

DISCRIMINATING REAL AND SYNTHETIC SUPER-RESOLVED AUDIO SAMPLES USING EMBEDDING-BASED CLASSIFIERS

Mikhail Silaev¹, Konstantinos Drossos², and Tuomas Virtanen¹

¹Tampere University, Tampere, Finland

²Nokia Technologies, Espoo, Finland

ABSTRACT

Generative adversarial networks (GANs) and diffusion models have recently achieved state-of-the-art performance in audio super-resolution (ADSR), producing perceptually convincing wideband audio from narrowband inputs. However, existing evaluations primarily rely on signal-level or perceptual metrics, leaving open the question of how closely the distributions of synthetic super-resolved and real wideband audio match. Here we address this problem by analyzing the separability of real and super-resolved audio in various embedding spaces. We consider both middle-band ($4 \rightarrow 16$ kHz) and full-band ($16 \rightarrow 48$ kHz) upsampling tasks for speech and music, training linear classifiers to distinguish real from synthetic samples based on multiple types of audio embeddings. Comparisons with objective metrics and subjective listening tests reveal that embedding-based classifiers achieve near-perfect separation, even when the generated audio attains high perceptual quality and state-of-the-art metric scores. This behavior is consistent across datasets and models, including recent diffusion-based approaches, highlighting a persistent gap between perceptual quality and true distributional fidelity in ADSR models. Code and demo are available at <https://github.com/msilaev/ADRS>.

Index Terms— GAN, discriminator, data distribution, separability, feature representations, bandwidth expansion, audio super-resolution

1. INTRODUCTION

The primary objective of generative adversarial networks (GANs) [1] is to generate synthetic ('fake') samples that closely resemble real data, effectively sampling from a distribution approximating the real one. Evaluating how well GANs achieve this goal remains a fundamental challenge, as traditional metrics often fail to capture perceptual quality, diversity and overfitting to the training data. Reliable evaluation is crucial for understanding model behavior and guiding improvements in generative quality. To evaluate GAN performance, various approaches have been proposed. For example, the representational quality of a GAN's discriminator is evaluated by reusing its frozen intermediate layers

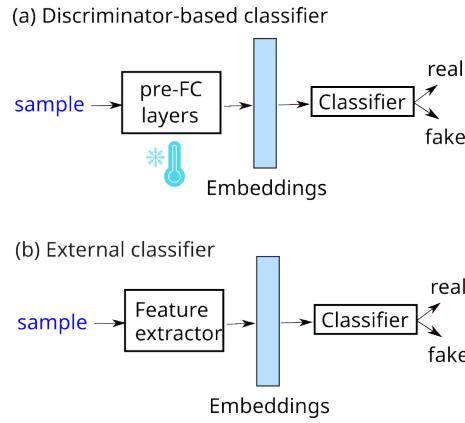


Fig. 1: Feature extraction and 'real'/'fake' classification task. (a) Discriminator-based classifier uses internal discriminator features produced by its pre-fully connected (pre-FC) layers with frozen weights. (b) External classifier operating on independent features extracted by a separate network, enabling an analysis of potential representation bias between the two classifiers.

as feature extractors for supervised downstream tasks, such as image classification [2]. Classifiers trained on learned representation embeddings have been used to measure privacy preservation in the speech and audio domains, including automatic speech recognition [3] and sound event detection [4].

A complementary direction compares the statistical similarity of real and generated samples in embedding spaces such as that induced by a pretrained Inception network [5, 6, 7]. Such embeddings enable quantitative metrics like the Fréchet Inception Distance [5], Fréchet Audio Distance [8], Deep Speech Distance [9], and precision-recall measures [6, 7, 10], which assess the overlap and coverage between real and synthetic data manifolds.

However, these evaluation strategies have not been systematically applied to GAN-based super-resolution tasks, either in image [11] or audio [12] domains. In this work, we extend the assessment of GANs and recently proposed diffusion models [13, 14] by addressing a straightforward yet practically important question: Can linear classifiers trained on ei-

ther frozen discriminator representations (Fig. 1a) or external embeddings (Fig. 1b) reliably distinguish real wideband from super-resolved audio samples?

We focus on the practically relevant audio super-resolution (ADSR) problem, which aims to enhance low-bandwidth audio signals by generating new high-frequency content, thereby expanding their spectral range. This topic has attracted considerable research interest in recent years [13, 14, 15, 16, 17, 18, 19, 12, 20, 21, 22]. The evaluation of GANs and diffusion models for ADSR has focused on signal-level metrics such as signal-to-noise ratio (SNR), log-spectral distance (LSD), and perceptual measures including mean opinion score (MOS) and subjective listener tests [12, 18]. In contrast, GAN-specific evaluations have not yet been explored. This gap is particularly relevant for full-band ADSR [12], which is the primary focus of the present work, where conventional objective metrics and human listeners often struggle to capture the subtle, yet perceptually important differences between real and synthesized signals.

The rest of the paper is organized as follows. The GAN-based ADSR and architectures are outlined in Sec. 2. The concept of ‘real’/‘fake’ classifiers is explained in Sec. 3. Evaluation results including objective metrics, MUSHRA (Multi-Stimulus Test with Hidden Reference and Anchor) listening test [23] for the full-band ADSR and classifier accuracy are reported in Sec. 4. Conclusions are presented in Sec. 5.

2. AUDIO SUPER-RESOLUTION MODELS

Given audio signal $\mathbf{x} = [x(n/f_s)]$, $n = 0, \dots, N$ the ADSR task aims to reconstruct an upsampled signal $\mathbf{y} = [y(m/rf_s)]$, $m = 0, \dots, rN$, where r is the upsampling ratio. The input and upsampled signals thus have lengths N and rN , respectively, but share the same duration N/f_s . Increasing the sampling rate $f_s \rightarrow rf_s$ expands the Nyquist frequency from $f_s/2$ to $rf_s/2$, allowing new frequency components in the band $[f_s/2, rf_s/2]$. The original and upsampled signals are referred to as narrowband (NB) and wideband (WB), respectively.

Deep learning methods for ADSR have advanced rapidly, evolving from supervised models such as AudioUNet [22] to GAN-based and diffusion-based generative models [18, 19, 20, 21, 24]. In this work, we adopt the MU-GAN (Multi-scale U-Net GAN) architecture, originally proposed for the $4 \rightarrow 16$ kHz ADSR [19]. We further extend MU-GAN to the full-band $16 \rightarrow 48$ kHz ADSR. Due to limited implementation details and the absence of publicly available code, we re-implemented, trained, and evaluated the model from scratch. As a supervised baseline, we employ the AudioUNet model [22] for both $4 \rightarrow 16$ kHz and $16 \rightarrow 48$ kHz upsamples. The trained MU-GAN discriminator features are used as embeddings for the ‘real’/‘fake’ classification task across all considered models.

Both AudioUNet and MU-GAN models were imple-

mented in PyTorch¹ and trained using the VCTK speech dataset [25] for $4 \rightarrow 16$ kHz and $16 \rightarrow 48$ kHz ADSR, and the FMA-small music dataset [26] for $16 \rightarrow 48$ kHz ADSR. We used the official train/validation/test splits for the FMA dataset [26] and the train/test split for the VCTK dataset, following the protocol used in previous works [12, 19, 22]. Training was performed on a single NVIDIA A100 GPU. For the VCTK dataset, both models were trained for 500 epochs. For the FMA-small dataset, AudioUNet and MU-GAN were trained for 100 and 80 epochs, respectively, with early stopping. The learning rate was set to 10^{-4} and the mini-batch size to 128. The generator and discriminator were optimized using Adam and stochastic gradient descent (SGD), respectively. To stabilize adversarial training, we adopted a scheduled update strategy, where the generator was updated more frequently than the discriminator. For the converged MU-GAN models, the discriminator achieved an accuracy of approximately 51% on the VCTK dataset and 49% on the FMA dataset, indicating that the model reached the desired equilibrium characteristic of well-trained GANs, where real and generated samples become nearly indistinguishable.

To enable comparison with state-of-the-art methods for full-band ADSR, we use HiFi-GAN, originally proposed for denoising and bandwidth extension [12], and two recent diffusion-based models, FlowHigh [13] and FlashSR [14], both capable of upscaling audio from arbitrary input sampling rates to 48 kHz. For our experiments, we use the publicly available unofficial HiFi-GAN implementation with pretrained weights for inference [24], and the official inference code released for FlowHigh² and FlashSR³.

3. ‘REAL’ / ‘FAKE’ AUDIO CLASSIFIERS

Classifiers capable of distinguishing real and synthetic audio clips are trained and evaluated using several labeled embedding datasets. These datasets are constructed by transforming real and synthetic audio signals into fixed-length embeddings using different feature extractors. Several types of embeddings are considered. First, to study the quality of learned representations, we use features extracted from the fixed pre-FC layer of the MU-GAN discriminator (Fig. 1a). In our implementation, an 8192-sample input is mapped to a 32-dimensional embedding vector by this discriminator layer.

In addition, two types of external embeddings are employed (Fig. 1b). The OpenL3 [27] model generates a 512-dimensional embedding vector from a 1-second audio segment. This model is suitable for the $4 \rightarrow 16$ kHz ADSR task but cannot be reliably applied to the $16 \rightarrow 48$ kHz setting due to its limited input bandwidth, which prevents it from capturing the full frequency content of 48 kHz audio.

¹<https://github.com/msilaev/ADRS>

²<https://github.com/resemble-ai/flowhigh>

³https://github.com/jakeoneijk/FlashSR_Inference

To address the bandwidth limitations of OpenL3, log-Mel spectrogram energies are used as an alternative feature representation. We use 256 Mel-frequency bins, an FFT size of 4096, and a hop length of 256. The upper frequency limit is set by the Nyquist frequency, corresponding to 8 kHz and 24 kHz for input sampling rates of 16 kHz and 48 kHz, respectively. To produce fixed-length embeddings that are independent of input duration, adaptive average pooling is applied along the temporal dimension.

Table 1: LSD and SNR metrics for different models

Model	VCTK			FMA		
	LSD	SNR	LSD	SNR	LSD	SNR
	4→16	4→16	16→48	16→48	16→48	16→48
AudioUnet	4.5	15.4	4.2	22	9.2	24.5
MU-GAN	3.9	14.6	4.2	20.8	6.7	27.3
HiFi-GAN	–	–	2.1	17.5	–	–
FlowHigh	–	–	3	-6.8	3.6	-3
FlashSR	–	–	3.9	16	8.4	18.2

4. EVALUATION AND RESULTS

To evaluate the performance and perceived realism of ADSR models, we combine objective signal comparison metrics, signal-to-noise ratio (SNR), and logarithmic spectral distance (LSD) with MUSHRA listening tests.

4.1. Evaluation of Signal-Level Performance

Table 1 summarizes the SNR and LSD results, mainly to show that no unexpectedly low or high values are observed. The $4 \rightarrow 16$ kHz ADSR results are consistent with previous work [19, 22], and, as in those studies, LSD is computed using the natural logarithm.

A notable observation is that the FlowHigh model yields negative SNR values while achieving lower LSD than AudioUNet and MU-GAN. The negative SNR does not indicate degraded perceptual quality but instead arises from a global amplitude scaling factor, which may vary across samples.

No significant differences are observed between HiFi-GAN, AudioUNet, and MU-GAN scores. However, as shown below, the SNR and LSD metrics correlate poorly with human listening results, which reveal clear perceptual differences between models’ outputs.

4.2. MUSHRA listening test

For evaluation, 12 recordings were randomly selected from the VCTK test set, ensuring an equal number of male and female speakers. Original WB recordings at 48 kHz were down-sampled to 16 kHz and then up-sampled back to 48

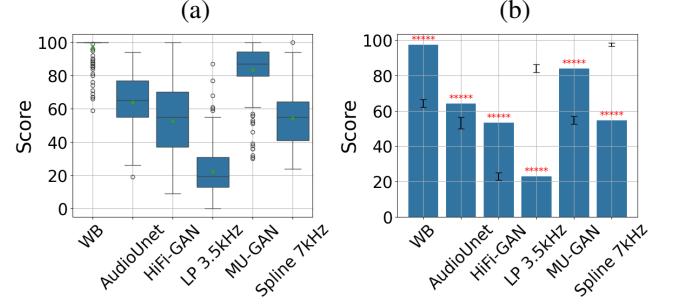


Fig. 2: Listener scores for different ADSR methods. MUSHRA scores for listening tests across different conditions WB, MU-GAN, AudioUnet, HiFi-GAN, LP 3.5 kHz, Spline-Up 7 kHz. (a) Inter-quartile range (IQR), medians, and mean values by green triangles. (b) Mean scores with error bars representing 95% confidence intervals.

kHz using three different ADSR models: AudioUnet, MU-GAN, and HiFi-GAN. Each recording was supplemented by two anchor signals: a low-pass filtered version at 3.5 kHz and a middle-pass filtered version at 7 kHz. Furthermore, the original NB recording was included. This yields six experimental conditions. The listening test was implemented using the MUSHRA listening test interface, which allowed listeners to set loops if they wanted to focus on particular short passages of the audio signal.

Listening test results are shown in Fig. 2 following the MUSHRA recommendations [23]. The box plot in Fig. 2a shows the distribution of MUSHRA scores for six conditions: WB, MU-GAN, AudioUnet, HiFi-GAN, LP 3.5 kHz, and Spline-Up 7 kHz, highlighting median, IQR, mean values, and outliers. The bar graph in Fig. 2b shows the mean scores with 95% confidence intervals. MU-GAN achieves the highest score, closely matching the WB reference, while HiFi-GAN performs the worst, closely resembling the 7 kHz anchor. AudioUnet performs slightly better. Non-overlapping confidence intervals in Fig. 2b show that listeners can reliably distinguish between real and synthetic audio, even though the perceptual quality of the MU-GAN outputs remains comparable to that of the target WB recordings. In the next section, we demonstrate that this distinction is also captured by binary classifiers.

4.3. ‘Real’/‘Fake’ Classifier Accuracy

The embedded datasets are constructed using audio clips from the test subsets of the VCTK (speech) and FMA-small (music) datasets, which were not seen during the training of any considered model. Each embedding dataset is randomly shuffled and split into training (80%) and testing (20%) subsets. Before classification, the feature vectors are standardized to have zero mean and unit variance. A linear discriminant analysis (LDA) classifier is then trained on the normalized train-

ing data and evaluated on the test data to assess its ability to distinguish between real and fake samples. Additionally, the test embeddings are projected into the discriminant subspace learned by LDA, where class separation is maximized. This projection enables a qualitative visualization of class overlap and separability along a single discriminant axis.

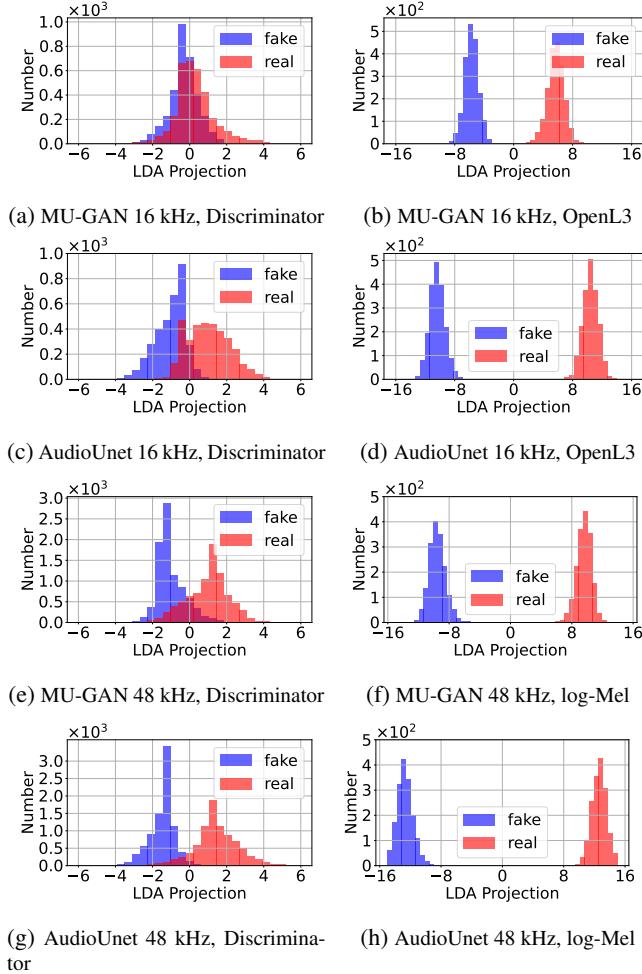


Fig. 3: LDA projections for audio clips from VCTK test dataset, calculated using MU-GAN discriminator features (left column), OpenL3 (right column b,d) and log-Mel (right column f,h) embeddings. (a-d) $4 \rightarrow 16$ kHz and (e-h) $16 \rightarrow 48$ kHz ADSR.

Example LDA distributions for different embedding spaces are shown in Fig. 3. Panels corresponding to (a,c,e) discriminator pre-FC feature embeddings show completely overlapping distributions for the $4 \rightarrow 16$ kHz MU-GAN, partial overlap for the $16 \rightarrow 48$ kHz MU-GAN, and almost complete separation for AudioUNet. On the one hand, this demonstrates that the discriminator learns useful features capable of distinguishing real from synthetic clips. On the other hand, the clear difference in distribution overlap be-

tween MU-GAN and AudioUNet indicates that the generator is trained to make these distributions more similar.

In contrast, panels (b,d) and (f,h) show that LDA projections of the OpenL3 and log-Mel embeddings demonstrate complete 'real'/fake' class separation, achieving 100% classification accuracy. Note that panels (e,f) and (g,h) correspond to MU-GAN and AudioUNet models for $16 \rightarrow 48$ kHz upsampling studied using MUSHRA test reported in Sec. 4.2.

Table 2: Binary 'fake'/'real' classifier accuracy (%) based on MU-GAN discriminator embeddings. Columns correspond to ADSR $4 \rightarrow 16$ kHz on VCTK, $16 \rightarrow 48$ kHz on VCTK, $16 \rightarrow 48$ kHz on FMA-small datasets.

Model	VCTK 4→16	VCTK 16→48	FMA 16→48
AudioUnet	80%	95%	78%
MU-GAN	56%	83%	70%
HiFi-GAN	–	93%	–
FlowHigh	–	85%	74%
FlashSR	–	88%	66%

These results hold qualitatively for all models considered. As shown in Table 2, the learned MU-GAN discriminator embeddings yield high accuracies (around 90%) for the AudioUNet, FlashSR, and FlowHigh models on the VCTK dataset, and slightly lower (around 70%–80%) for the FMA dataset. At the same time, the log-Mel and OpenL3 embeddings achieve perfect classification accuracy (100%) across all tasks and models considered.

5. CONCLUSIONS

This study highlights a gap between traditional signal metrics, perceptual quality, and the separability of real and super-resolved audio distributions produced by generative ADSR models. During stable GAN training, the discriminator accuracy converges to around 50%, yet its learned representations can still distinguish 'real' and 'fake' audio in a supervised benchmark, indicating that the GAN captures comprehensive data features. Listener scores in the MUSHRA test closely match the target wideband audio.

Classifiers trained on external embeddings, such as OpenL3 or log-Mel spectrograms, achieve nearly perfect separation between real and generated clips. This behavior is consistent across domains, sampling rates ($4 \rightarrow 16$ kHz and $16 \rightarrow 48$ kHz), and extends to state-of-the-art diffusion models [13, 14]. These results suggest that high perceptual quality does not necessarily imply accurate distribution modeling, revealing a persistent gap between perceptual realism and representational fidelity—an open challenge for future ADSR research.

6. REFERENCES

- [1] I. Goodfellow et al., “Generative adversarial nets,” *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.
- [2] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv:1511.06434*.
- [3] B. Srivastava et al., “Privacy-preserving adversarial representation learning in asr: Reality or illusion?,” in *Interspeech*, 2019, pp. 3700–3704.
- [4] S. Gharib et al., “Adversarial representation learning for robust privacy preservation in audio,” *IEEE Open J. Signal Process.*, vol. 5, pp. 294–302, 2024.
- [5] M. Heusel et al., “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [6] M. Lucic et al., “Are GANs created equal? a large-scale study,” *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
- [7] M. Sajjadi et al., “Assessing generative models via precision and recall,” *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
- [8] K. Kilgour et al., “Frechet audio distance: A metric for evaluating music enhancement algorithms,” *arXiv:1812.08466*.
- [9] M. Binkowski et al., “High fidelity speech synthesis with adversarial networks,” *arXiv:1909.11646*.
- [10] T. Kynkänniemi et al., “Improved precision and recall metric for assessing generative models,” *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
- [11] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *CVPR - IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
- [12] J. Su et al., “Bandwidth extension is all you need,” in *ICASSP - IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2021, pp. 696–700.
- [13] J. Im and J. Nam, “Flashsr: One-step versatile audio super-resolution via diffusion distillation,” in *ICASSP - IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2025, pp. 1–5.
- [14] J. Yun, S. Kim, and S. Lee, “Flowhigh: Towards efficient and high-quality audio super-resolution with single-step flow matching,” in *ICASSP - IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2025, pp. 1–5.
- [15] D. Gupta and H. Shekhawat, “Artificial bandwidth extension using H_∞ optimization, deep neural network, and speech production model,” in *SPCOM - IEEE Int. Conf. Signal Process. Commun.*, 2022, pp. 1–5.
- [16] H. Liu et al., “Audiosr: Versatile audio super-resolution at scale,” in *ICASSP - IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1076–1080.
- [17] X. Li et al., “Speech audio super-resolution for speech recognition,” in *Interspeech*, 2019, pp. 3416–3420.
- [18] E. Moliner and V. Välimäki, “Behm-gan: Bandwidth extension of historical music using generative adversarial networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 943–956, 2022.
- [19] S. Kim and V. Sathe, “Bandwidth extension on raw audio via generative adversarial networks,” *arXiv:1903.09027*.
- [20] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 17022–17033, 2020.
- [21] Y. Li et al., “Real-time speech frequency bandwidth extension,” in *ICASSP - IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 691–695.
- [22] V. Kuleshov, S. Z. Enam, and S. Ermon, “Audio super resolution using neural networks,” *arXiv:1708.00853*.
- [23] ITU-R, “Method for the subjective assessment of intermediate quality level of audio systems,” Tech. Rep. BS.1534-3, International Telecommunication Union, Geneva, Switzerland, Oct 2015.
- [24] B. Spell, “Hifi-gan-bwe: High-fidelity bandwidth extension,” <https://pypi.org/project/hifi-gan-bwe/>, 2025.
- [25] J. Yamagishi et al., “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” *Univ. Edinburgh CSTR*, pp. 271–350, 2019.
- [26] M. Defferrard et al., “FMA: A Dataset for Music Analysis,” in *ISMIR - Int. Soc. Music Inf. Retr. Conf.*, 2017, pp. 316–323.
- [27] A. Cramer et al., “Look, listen, and learn more: Design choices for deep audio embeddings,” in *ICASSP - IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 3852–3856.