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

TaeSoo Kim*† ‡, Daniel Rho* ‡, Gahui Lee ‡, JaeHan Park ‡, and Jong Hwan Ko †

KT Corporation ‡
Sungkyunkwan University †

Introduction

- **Audio signal representations with the neural fields**
  - Known as coordinate-based or implicit neural representations, are versatile tools used for various signal types such as audio, images, and videos
  - Challenges remain in representing audio signals with noise and reduced high-frequency component quality
- **Contributions**
  - Employing softmax distribution for training neural field models instead of continuous values
  - Enhancing the quality of represented audio signals
  - Minimizing subsequent noises

Related Works

Neural Radiance Fields (NeRF) [1]

- Utilizing positional encoding to map low dimensional coordinates into a much higher dimensional space
- Making it possible for neural fields to approximate a higher frequency function

Sinusoidal Representation Networks (SIREN) [2]

- Designing the inner structure of neural networks, such as weight initialization and activation function
- Using periodic activation function with calculated initialization policy

Fourier Neural Network (FNN) [3]

- Utilizing the fourier feature mapping to overcome the spectral bias of MLPs that prevent learning high-frequency details

Proposed Method

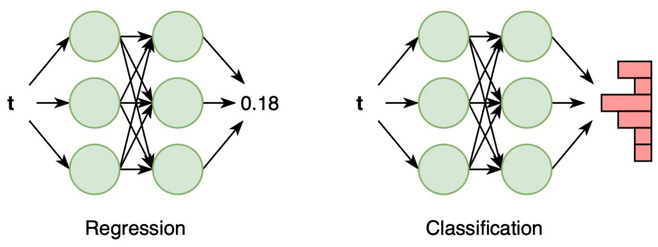

Fig. 1. Overall illustration of our proposed method. Ours solves the audio signal representation task using classification instead of regression.

Training neural fields models with discrete values

- Use discrete values instead of real numbers for temporal coordinates in neural fields models
- Involves predicting the probability of each discrete value class using techniques like softmax
- Softmax is flexible and can handle arbitrary distributions, which is useful when data is implicitly continuous

μ-law companding

- Encoding Function

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

- Decoding Function

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

A-law companding

- Encoding Function

$$A(x) = \operatorname{sgn}(x)\begin{cases}\frac{A|x|}{\ln(A)+1}, |x| < \frac{1}{A} \\ \frac{\ln(A|x|)+1}{\ln(A)+1}, \frac{1}{A} \le |x| \le 1\end{cases}$$

- Decoding Function

$$A^{-1}(y) = \operatorname{sgn}(y)\begin{cases}\frac{|y|(\ln(A)+1)}{A}, |y| < \frac{1}{\ln(A)+1} \\ \frac{e^{|y|(\ln(A)+1)-1}}{A}, \frac{1}{\ln(A)+1} \le |y| \le 1\end{cases}$$

- Using standard μ and A is setted to 255, which is a standard value
- Cross entropy loss is used as the objective function

Experiments

- Two experiments were conducted to evaluate the performance of neural fields models. : audio signal representation and audio super-resolution.
- Three neural fields models were used in the experiments: SIREN, FNN, and MLP with positional encoding (Pos Enc.).
- Two datasets were used to train and test the models: VCTK [4] and ESC-50 [5].
- The performance of the models was evaluated using three metrics: perceptual evaluation of speech quality (PESQ) [6], signal-to-noise ratio (SNR), and log-spectral distance (LSD) [7].

Results

Audio signal representation

- Introducing waveform encoding enhances audio quality significantly, and FNNs with A-law companding perform best for the VCTK dataset.
- For the ESC-50 dataset, waveform-encoding models produce significantly higher quality outputs in terms of SNR and PESQ.
- Overall, neural field models represent audio signals with higher quality when waveform encoding is applied.

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.

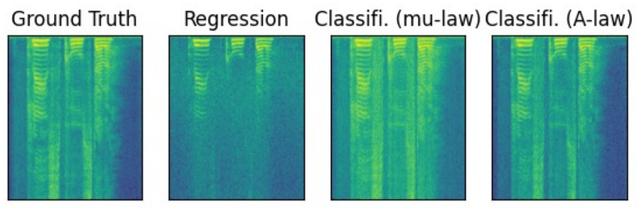

Fig. 2 The target log-magnitude spectrogram and three generated spectrograms

Audio super-resolution

- All neural networks were trained at a low sampling rate of 16 kHz and tested at a high sampling rate of 44.1 kHz.
- Introducing waveform encoding improves the value of PESQ in both datasets, VCTK and ESC-50.
- However, all models have lower scores for all assessment methods across all datasets, indicating high-frequency component prediction errors when predicting missing high-resolution components.

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 → 44.1 kHz).

Conclusion

- This paper proposes using discrete outputs instead of continuous values for representing audio signals with neural field models
- Waveform encoding allows audio signals to be expressed using a limited set of integers, which can be treated as a class for classification-based training of neural field models.
- Waveform encoding enhances the capacity of neural field models to represent high-quality audio signals across datasets and neural architectures.
- Audio super-resolution experiments demonstrate that even with softmax distributions, the distribution of continuous data can still be easily modeled using waveform encoding.

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]Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Sing- hal, 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.

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

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

[6] 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.

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