing*, vol. 19, 2015.

[7] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-Intrusive Speech Intelligibility Prediction for Hearing-Impaired Users using Intermediate ASR Features and Human Memory Models," in *Proc. ICASSP'24*, (Seoul, South Korea), Apr. 2024.

[8] R. Sutherland, G. Close, T. Hain, S. Goetze, and J. Barker, "Using speech foundational models in loss functions for hearing aid speech enhancement," in *European Signal Processing Conf. (EUSIPCO)*, 2024.

[9] F. Xiong, B. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-end technologies for robust ASR in reverberant environments - spectral enhancement-based dereverberation and auditory modulation filterbank features," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, 2015.

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: dataset, task and baselines," in *Proc. ASRU*, 2015.

[11] W. Ravenscroft, S. Goetze, and T. Hain, "Att-TasNet: Attending to Encodings in Time-Domain Audio Speech Separation of Noisy, Reverberant Speech Mixtures," *Frontiers in Signal Processing*, vol. 2, 2022.

[12] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement," in *Proc. ICASSP*, 2021.

[13] N. Moritz, K. Adiloğlu, J. Anemüller, S. Goetze, and B. Kollmeier, "Multi-channel speech enhancement and amplitude modulation analysis for noise robust automatic speech recognition," *Computer Speech & Language*, vol. 46, 2017.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[15] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement," in *Proc. 36th Int. Conf. on Machine Learning*, Jun 2019.

[16] S. Fu, C. Yu, T. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," *CoRR*, vol. abs/2104.03538, 2021.

[17] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17022–17033, 2020.

[18] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *ICASSP*, pp. 5024–5028, 2018.

[19] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[20] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[21] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, vol. 2013, 2013.

[22] G. Close, T. Hain, and S. Goetze, "MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data," in *EUSIPCO 2022*, 2022.

[23] G. Close, T. Hain, and S. Goetze, "PAMGAN+/-: Improving Phase-Aware Speech Enhancement Performance via Expanded Discriminator Training," in *AES 154th Conv.*, May 2023.

[24] G. Close, W. Ravenscroft, T. Hain, and S. Goetze, "CMGAN+/+: The University of Sheffield CHiME-7 UDASE Challenge Speech Enhancement System," in *Proc. 7th Int. Workshop on Speech Processing in Everyday Environments (CHiME 2023)*, Aug. 2023.

[25] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *SSW*, pp. 146–152, 2016.

[26] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," *arXiv preprint arXiv:1312.6211*, 2013.

[27] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "Kan: Kolmogorov-Arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

[28] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, 12 1997.

[29] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, 2001.

[30] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[31] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497, 2021.

[32] A. Avila, B. Cauchi, S. Goetze, S. Doclo, and T. Falk, "Performance Comparison of Intrusive and Non-Intrusive Instrumental Quality Measures for Microphone-Array Processed Speech," in *Proc. International Workshop on Acoustic Signal Enhancement IWAENC*, 2016.

[33] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, 1986.

[34] A. N. Kolmogorov, "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition," in *Doklady Akademii Nauk*, vol. 114, pp. 953–956, Russian Academy of Sciences, 1957.

[35] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *CoRR*, vol. abs/1702.03118, 2017.

[36] I. Drokin, "Kolmogorov-Arnold convolutions: Design principles and empirical studies," *arXiv preprint arXiv:2407.01092*, 2024.

[37] W. J. Gordon and R. F. Riesenfeld, "B-spline cruves and surfaces," in *Computer Aided Geometric Design*, pp. 95–126, Academic Press, 1974.

[38] C. Boor, "Subroutine package for calculating with b-splines," 1971.

[39] A. D. Bodner, A. S. Tepsich, J. N. Spolski, and S. Pourteau, "Convolutional Kolmogorov-Arnold networks," 2024.

[40] R. Genet and H. Inzirillo, "Tkan: Temporal Kolmogorov-Arnold networks," *arXiv preprint arXiv:2405.07344*, 2024.

[41] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Int. conf. oriental COCOSDA, jointly with 2013 Conf. on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, 2013.

[42] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," *Proc. of Meetings on Acoustics*, vol. 19, no. 1, 2013.

[43] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021. arXiv:2106.04624.

[44] Blealtan and A. Dash, "efficient-kan." https://github.com/Blealtan/efficient-kan, 2024.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *CoRR*, vol. abs/1502.01852, 2015.

[46] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[47] G. Close, T. Hain, and S. Goetze, "The effect of spoken language on speech enhancement using self-supervised speech representation loss functions," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.