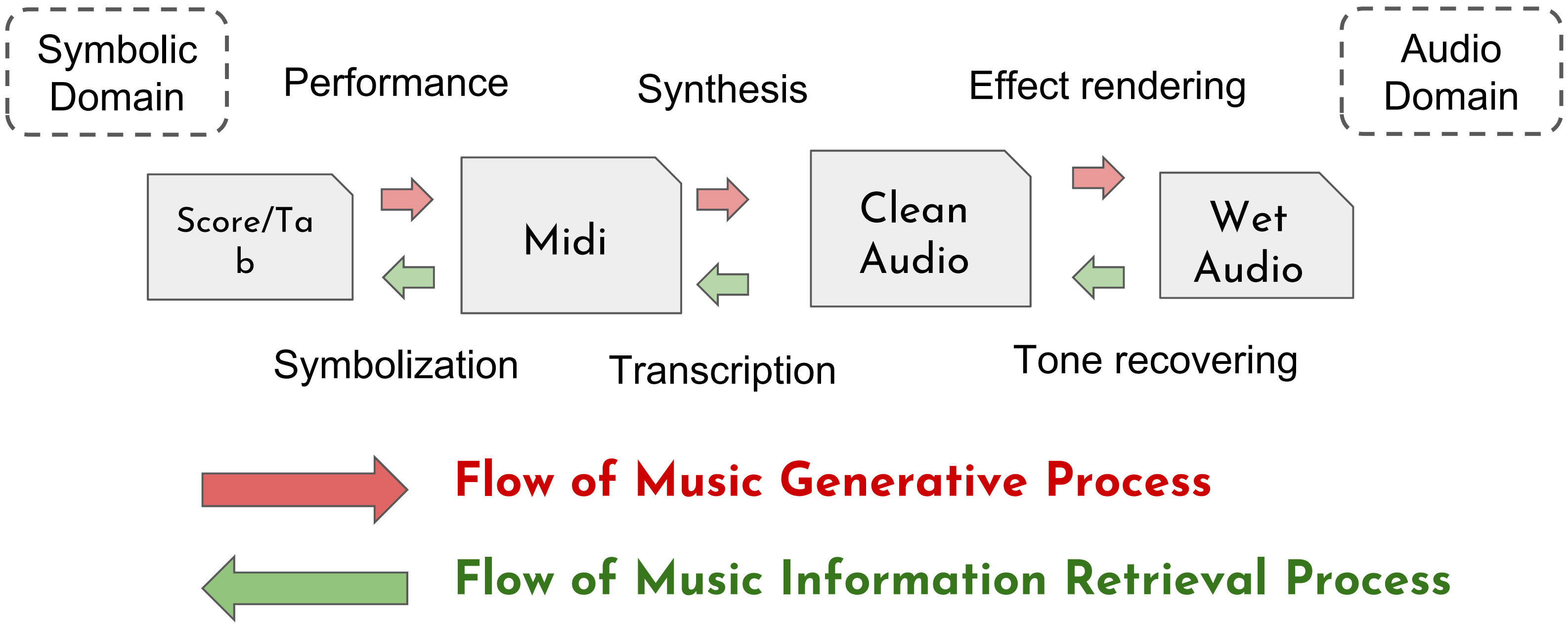
Deep Learning for Music Analysis and Generation

Guitar x ML/DL

Yu-Hua Chen Ph.D. candidate
f08946011@ntu.edu.tw

How we represent *guitar*



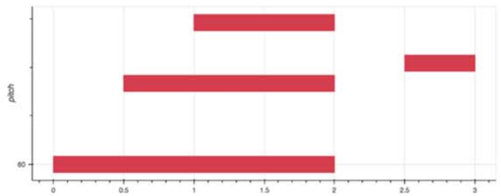
Outline

1. Tabulature modeling and generation
2. Electric guitar transcription (wet audio -> tab)
3. Effect rendering (clean audio -> wet audio)

Tabulature modeling and generation

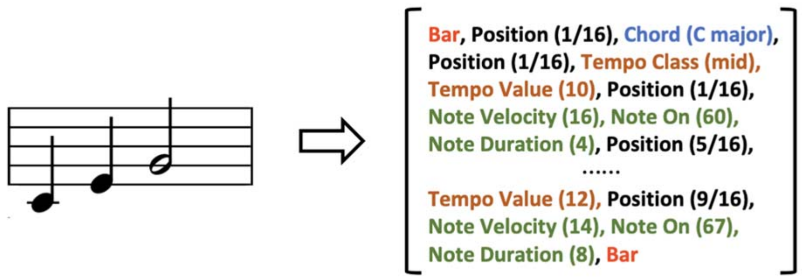
Background

1. Music Transformer
2. REMI



```
SET_VELOCITY<80>, NOTE_ON<60>
TIME_SHIFT<500>, NOTE_ON<64>
TIME_SHIFT<500>, NOTE_ON<67>
TIME_SHIFT<1000>, NOTE_OFF<60>, NOTE_OFF<64>,
NOTE_OFF<67>
TIME_SHIFT<500>, SET_VELOCITY<100>, NOTE_ON<65>
TIME_SHIFT<500>, NOTE_OFF<65>
```

Represent MIDI file using **token**, easier to make it as input for language model

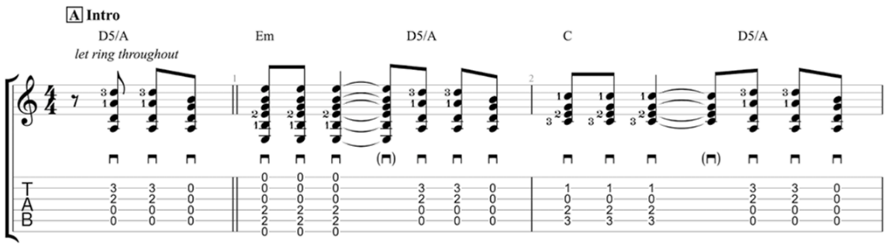


Difference between Tab and MIDI

Tabulature

1. Note onset
2. Note offset
3. Note on fingerboard
4. Chord information (not at all)
5. Quantized time grid
6. No velocity
7. etc..

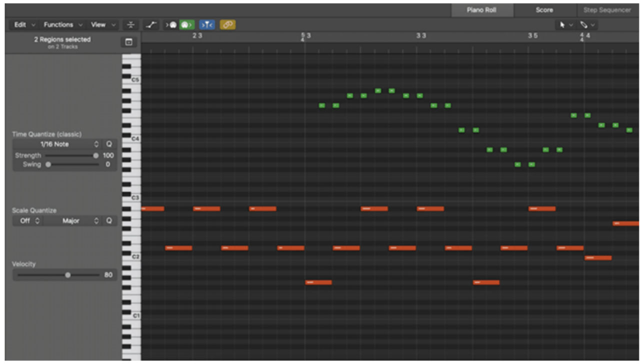
More like a score



Easy Solo Guitar Arrangement

MIDI

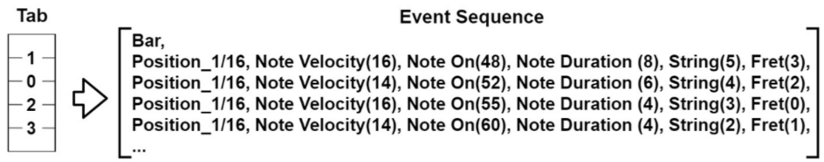
1. Note onset
2. Note offset
3. **With velocity**
4. etc..

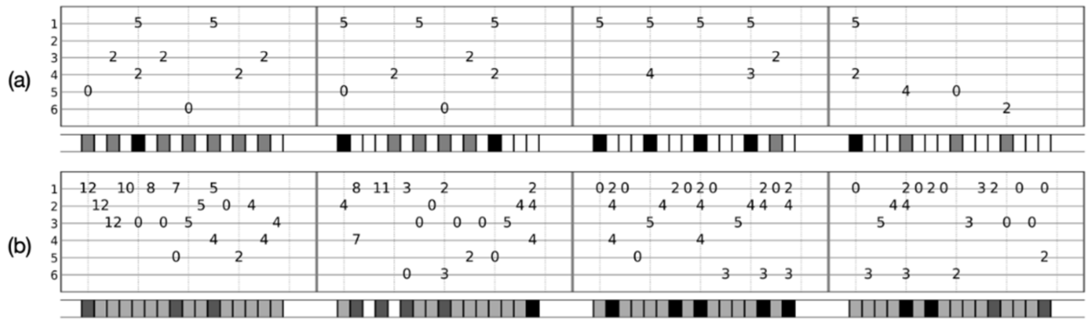


Tabulature modeling and generation

Inspired by REMI and Music Transformer,

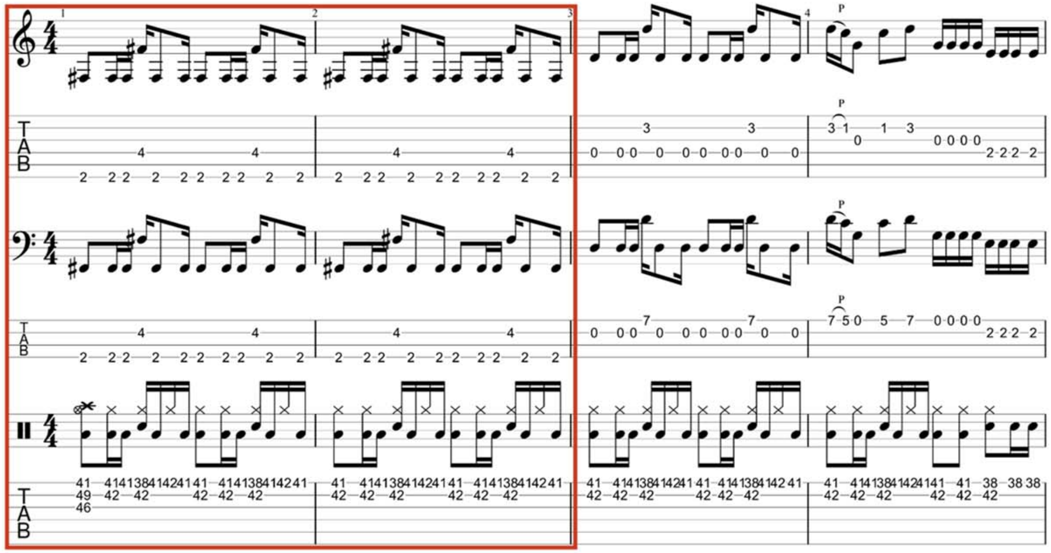
We proposed a **Guitar Tab Generation Model** (ISMIR’20)





Followup works

1. DadaGP dataset (ISMIR’21)
   a. 26181 song scores in the GuitarPro format covering 739 musical genres
2. GTR-CTRL (EvoMUSART'23)
   a. instrument and genre Conditioning
   b. multi tracks
   c. more token types

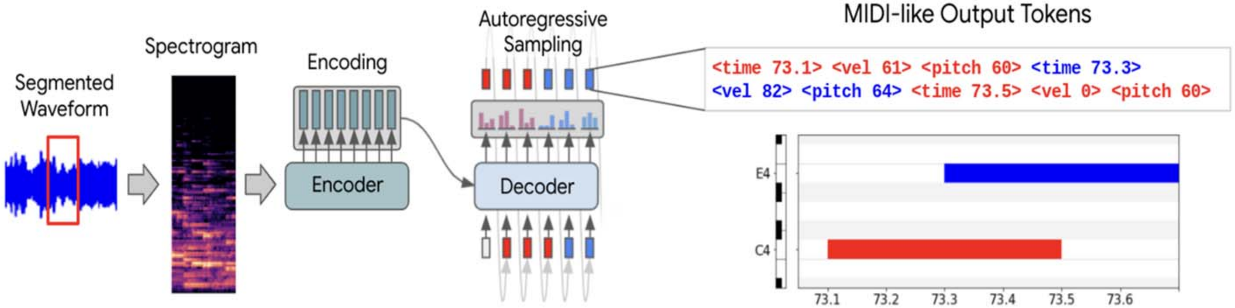


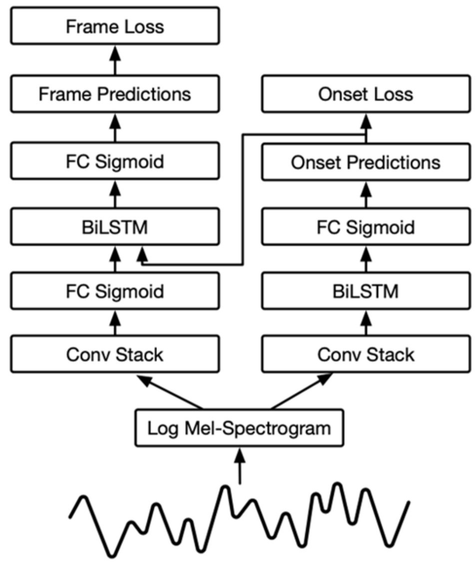
Electric guitar transcription (wet audio -> tab)

Transcription

Transcription - Transcribe Audio into MIDI

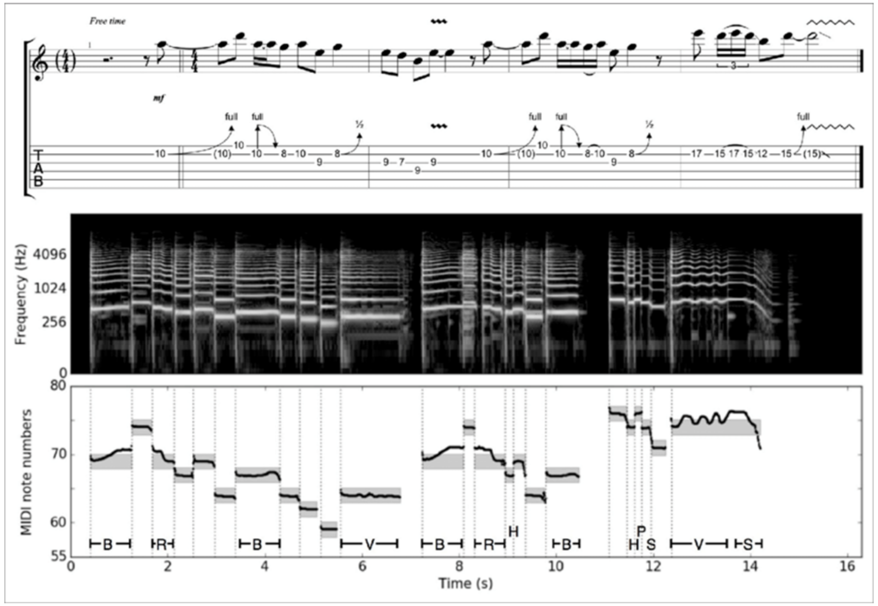
1. Onset and Frames (O&F)
   (by Curtis Hawthorne, ISMIR'17)
1. Sequence-to-Sequence Piano Transcription with Transformers
   (by Curtis Hawthorne, ISMIR'21)
1. MT3: Multi-Task Multitrack Music Transcription
   (by Josh Gardner, ICLR'22)





Challenge in Transcription

1. electric guitar audio often comes with multi-effect (amplifier, modulation..)
2. most of transcription model is designed for piano in MIDI, which is in a discrete format



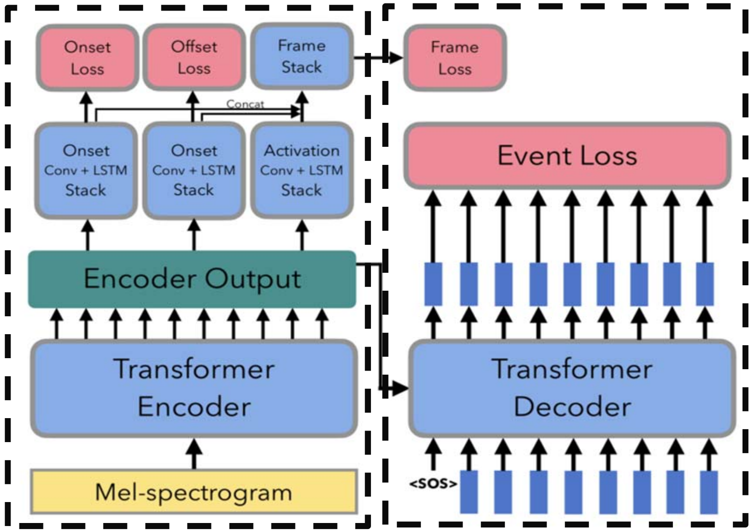
TENT: Technique-Embedded Note Tracking for Real-World Guitar Solo Recordings, Ting-Wei Su, Yuan-Ping Chen, Li Su, Yi-Hsuan Yang, ISMIR’19

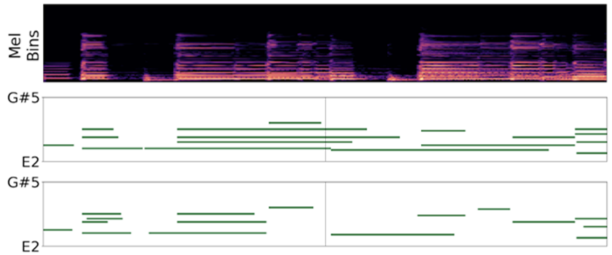
Transcription

Inspired previous mentioned works

We proposed a **Guitar Transcription Model** (ICASSP'22)

transcribing amplifier effected guitar audio

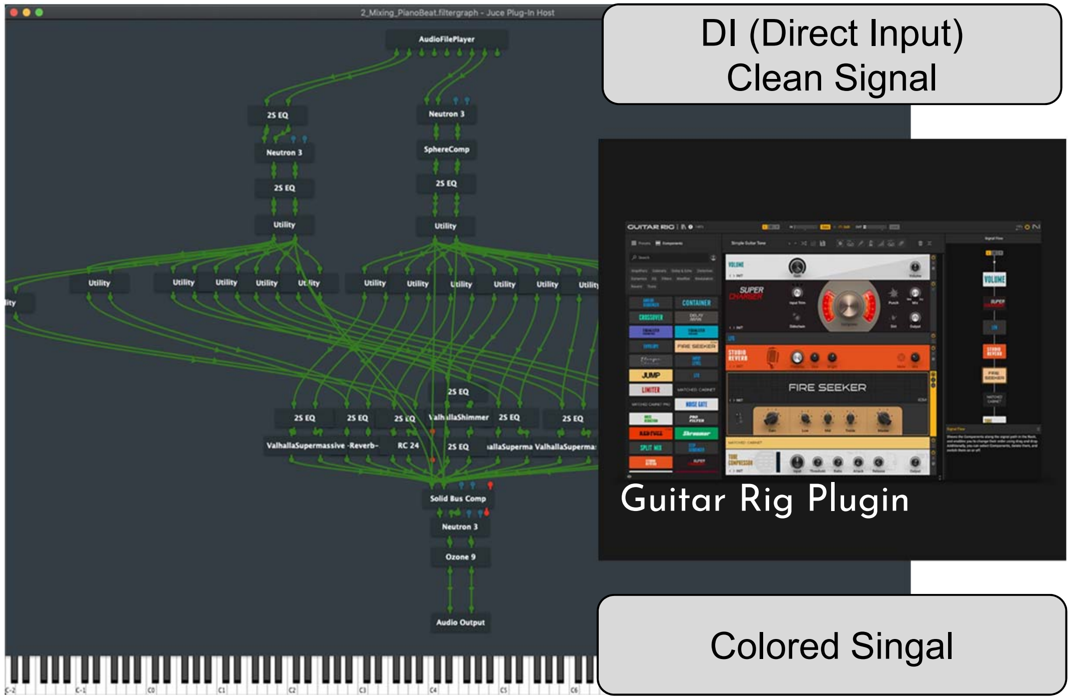




Model	(Encoder output) Onset F1	Frame F1
Onset and Frame (OAF) [14]	0.591	0.583
CE-only Transformer [17]	0.543	0.523
	0.554	0.524
	0.568	0.537
Proposed multi-loss Transformer	0.598	0.579
	0.604	0.573
	0.613	**0.582**

Transcription

Also a new Guitar Dataset - EGDB



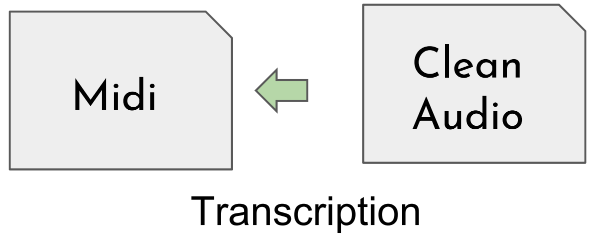
- The DI (input signal) is recorded by musician
  - Given a Tab
  - Sight Reading Performance
    - w/ a special pickup
  - Post-processing
  - Human Curation
  - Rendered by JUCE
    - w/ different tones
- We have (tab, DI, color) pairs
  - Audio of individual string

Followup works

1. Sequence-to-Sequence Network Training Methods for Automatic Guitar Transcription With Tokenized Outputs (ISMIR'23)
2. Note and Playing Technique Transcription of Electric Guitar Solos in Real-World Music Performance (ICASSP'23) *from Li Su's lab*
3. FretNet: Continuous-valued pitch contour streaming for polyphonic guitar tablature transcription (ICASSP'23)
4. SynthTab: Leveraging Synthesized Data for Guitar Tablature Transcription (ICASSP'24)
5. GAPS: A Large and Diverse Classical Guitar Dataset and Benchmark Transcription Model (ISMIR'24)

Large Dataset!

Followup works (Large dataset)



Midi Guitar	# Tracks	Variations	Total Hours
Acoustic Nylon (25)	5501	7	1620.40
Acoustic Steel (26)	5149		1890.95
Electric Jazz (27)	1572	16	1305.73
Electric Clean (28)	2989		2793.04
Electric Muted (29)	504		467.47
Overdriven (30)	1556		1534.21
Distortion (31)	3444		3501.09

Table 2. SynthTab track distribution by MIDI instrument.

Both of them are rendered from DadaGP
Collecting electric guitar dataset is a challenge task (not in algorithmic manner)

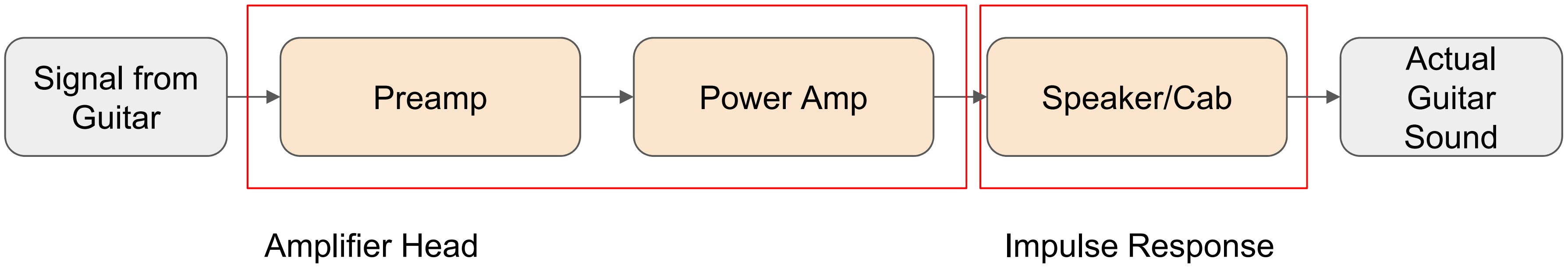
1. Pre-train
2. Finetune on their dataset

Name	Audio type	Track count	Duration (m)	Note count	Scores
GuitarSet [5]	Real	360	180	62,476	No
IDMT-SMT-Guitar [7]	Real	1173	340	*5,767	No
EGDB [6]	Real	240	118	35,700	No
FrançoisLeduc [4]	Real	79	240	75,312	Yes (commercial)
GAPS (ours)	Real	300	843	259,410	Yes
SynthTab [8]	Synthetic	20,715	786,774	-	Yes, via DadaGP

Effect rendering (clean audio -> wet audio)

Neural amplifier (effect) modeling

- How electric guitar produce sound:

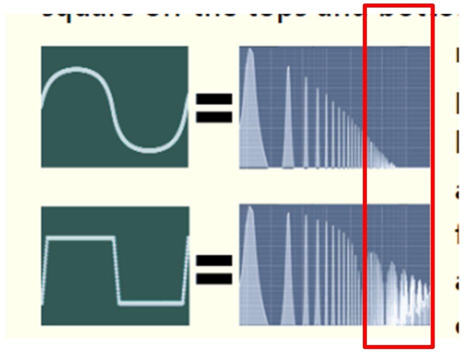


Background of overdrive and distortion

- From guitarist perspective
  - Overdrive: adding a little gain boost or boosting the signal, it also make the dynamic more sensitive
  - Distortion: produces way more gain than overdrive
- From signal perspective
  - They are both the result of a non-linear characteristic curve(clipping), which causes the input signal to be clipped at the output
  - Overdrive or distortion?
    depends on the extent of **non-linearity**

Sin Wave

Clipped
Sin Wave



Traditional Method : Circuit Analysis

- White-Box
- Nodal Analysis
  - Rewrite the schematic into equations
- pros:
  - Accurate
  - User control
- cons:
  - Slow and infeasible for large circuit
  - Re-design everytime
  - Need to open up the hardware

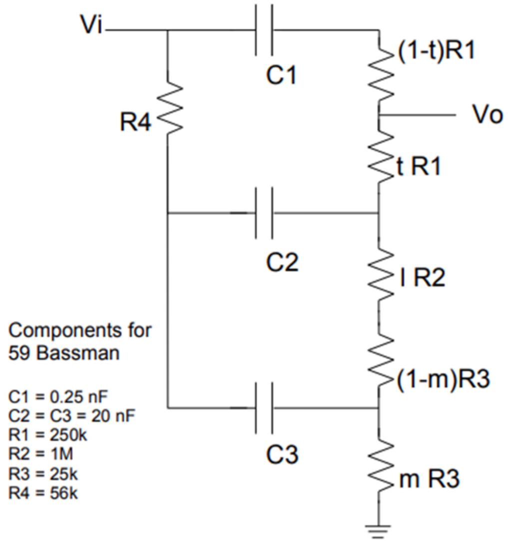
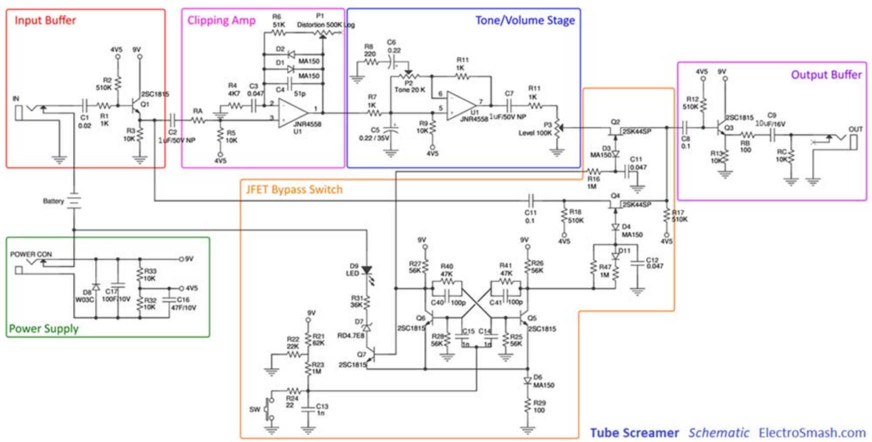


Figure 1: *Tone stack circuit with component values.*

(DAfx'06) DISCRETIZATION OF THE '59 FENDER BASSMAN TONE STACK, David T. Yeh

Traditional Method : Circuit Analysis



TS808

Klon Centaur

Need to know the circuit itself, then model it.

Neural Network on amps or effect modelling

- Notable Researcher
    - Alec wright
        - Ph.D@Aalto University
    - Vesa Välimäki
        - Professor@Aalto University
    - Lauri Juvela
        - Assistant Professor@Aalto University
    - Marco A. Martinez-Ramirez
        - Researcher@Sony AI
    - Christian J. Steinmetz
        - Ph.D@QUML

- Company
    - Neural DSP
    - Native Instrument
    - Adobe
    - Positive Grid
    - many pedal and plugin company…

CNN on amps modelling

(ICASSP'19)Deep Learning for Tube Amplifier Emulation

from Eero-Pekka Damskagg, **Lauri Juvela**, Etienne Thuillier, and **Vesa Välimäki**

Dataset : audio tagging dataset (~4 hours)

Model : Wavenet - with condition module

$$\boldsymbol{z} = \tanh(W_f * \boldsymbol{x} + V_f * \boldsymbol{c}) \odot \sigma(W_g * \boldsymbol{x} + V_g * \boldsymbol{c}), \quad (2)$$

Target Tone : Fender Bassman preamplifier which rendered from SPICE

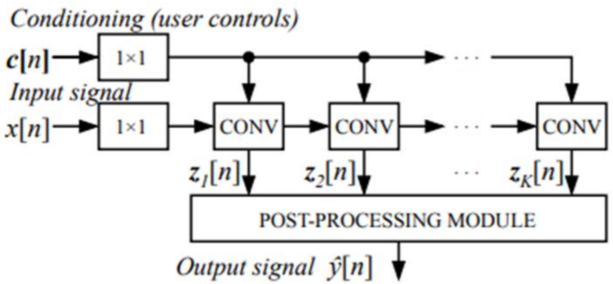


Fig. 2. Proposed deep neural network architecture.

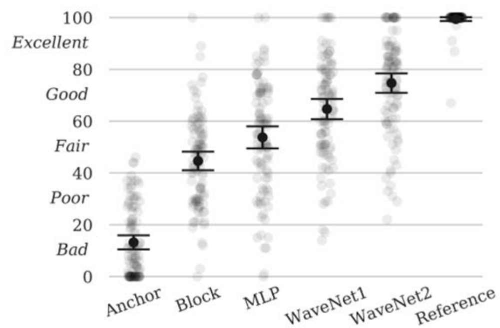


Fig. 3. Mean results of the MUSHRA listening test.

RNN on amps modelling

(DAFx'19)[Real-Time Black-Box Modelling With Recurrent Neural Networks]

from **Alec Wright**, Eero-Pekka Damskägg, and **Vesa Välimäki**

Dataset : IDMT-guitar - 8 minutes and 10 seconds of audio

Model : RNN

Target Tone : Electro-Harmonix Big Muff Pi distortion/fuzz pedal and the Blackstar HT-1 combo guitar amplifier.

Table 1: *Error-to-signal ratio and processing speed for the Wavenet and the proposed GRU/LSTM models of the HT-1 Amplifier. The best results are highlighted.*

Model	Hidden Size	Layers	Number of Parameters	ESR	Time (s) / s of Output
WaveNet	16	10	24065	2.2%	0.53
WaveNet	8	18	11265	1.2%	0.63
WaveNet	16	18	43265	**0.79%**	0.91
GRU	32	1	3393	3.3%	**0.097**
LSTM	64	1	17217	1.8%	0.24
LSTM	96	1	38113	1.1%	0.41

Table 2: *Error-to-signal ratio and processing speed for the Wavenet and proposed LSTM models of the Big Muff pedal. The best results are highlighted.*

Model	Hidden Size	Layers	Number of Parameters	ESR	Time (s) / s of Output
WaveNet	16	10	24065	11%	0.53
WaveNet	8	18	11265	9.9%	0.63
WaveNet	16	18	43265	9.2%	0.91
LSTM	32	1	4513	10%	**0.12**
LSTM	48	1	9841	6.1%	0.18
LSTM	64	1	17217	**4.1%**	0.24

If there's no paired data.
Can we learn the transformation between these two distributions?

Unsupervised learning on amps modelling

(ICASSP'22)ADVERSARIAL GUITAR AMPLIFIER MODELLING WITH UNPAIRED DATA

from **Alec Wright**, **Vesa Välimäki** and **Lauri Juvela**

Dataset : IDMT-guitar - 40 min of audio from the Ibanez 2820 guitar, and 30 min from the Career-SG

Generator : Wavenet

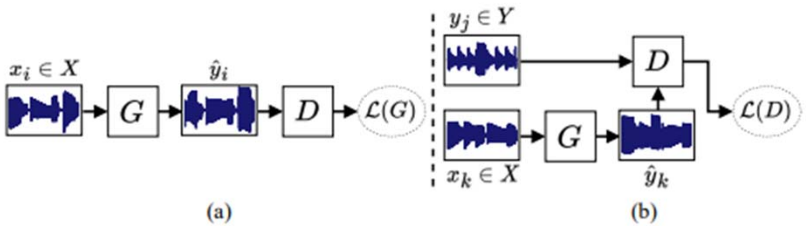


Fig. 2. Training setup for (a) the Generator and (b) Discriminator. Generator inputs, taken from the input domain X, are processed to emulate the timbre, but not content, of the target domain Y.

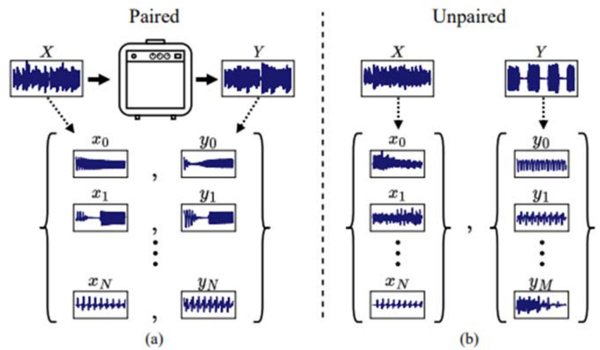


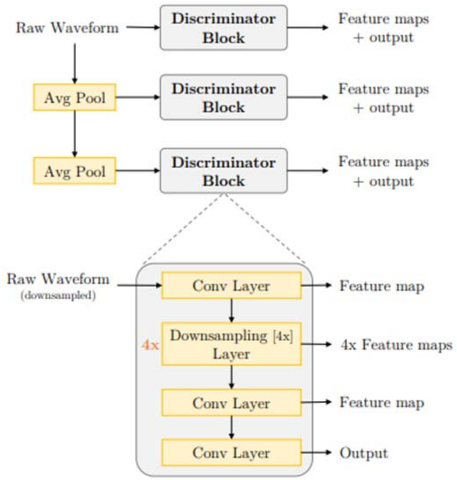
Fig. 1. (a) Supervised black-box modelling is based on *paired* audio data $\{x_i, y_i\}_{i=0}^{N}$, where the target audio y_i is obtained by processing the input audio x_i with the target device. When paired data is unavailable, we propose to use (b) *unpaired* data, made up of examples of a *source timbre* $\{x_i\}_{i=0}^{N} \in X$ and examples of a *target timbre* $\{y_j\}_{j=0}^{M} \in Y$, where neither the content nor the timbre contained in x_i match those contained in y_j.

Unsupervised learning on amps modelling

(ICASSP'22)[ADVERSARIAL GUITAR AMPLIFIER MODELLING WITH UNPAIRED DATA]

from **Alec Wright**, **Vesa Välimäki** and **Lauri Juvela**

Discriminator: [MelGAN]



(b) Discriminator

Table 2. Objective and subjective results for the Single Guitar experiment. For validations losses, bold indicates best performing unsupervised model. For the listening test result bold indicates best performing of all models and 95% confidence intervals are shown.

Model		Validation Loss $\mathcal{E}_{\text{ms}}$	$\mathcal{E}_{\text{lms}}$	$\mathcal{E}_{\text{mel}}$	$\mathcal{E}_{\text{lmel}}$	$\mathcal{E}_{\text{ESR}}$	Listening Test
Target Tone: Clean							
Supervised		5.12	0.76	0.57	0.12	0.003	81±4.1
MelGAN		**37.5**	1.47	**2.75**	**0.17**	2.38	71±4.8
Spectral Domain Input	# Disc.						
Spect.	1	39.2	3.27	3.39	0.39	2.55	32±4.7
Mel	1	40.0	1.51	2.88	0.28	1.27	46±4.4
Log Spect.	3	44.1	**0.81**	3.76	0.18	2.71	82±4.5
Log Mel	3	46.9	0.93	4.07	0.19	**1.04**	**83±3.9**
Target Tone: Light Distortion							
Supervised		2.57	0.81	0.28	0.09	0.001	**93±3.0**
MelGAN		**25.2**	2.18	**1.32**	**0.18**	2.51	73±5.4
Spectral Domain Input	# Disc.						
Spect.	1	32.5	4.26	2.39	0.45	**1.49**	35±4.0
Mel	1	34.4	4.12	2.57	0.48	2.43	34±4.0
Log Spect.	1	45.3	**1.11**	4.51	0.23	2.18	81±4.8
Log Mel	3	38.1	1.17	3.36	0.21	2.50	88.7±3.9
Target Tone: Heavy Distortion							
Supervised		6.33	2.53	0.60	0.19	0.03	57±4.6
MelGAN		**22.4**	**2.49**	**1.81**	**0.22**	2.04	**92±2.8**
Spectral Domain Input	# Disc.						
Spect.	1	28.9	4.14	2.70	0.37	2.33	54±5.7
Mel	1	25.5	7.15	2.36	0.60	**0.86**	28±3.4
Log Spect.	1	32.1	2.52	3.25	0.29	3.17	81±4.8
Log Mel	3	24.5	2.55	2.21	0.23	2.37	85±3.8

If we consider this task as a audio generation problem with clean guitar as conditional input?

Unsupervised learning on amps modelling

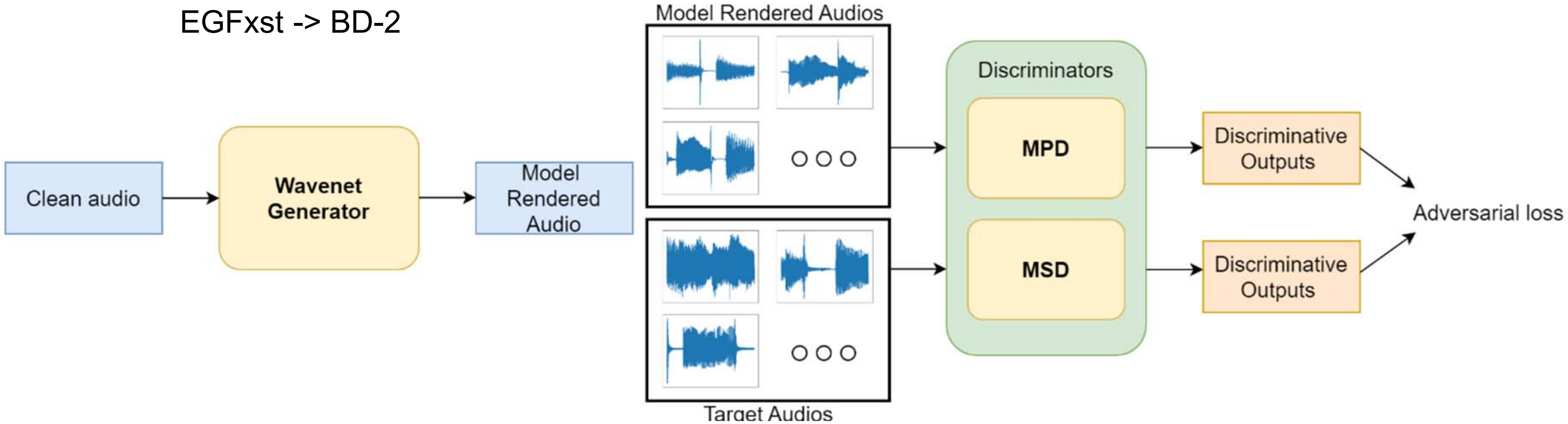
(DAFx'24) Improving unsupervised clean-to-rendered guitar tone transformation using GANs and integrated unaligned clean data

(ISMIR-LBD'23) Neural amplifier modelling with several GAN variants

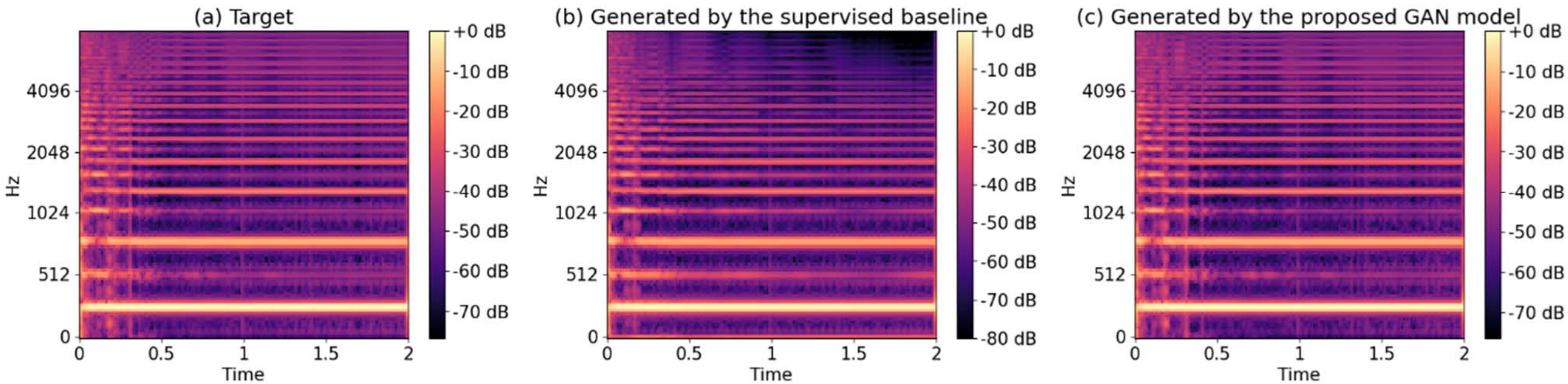
Dataset : EGDB, EGFxset

Generator : Wavenet

Target Tone : EGDB -> Mesa Boogie(Mesa), Marshall, Fender Twin Reverb (Ftwin),
EGFxst -> BD-2



BD-2 as target tone



Way to represent tone is undefined

Way to represent tone is undefined

By text description? Clearly, insufficient for a diverse range of tones

"Muddy" = a tone with too much bass.

"Bright" = a tone with a lot of treble, sometimes *too much*

"Thick" = a tone with a well-crafted midrange and bass

"Thin" = a tone with not enough definition to it.

"Fuzzy" = a tone with a lot of distortion, potentially affecting the clarity of it.

Zero-shot amp modeling

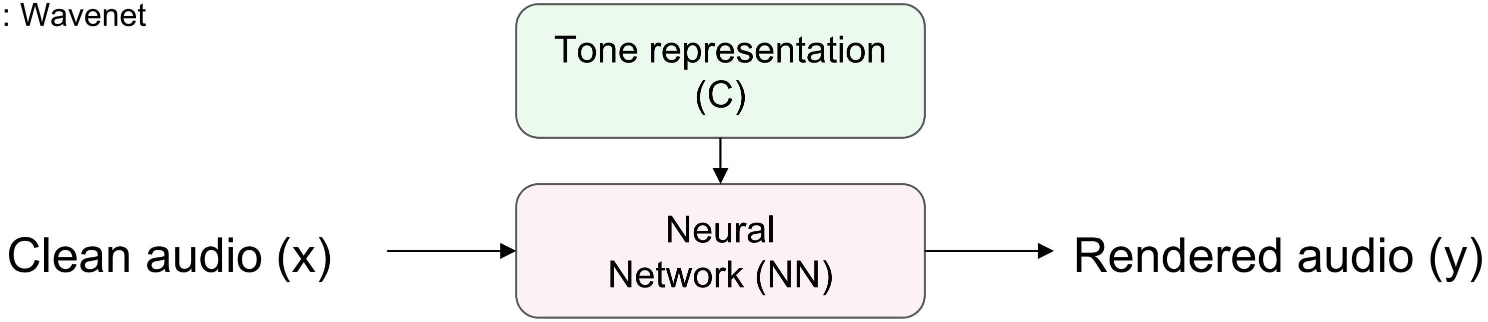
(ISMIR'24) Towards Zero-Shot Amplifier Modeling: One-to-Many Amplifier Modeling via Tone Embedding Control

Positive Grid®

from **Yu-Hua Chen, Yen-Tung Yeh, Yuan-Chiao Cheng, Jui-Te Wu, Yu-Hsiang Ho, Jyh-Shing Roger Jang, Yi-Hsuan Yang**

Dataset : Internal dataset

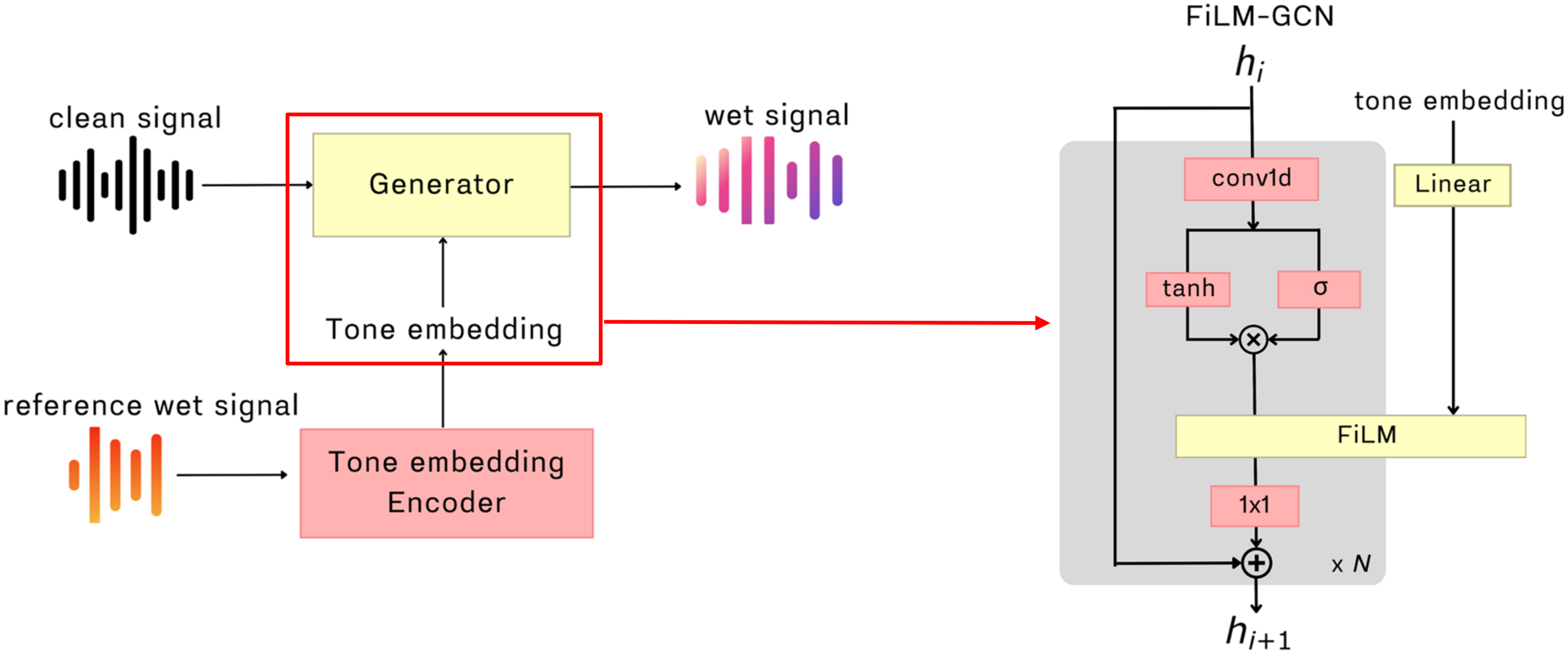
Generator : Wavenet



$$y = NN(x, c)$$

Imitate the transform in device by neural network

General concept of our proposed model



Dataset

- 30 mins of clean audio (x)
- 9 guitar amplifiers rendered audio (y)
  - High-gain
    - Boogie Mark IV (amp1), PRS Archon 100 (amp2), and Soldano SLO-100 (amp3)
  - Low-gain
    - Fender Tweed Deluxe (amp4), Vox AC30 (amp5), and Matchless DC30 (amp6)
  - Crunch
    - Vox AC30 Hand wired Overdriven (amp7), Friedman BE100 (amp8), and Overdriven Marshall JTM45 (amp9).

Tone embedding visualization

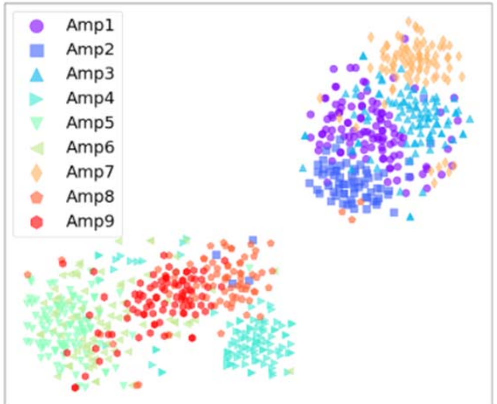
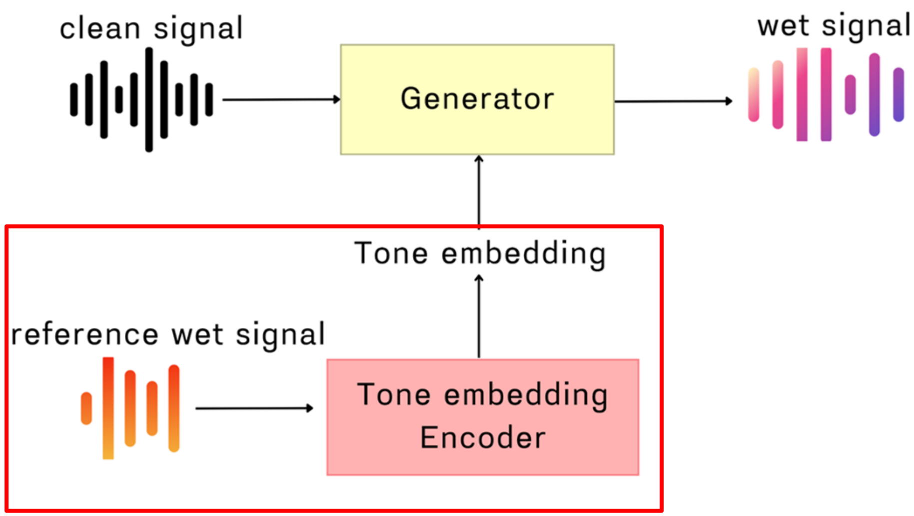


Figure 4: A t-SNE visualization of the tone embeddings from the wet signals of the $N = 9$ amps. Each point represents a tone embedding extracted from a wet signal, with color and shape indicating the category of the amp tone. We see 2 big cross-amp clusters and 9 small clusters for each amp, suggesting the ability of the encoder $\mathcal{E}$ to distinguish between different tones based on their embeddings.

- Tone embedding encoder is trained on a larger dataset and rendered with a great diversity of amplifiers
- The t-SNE visualization shows the efficacy of capturing tone information and clustering tone from same amplifiers.

Rerference wet signal of tone embedding module



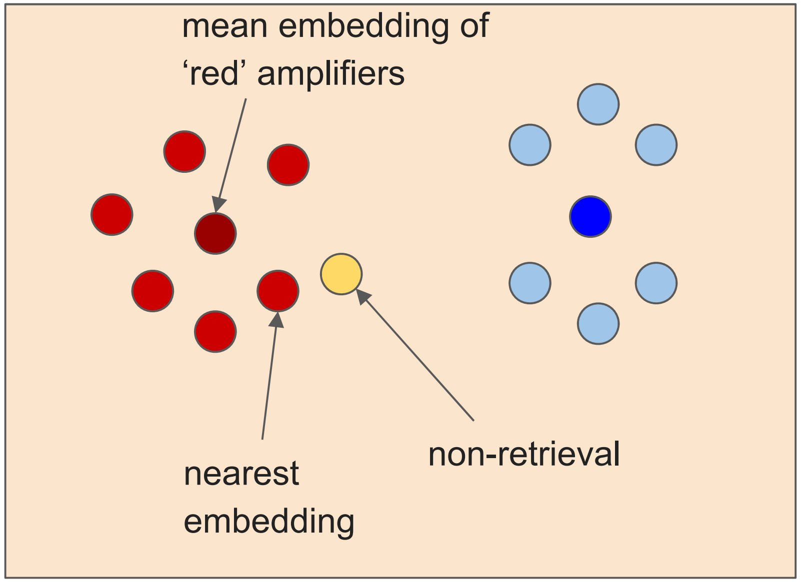
- Paired reference
    - content and tone are identical to wet signal
- Unpaired reference
    - different content
    - identical tone

Objective evaluation in STFT loss

		GCN	FiLM-GCN			Concat-GCN	
		one-to-one	LUT	ToneEmb (paired)	ToneEmb (unpaired)	LUT	ToneEmb (paired)
high-gain	Amp1	0.0420	0.1441	0.1189	**0.0777**	0.1593	0.1523
	Amp2	0.0268	0.1951	**0.0670**	0.1189	0.1741	0.1208
	Amp3	0.0527	0.1659	0.1254	**0.1143**	0.1777	0.1304
low-gain	Amp4	0.0087	0.0698	**0.0230**	0.0275	0.0618	0.0775
	Amp5	0.0004	0.0813	0.0167	**0.0138**	0.0334	0.0166
	Amp6	0.0014	0.0947	0.0169	**0.0121**	0.0779	0.0275
crunch	Amp7	0.0393	0.1022	0.0860	0.0885	**0.0733**	0.0988
	Amp8	0.0124	0.1583	0.0760	**0.0604**	0.1562	0.0775
	Amp9	0.0035	0.1593	0.0375	**0.0290**	0.1211	0.0407

Tone embedding of unseen amplifiers

- **non-retrieval**
  - directly take the output of tone embedding encoder
- **nearest embedding**
  - closet embedding seen in training data
- **mean embedding**
  - closet embedding among mean embedding of 9 amplifiers



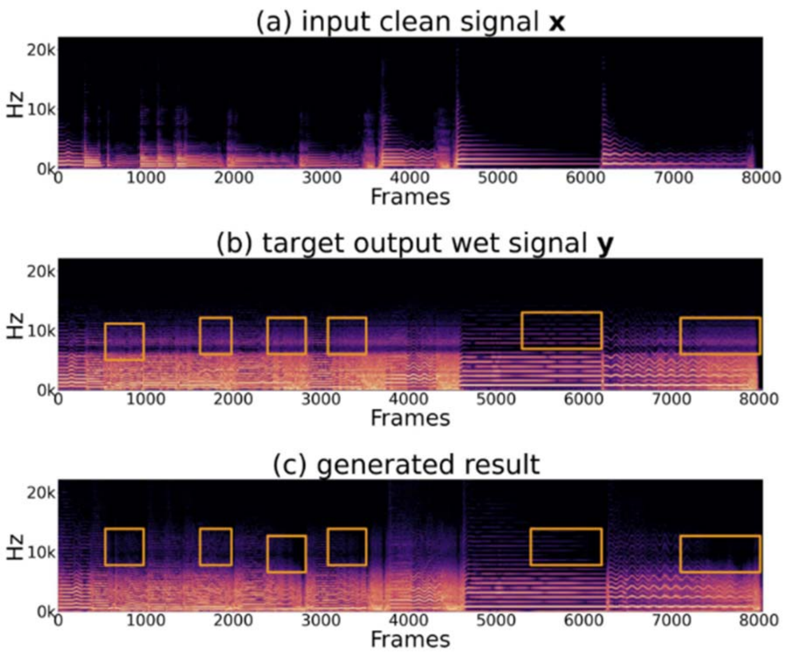
Evaluation of zero shot tone modelling

	non-retrieval	retrieval-based	
	$(\phi^* = \mathcal{E}(\mathbf{z}^*))$	nearest	mean
unseen high gain	**0.2511**	0.2560	0.2593
unseen low gain	0.0338	**0.0274**	0.0404

Table 2: Efficacy of using different methods for FiLM-GCN (cf. Section 3.4) for zero-shot modeling of two unseen amps, measured again in complex STFT loss.

- non-retrieval embedding slightly outperform other methods on modelling high-gain target
- A more versatile embedding condition can enhance the performance of unseen amplifiers

Case study in zero shot modelling scenario



- The high-frequency content is not modelled accurately in the orange square area.
- For the quick string-bending content around frames 6,000 to 7,000, the generated harmonics are correctly damped.
- The characteristic of high gain is accurately modeled.

Conclusion

1. There is opportunity for further improvement and investigation in each transformation or modeling on the below roadmap
2. Guitar-related tasks can use piano-related tasks as a reference, but modifications should be made or highlighted (e.g., tablature vs. MIDI)