REGRESSION TO CLASSIFICATION: WAVEFORM ENCODING FOR NEURAL FIELD-BASED AUDIO SIGNAL REPRESENTATION

TaeSoo Kim[*1,2], *Daniel Rho*[*1], *Gahui Lee*[1], *JaeHan Park*[1], *and Jong Hwan Ko*[2]
[1]KT, Republic of Korea
[2]Department of Electrical and Computer Engineering, Sungkyunkwan University, Republic of Korea

ABSTRACT

Neural fields, also known as coordinate-based representations, are an emerging signal representation framework. This approach has also been used to represent audio signals, but the generated audio often contains noise. To reduce noise and improve representation quality, we propose using waveform encoding in the neural field. Instead of yielding real numbers for each temporal coordinate, this involves using discrete integers as outputs, with waveform-encoded integers as target classes, and treating the representation problem as a classification task rather than a regression problem. The experimental results show that waveform encoding can improve the audio quality of neural fields across a variety of audio datasets.

***Index Terms*—** neural fields, implicit neural representation, audio representations, waveform coding

1. INTRODUCTION

The temporal resolution of audio signals created by neural networks is often fixed during training and testing. However, neural fields, also referred to as coordinate-based representations, enable the generation of signals at any resolution by inserting corresponding continuous coordinates during inference without requiring additional training. Not only that, neural fields have demonstrated their ability to accurately represent arbitrary signals with fine details and have been successfully used to represent various types of signals such as audio, image, and video signals [1, 2]. Also, it is widely used for neural rendering, which aims to represent the whole 3D or even 4D scene [1, 3].

This neural representation scheme is still in its early stages in the audio domain. More specifically, neural fields have just started being used for a variety of audio-related tasks, including audio signal representation [2], super resolution [4], and audio synthesis [5]. However, noise in the audio signals produced by neural fields is a common problem, and there are currently only a few studies on how to improve the quality of the audio signals produced by neural fields.

We propose to utilize waveform encoding in the neural field to reduce noise and represent high-quality audio signals.

*Equal Contribution

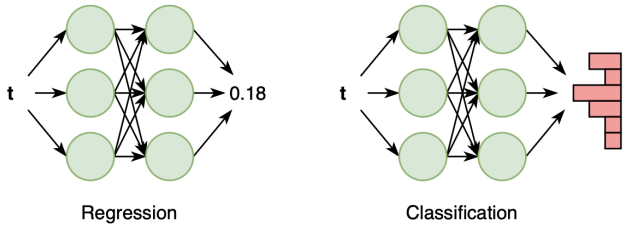

Fig. 1. Overall illustration of our proposed method. Ours solves the audio signal representation task using classification instead of regression.

Noise, which is often perceptually unpleasant, is a persistent issue for neural fields. Waveform encoding, which converts continuous analog signals to digital signals, is widely used in telecommunication systems to prevent signal quality from being degraded by noise during transmission. Since waveforms can be efficiently quantized with low bit through waveform encoding, we used the softmax distribution as the output of neural fields. Fig 1 illustrates the overall structure of the proposed method. As a result, we were able to train the neural fields model to represent higher-quality audio signals than the baseline models.

2. RELATED WORKS

2.1. Neural fields

Neural fields, also known as implicit neural representation, is one of the neural network-based signal representation schemes. It uses neural networks to model a function of coordinates, $f(x) = y$, mapping coordinates to their corresponding values. To model the function of coordinates, neural fields frequently employ variants of simple multilayer perceptrons (MLPs) with non-linearity activation. Neural fields sample values for each coordinate, which frequently requires the use of complex and heavy networks that consume a lot of computing power. However, naively implementing and training neural networks to map coordinates to values fails to capture high-frequency details due to extremely low-dimensional inputs (for example, one for audio and two for an image) and spectral biases [2, 6]. Numerous lines of

study have been proposed to address this problem, including positional encoding [6, 1, 7, 8], internal activation function and parameter initialization [2], to name a few. NeRF [1] proposed to use positional encoding [9], which was first proposed in the natural language understanding, for mapping coordinates into a much higher dimensional space. This made it possible for neural fields to approximate a higher frequency function. *Tancik et al.* [6] propose to use the fourier feature mapping to overcome the spectral bias of MLPs that prevent learning high-frequency details. On the other hand, other lines of research, such as SIREN [2], that do not preprocess nor map low dimensional coordinates into a much higher dimensional space, address this problem by designing the inner structure of neural networks, such as weight initialization and activation function. More specifically, SIREN propose to use periodic activation function with calculated initialization policy. The widespread use of neural fields in various types of signals is a result of these advancements.

In audio domain, SIREN [2] has shown that neural fields can accurately express an arbitrary audio signal with fine details. *Zuiderveld et al.* [4] demonstrated the utility of this representation scheme in the audio synthesis task by incorporating conditioning methods and employing neural fields as an auto-decoder [10]. LISA [5] utilizes neural fields to provide arbitrary scale super-resolution. They employ an encoder and a decoder to extract local latent codes from audio signal chunks and, respectively, predict signal value from coordinates.

2.2. Waveform Coding

Waveform encoding has been proposed for low-bit audio signal transmission while maintaining high perceptual audio quality. Simple linear quantization of audio signals frequently results in undesirable noise. Additionally, due to a number of unexpected or expected factors, analog signal transmission is frequently accompanied by unwanted noise. To alleviate the effects of noise, waveform encoding consists of two steps. Waveform encoding initially applies a non-linear function, also known as companding. After that, these converted signals are quantized into finite segments. Through these processes, waveform encoding translates continuous analog audio signals into finite, discrete digital signals [11].

In the early stages of audio synthesis using deep neural networks, WaveNet [12] was the first to employ waveform encoding, specifically μ-law companding. They generate new audio by auto-regressively predicting quantized audio samples based on the softmax distribution of the subsequent timestamp. But these are limited to non-neural field-methods and yet to be applied in neural fields.

3. METHOD

We propose using finite discrete integers as outputs rather than generating a real number for each temporal coordinate. To convert continuous signal values as quantized integers, μ-law and A-law encoding were used. Each quantized output is designated to a class, and neural networks are trained to predict the probability of each class rather than the scaler value. In short, we propose to change the neural field-based audio representation from a regression task to a classification task. Even when the data is implicitly continuous, a softmax distribution usually performs well as it is more flexible and can easily model arbitrary distributions [13, 12]. Therefore, we used cross entropy loss to train our proposed method.

3.1. Encoding

There are two standard companding methods; μ-law companding and A-law companding. μ-law companding is defined as follows,

$$M(x) = sgn(x)\frac{ln(\mu|x|+1)}{ln(\mu+1)}, -1 \leq x \leq 1 \tag{1}$$

sgn, $|\cdot|$ and $\lfloor\cdot\rceil$ are sign, absolute and rounding functions, respectively. The standard μ-law companding sets μ to 255.

A-law companding is defined as follows,

$$A(x) = sgn(x)\begin{cases}\frac{A|x|}{ln(A)+1}, |x| < \frac{1}{A} \\ \frac{ln(A|x|)+1}{ln(A)+1}, \frac{1}{A} \leq |x| \leq 1.\end{cases} \tag{2}$$

ln denotes natural logarithm, and A is a constant.

We employ waveform encoding to discretize and categorize continuous values. For μ-law encoding, the quantization equation is as follows,

$$f(x) = \lfloor\frac{(C(x)+1)\nu}{2} + 0.5\rceil. \tag{3}$$

$C(\cdot)$ denotes a companding function. In the case of μ-law encoding, it is $M(\cdot)$ in Eq. 1, and in the case of A-law encoding, it is $A(\cdot)$ in Eq. 2. This function maps an arbitrary continuous value between -1 and 1 to an integer between 0 to ν. Integers will be used as class labels respectively in our proposed method.

3.2. Decoding

The integer with the highest prediction probability is chosen in the inference phase and converted back into a real value using the following equation:

$$f^{-1}(\epsilon) = \frac{2\epsilon}{\nu} - 1. \tag{4}$$

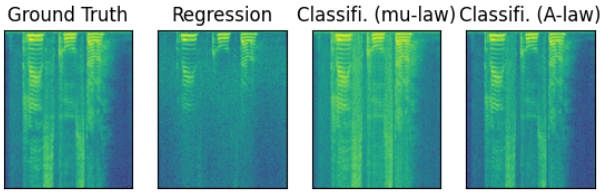

Fig. 2. The target log-magnitude spectrogram and three generated spectrograms. Using continuous values as outputs (regression) results in less accurate high-frequency components, compared to using discrete values as outputs (classifi.).

ϵ is the quantized value obtained from Eq. 3, and ν is the same ν in Eq. 3. Because Eq. 4 only converts integers to real numbers between -1 and 1, it should be possible to decode quantized integers into real numbers depending on which encoding function had been used.

We used the official inverse function of each waveform companding function for decoding [11]. The inverse of μ-law companding (Eq. 1) is as follows,

$$M^{-1}(y) = sgn(y)\frac{(\mu+1)^{|y|}-1}{\mu}. \tag{5}$$

Similarly, the inverse of A-law companding (Eq. 2) is as follows,

$$A^{-1}(y) = sgn(y)\begin{cases}\frac{|y|(ln(A)+1)}{A}, |y| < \frac{1}{ln(A)+1} \\ \frac{e^{|y|(ln(A)+1)-1}}{A}, \frac{1}{ln(A)+1} \leq |y| \leq 1.\end{cases} \tag{6}$$

3.3. Objective Function

We used the mean square error (MSE) loss as the objective function on the regression task, in which neural fields generate continuous audio signal values. Because neural network outputs are probabilities of classes, or quantized integers in our case, we used cross entropy for the proposed method. In other words, we optimized neural fields by classification rather than regression.

4. EXPERIMENTS AND RESULTS

We evaluate our proposed method on two main tasks; audio signal representation and super-resolution. We used three different models to see if the proposed method could be applied to a range of network structures; SIREN, Fourier neural networks (FNN) [6], and lastly, MLP with positional encoding (PosEnc) [1]. We used VCTK [14] and ESC-50 [15] as train and evaluation datasets.

Data	***Outputs***	***Model***	***PESQ*↑**	***SNR*↑**	***LSD*↓**
VCTK	*Floats*	*SIREN*	2.299	20.154	**0.299**
		FNN	1.342	14.785	0.308
		Pos Enc.	1.701	17.148	0.669
	μ-Law	*SIREN*	**2.948**	**21.042**	0.981
		FNN	3.893	36.585	0.084
		Pos Enc.	**2.434**	**20.926**	**0.505**
	A-Law	*SIREN*	2.921	20.627	1.032
		FNN	**3.937**	**36.857**	**0.080**
		Pos Enc.	2.416	20.737	0.517
ESC50	*Floats*	*SIREN*	3.253	19.331	0.399
		FNN	2.534	15.099	**0.444**
		Pos Enc.	2.569	13.381	**1.6342**
	μ-Law	*SIREN*	4.074	29.854	0.329
		FNN	**3.579**	19.886	0.789
		Pos Enc.	**3.049**	**15.812**	2.827
	A-Law	*SIREN*	**4.089**	**29.982**	**0.329**
		FNN	3.566	**19.967**	0.798
		Pos Enc.	3.039	15.499	2.741

Table 1. Performance in the audio signal representation task. *Floats* in the "Outputs" column indicates neural fields that have been trained with regression loss.

4.1. Datasets

To validate the proposed method, we used the voice bank corpus (VCTK version 0.92) [14] as a human speech dataset and the environmental sound classification dataset (ESC-50) [15] as an environmental sound dataset. The VCTK corpus was recorded by 110 English speakers with various accents and has a sampling rate of 48 kHz. We used audio samples recorded with *mic 1* in the official test dataset. The recordings of the speakers *p280* and *p315* were excluded due to technical issues. The ESC-50 dataset contains 2,000 audio streams with a 5-second length containing 50 classes for environmental acoustic events, each recorded at 44.1 kHz. We used all the audio samples from ESC-50 for evaluation.

4.2. Experimental Setup

For experiments, we used a 4-layer MLP with a width size of 512 as a backbone model. We use the same network architecture for SIREN, FNN, and PosEnc since the only differences among these methods are the initialization policy and inner activation functions, as well as how coordinates are processed. For SIREN, we set ω_0 and ω to 3,000 and 30 respectively, following the original paper. In the case of FNN, we set the σ and scale to 8,000 and 1,000, respectively. The number of frequencies was automatically set so that the maximum frequency of positional encoding does not exceed the Nyquist frequency of the target audio signal. More precisely, the number of frequencies was set to $\lfloor log_2(n\_samples/2) \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function and $n\_samples$ is the total number of samples in the target audio sample. Using the Adam op-

timizer, we train each method for 2,000 iterations, decaying the learning rate by 0.95 every 100 iterations.

We used three audio evaluation metrics to evaluate the performance of each method: perceptual evaluation of speech quality (PESQ) [16], signal-to-noise ratio (SNR), and log-spectral distance (LSD) [17]. PESQ is an optimized evaluation method for evaluating the quality of human voices. LSD, a commonly used metric for audio super-resolution or quantization, measures the distance between two audio samples in the frequency domain.

4.3. Audio Signal Representation

To evaluate the effect of output types on the performance of audio signal representation, we trained several neural fields (SIREN, FNN, and PosEnc) using different outputs. The sampling rate of every audio sample was set to 16 kHz. As shown in the Table 1, introducing waveform encoding enhances audio quality. In VCTK, waveform encoding improves the performance by a significant margin regardless of the neural architecture. Additionally, in terms of all three criteria, FNNs with A-law companding perform the best in representing the VCTK dataset (PESQ from 1.342 to 3.937, SNR from 14.785 to 36.857, and LSD from 0.308 to 0.080 in the case of FNN). Overall, μ-law and A-law encodings show similar representation performance. Even for the ESC-50 dataset, waveform-encoding models produce outputs that are significantly more qualitatively enhanced as measured in SNR and PESQ. Since the quantization levels chosen for waveform encoding are highly adapted to low sampling rates, it is inevitable that high-frequency components will be difficult to represent, leading to high LSD. The FNN and PosEnc models only differ in the input preprocessing methods they use, which demonstrates that the model's capacity to represent high frequency components has a significant impact on it regardless of whether waveform encoding is employed. According to the overall results, it can be seen that neural field models represent audio signals with higher quality when waveform encoding is applied.

In addition, we qualitatively compared regression and classification-based methods. Fig. 2 shows the log magnitude spectrogram of each method and the ground truth. We used PosEnc and ran it on an audio sample from VCTK. Although the classification-based method uses quantized outputs, as shown in the figure, the outputs produce a pattern that is similar to the ground truth at both low and high frequencies. Additionally, using waveform encoding and switching from a regression task to a classification task demonstrates a far better capacity to suppress signals when they should be, as seen in the right part of each spectrogram.

4.4. Audio Super-resolution

In this experiment, we evaluate the effect of waveform encoding on neural field on the audio super-resolution. We trained each neural network at the low sampling rate of 16 kHz and tested them at the high sampling rate of 44.1 kHz. In the same way as the audio presentation experiment, all experimental conditions—aside from sampling rate—were carried out. As demonstrated in Table 2, introducing waveform encoding improves the value of PESQ in both datasets. When compared to Table 1, it can be seen that all models have lower scores for all assessment methods across all datasets. These results imply that the neural field models render a high-frequency component prediction error when predicting missing high-resolution components. Plus, PESQ has a stronger correlation to perceptual quality than LSD and SNR [18, 7]. Therefore, it can be seen that the neural field that employs waveform encoding is able to predict the super-resolution component of the audio signal in a more perceptually friendly manner.

Data	***Outputs***	***Model***	***PESQ*↑**	***SNR*↑**	***LSD*↓**
VCTK	*Floats*	*SIREN*	2.296	**16.532**	**0.516**
		FNN	1.104	2.598	2.938
		Pos Enc.	1.726	**14.857**	**0.885**
	μ-Law	*SIREN*	**2.577**	13.926	1.403
		FNN	3.199	15.732	0.584
		Pos Enc.	1.583	11.062	1.966
	A-Law	*SIREN*	2.542	13.774	1.462
		FNN	**3.248**	**15.739**	**0.579**
		Pos Enc.	**2.213**	13.987	1.117
ESC50	*Floats*	*SIREN*	3.226	**12.653**	**0.681**
		FNN	2.171	6.247	**1.192**
		Pos Enc.	2.564	**10.844**	**1.717**
	μ-Law	*SIREN*	**3.797**	10.444	1.753
		FNN	**3.967**	10.831	1.658
		Pos Enc.	2.962	8.211	3.628
	A-Law	*SIREN*	3.754	10.332	1.768
		FNN	3.949	**10.837**	1.658
		Pos Enc.	**2.972**	8.234	3.648

Table 2. Performance in the audio super resolution task (16 kHz $\rightarrow$ 44.1 kHz). *Floats* in the "Outputs" column indicates neural fields that have been trained with regression loss.

5. CONCLUSION

In this paper, we propose using discrete outputs instead of continuous values for representing audio signals with neural field models. With the aid of waveform encoding, audio signals can be expressed using a limited set of integers, which can be regarded as a class. To this end, waveform encoding allows neural field models to be trained in a classification approach. It enhances the capacity of neural field models to represent high-quality audio signals across audio datasets, regardless of neural architecture. Furthermore, it has been demonstrated through audio super-resolution experiments that the distribution of continuous data can still be easily modeled even with softmax distributions.

6. REFERENCES

[1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[2] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 7462–7473, Curran Associates, Inc.

[3] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, July 2022.

[4] Jan Zuiderveld, Marco Federici, and Erik J Bekkers, "Towards lightweight controllable audio synthesis with conditional implicit neural representations," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[5] Jaechang Kim, Yunjoo Lee, Seunghoon Hong, and Jungseul Ok, "Learning continuous representation of audio for arbitrary scale super resolution," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3703–3707.

[6] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 7537–7547, Curran Associates, Inc.

[7] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

[8] Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-hornung, and Daniel Cohen-or, "Sape: Spatially-adaptive progressive encoding for neural optimization," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, pp. 8820–8832, Curran Associates, Inc.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[10] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[11] ITU-T, "Recommendation g. 711. pulse code modulation (pcm) of voice frequencies," 1988.

[12] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[13] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu, "Pixel recurrent neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1747–1756.

[14] Cassia Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[15] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[16] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P 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. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[17] Augustine Gray and John Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[18] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 3013–3027, 2011.