

Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck et al. 2017

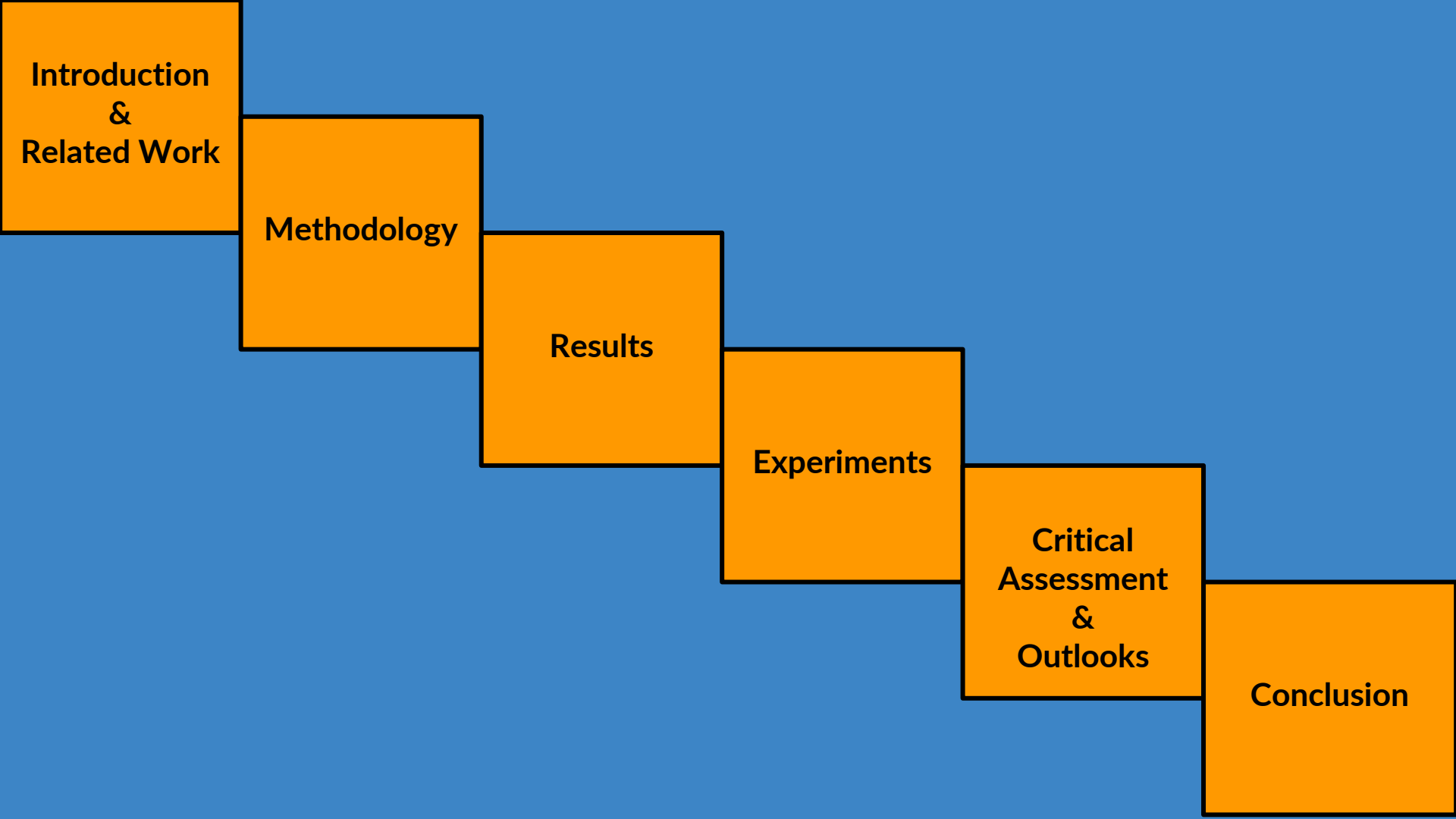
Student - Matteo MARENGO ¹

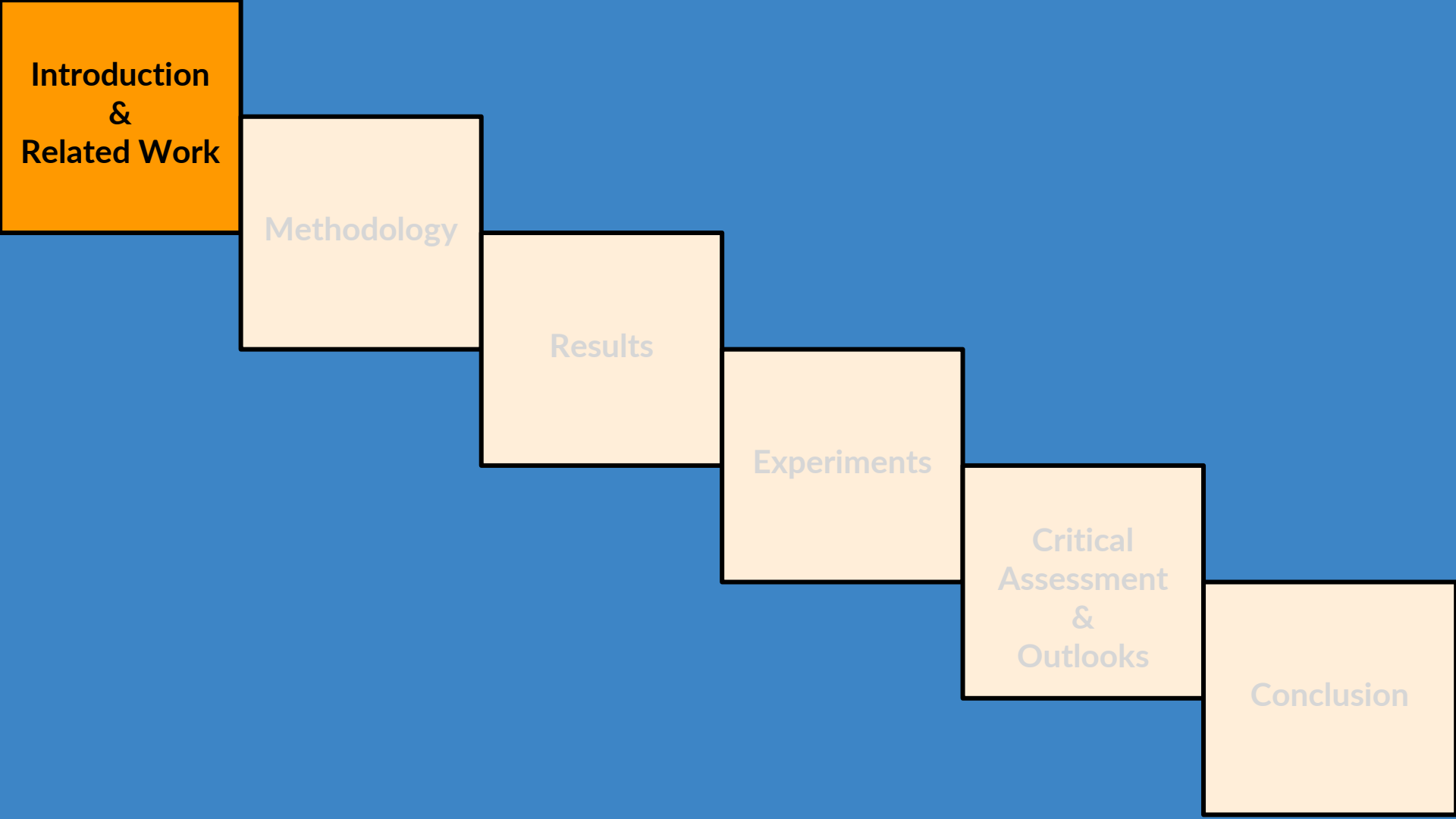
Teacher - Emmanuel BACRY ²

27/03/2024

¹ ENS Paris-Saclay, Msc MVA (*Mathématiques Vision Apprentissage*)

² CEREMADE, Université Paris-Dauphine, PSL





Audio Synthesis: Recent progresses



Angèle - Saiyan [IA] (prod. Lnkhey)



Lnk 🎵
17,9 k a

Rejoindre

S'abonner

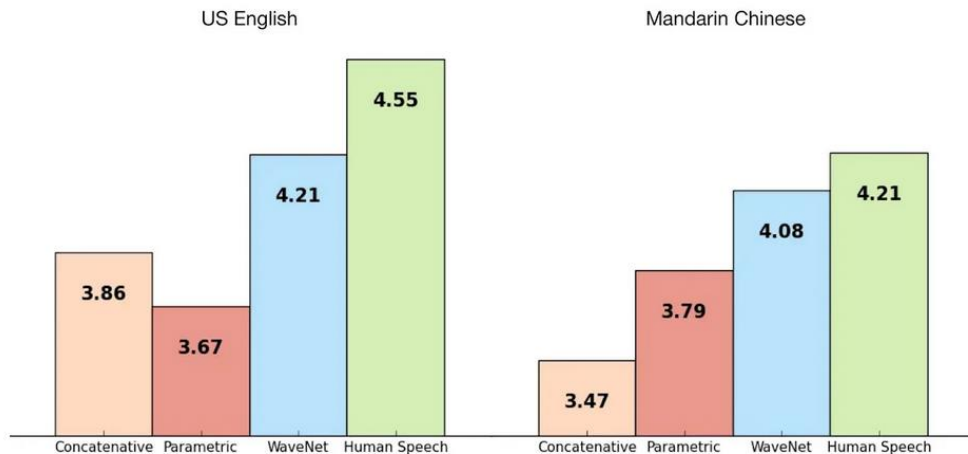
10 M de vues il y a 7 mois

Générique Pokémon - Johnny Hallyday (AI Cover)

1,1 M de vues • il y a 8 mois

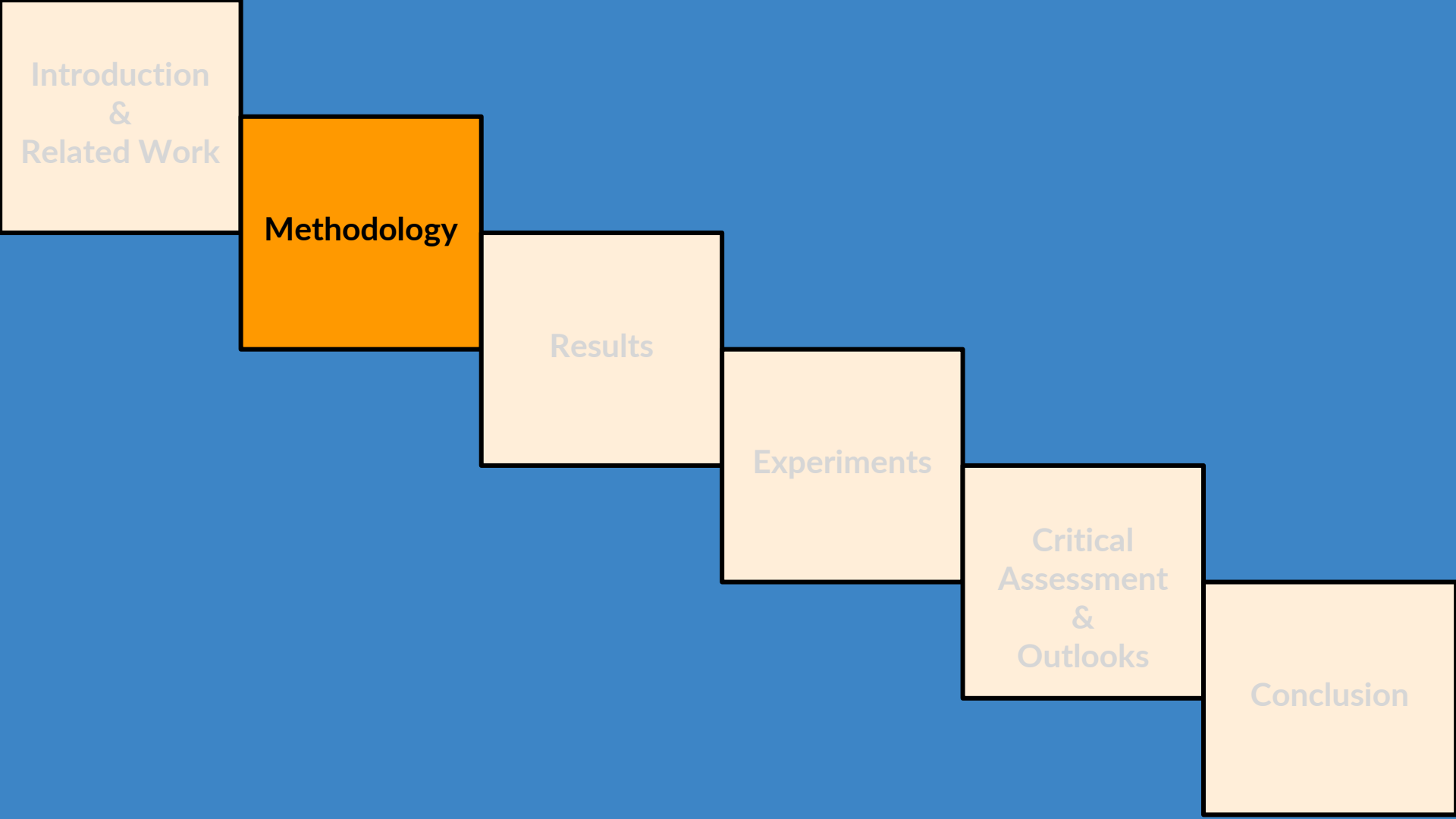
Audio Synthesis: Related Work

- Vocoders for TTS systems
- Synthesizers for Music
- Frequency Modulation (FM)
- WaveNet 2016



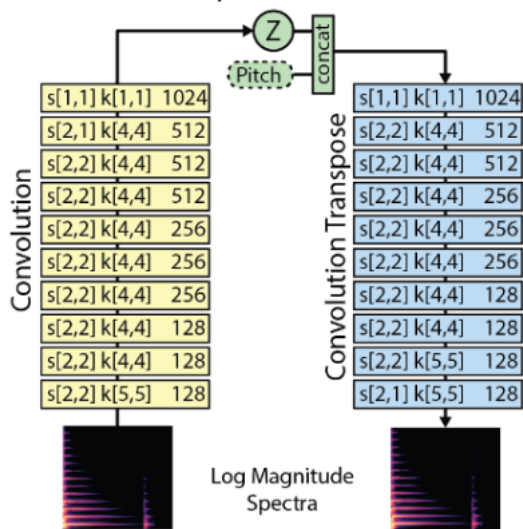
Contributions of the paper

- Limitations: Need of a temporal dependency
- No Dataset for audio
- Two main contributions of the paper:
 - 1/ Novel WaveNet autoencoder architecture
 - 2/ NSYNTH dataset



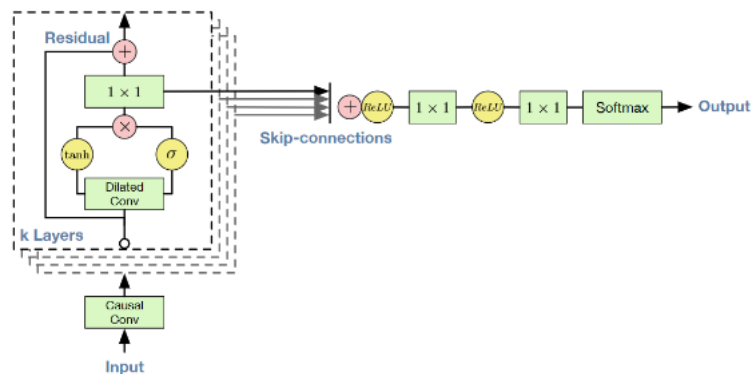
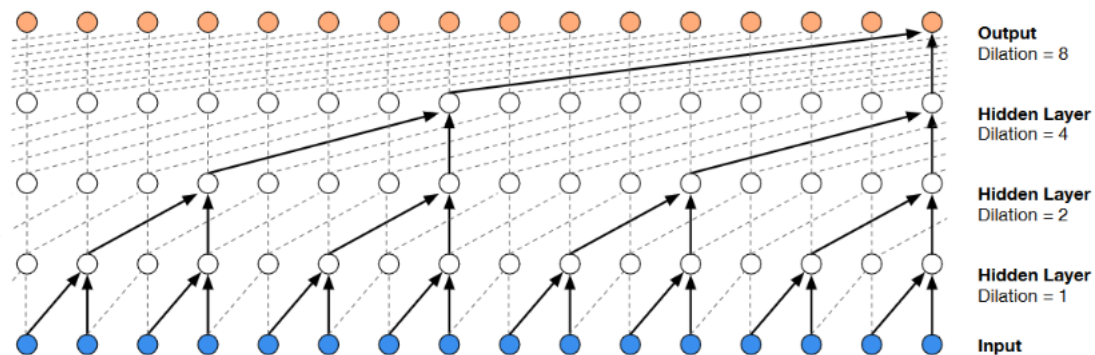
Spectral Autoencoder

a) Baseline Spectral Autoencoder

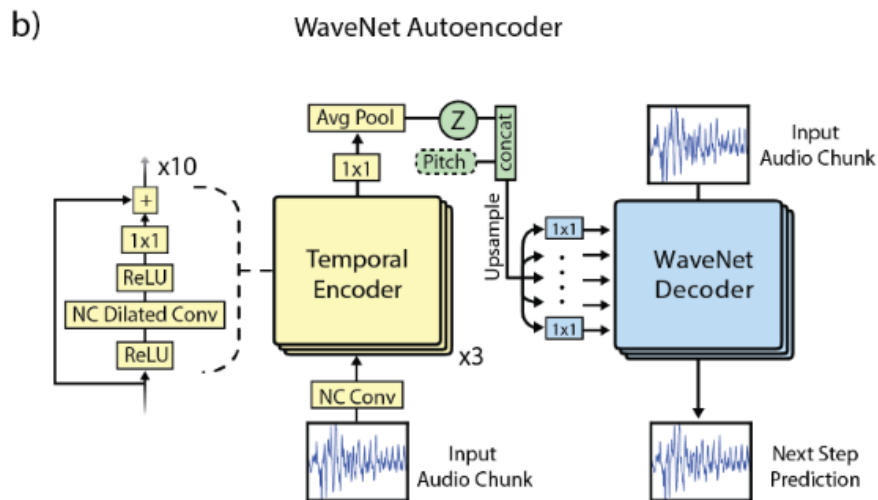


- Baseline autoencoder composed of convolutional structures
- Inspired by models in Computer Vision

WaveNet architecture



AutoEncoder WaveNet architecture

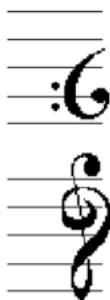


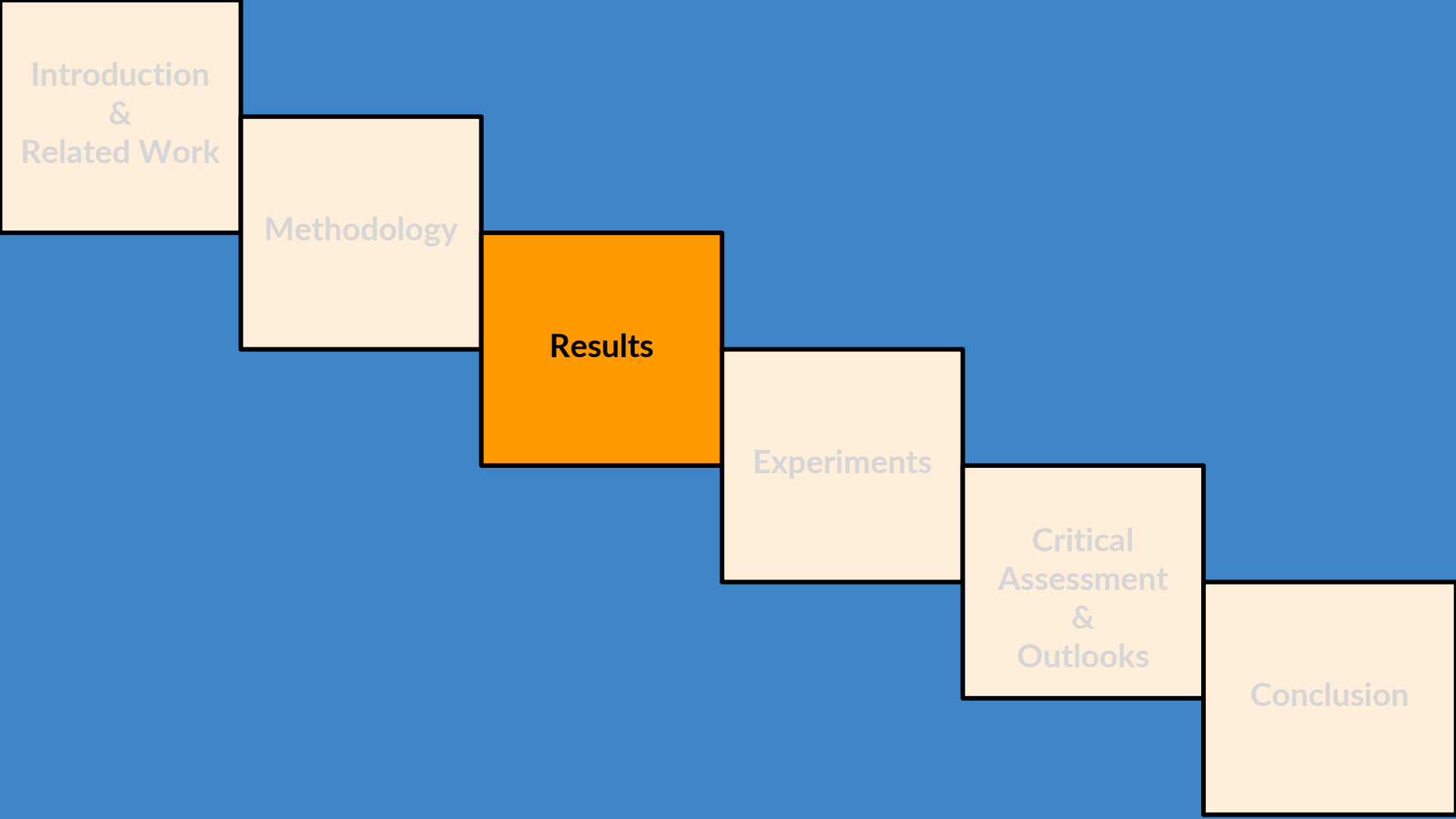
- Temporal encoder that captures hidden embeddings distributed in time.
- 30-layer residual network of dilated convolutions → generates a sequence of hidden codes that represent the audio temporally.

CQT spectrogram

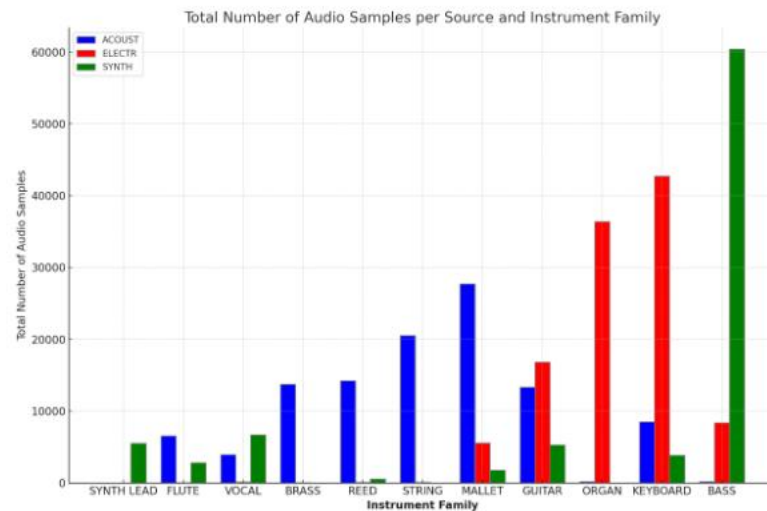
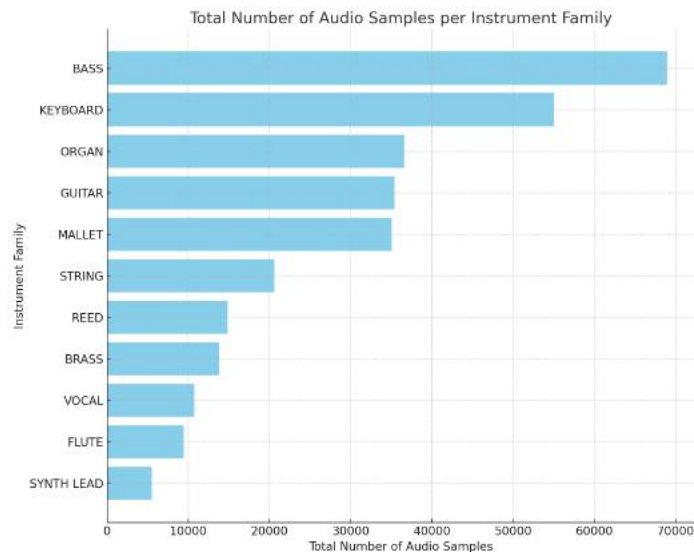
- Constant-q transform.
- CQT spectrogram uses filters spaced logarithmically in frequency
- Greater frequency resolution at lower frequencies and better temporal resolution at higher frequencies.

MIDI number	Note name	Keyboard	Frequency Hz	Period ms
21	A0		27.500	36.36
22	B0		30.868	32.40
23	C1		32.703	30.58
24	D1		36.708	27.24
25	E1		41.203	24.27
26	F1		43.654	22.91
27	G1		48.999	20.41
28	A1		55.000	18.18
29	B1		61.735	16.20
30	C2		65.406	15.29
31	D2		73.416	13.62
32	E2		82.407	12.13
33	F2		87.307	11.45
34	G2		97.999	10.20
35	A2		110.00	9.091
36	B2		123.47	8.099
37	C3		130.81	7.645
38	D3		146.83	6.811
39	E3		164.81	6.068
40	F3		174.61	5.727
41	G3		196.00	5.102
42	A3		220.00	4.545
43	B3		246.94	4.050
44	C4		261.63	3.822
45	D4		293.67	3.405
46	E4		329.63	3.034
47	F4		349.23	2.865
48	G4		392.00	2.551
49	A4		440.00	2.273
50	B4		493.88	2.025
51	C5		523.25	1.910
52	D5		587.33	1.703
53	E5		659.26	1.517
54	F5		698.46	1.432
55	G5		783.99	1.276
56	A5		880.00	1.136
57	B5		987.77	1.012
58	C6		1046.5	0.9556
59	D6		1174.7	0.8513
60	E6		1318.5	0.7584
61	F6		1396.9	0.7159
62	G6		1568.0	0.6378
63	A6		1760.0	0.5682
64	B6		1975.5	0.5062
65	C7		2093.0	0.4778
66	D7		2349.3	0.4257
67	E7		2637.0	0.3792
68	F7		2793.0	0.3580
69	G7		3136.0	0.3189
70	A7		3520.0	0.2841
71	B7		3951.1	0.2531
72	C8		4186.0	0.2389



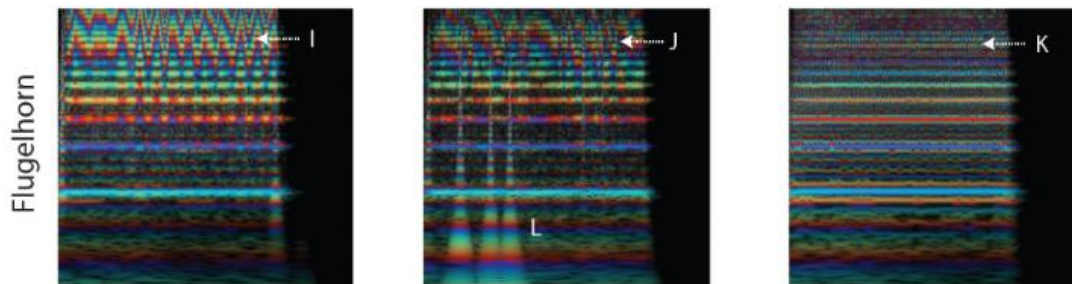


Nsynth dataset



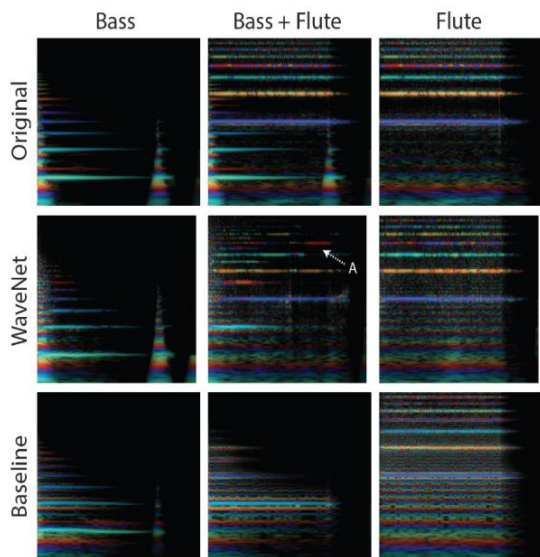
Reconstruction

- **WaveNet autoencoder:** captures key characteristics (fundamental frequency, noise on the attack)
- **Baseline Model:** adds percussive sounds, suffers from noisy phase estimation



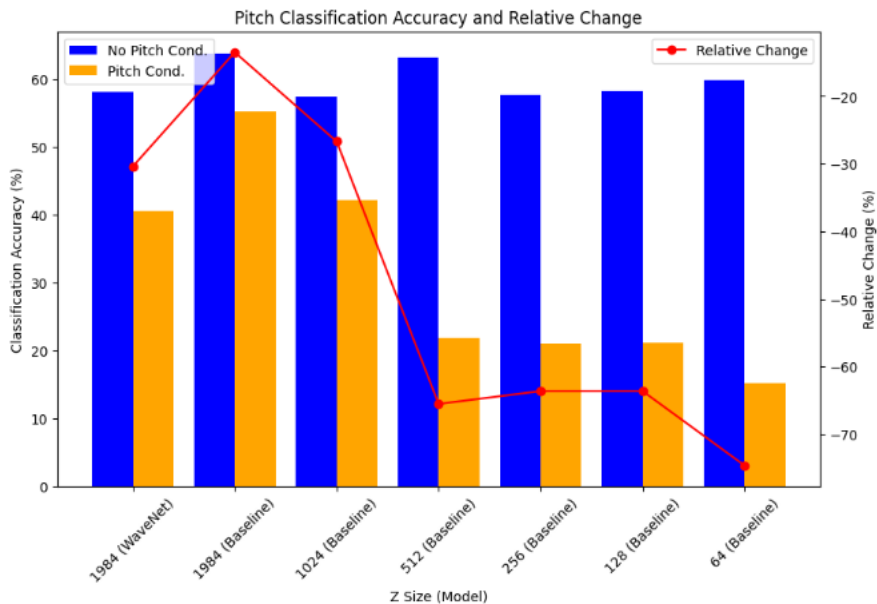
Interpolation in Timbre and Dynamics

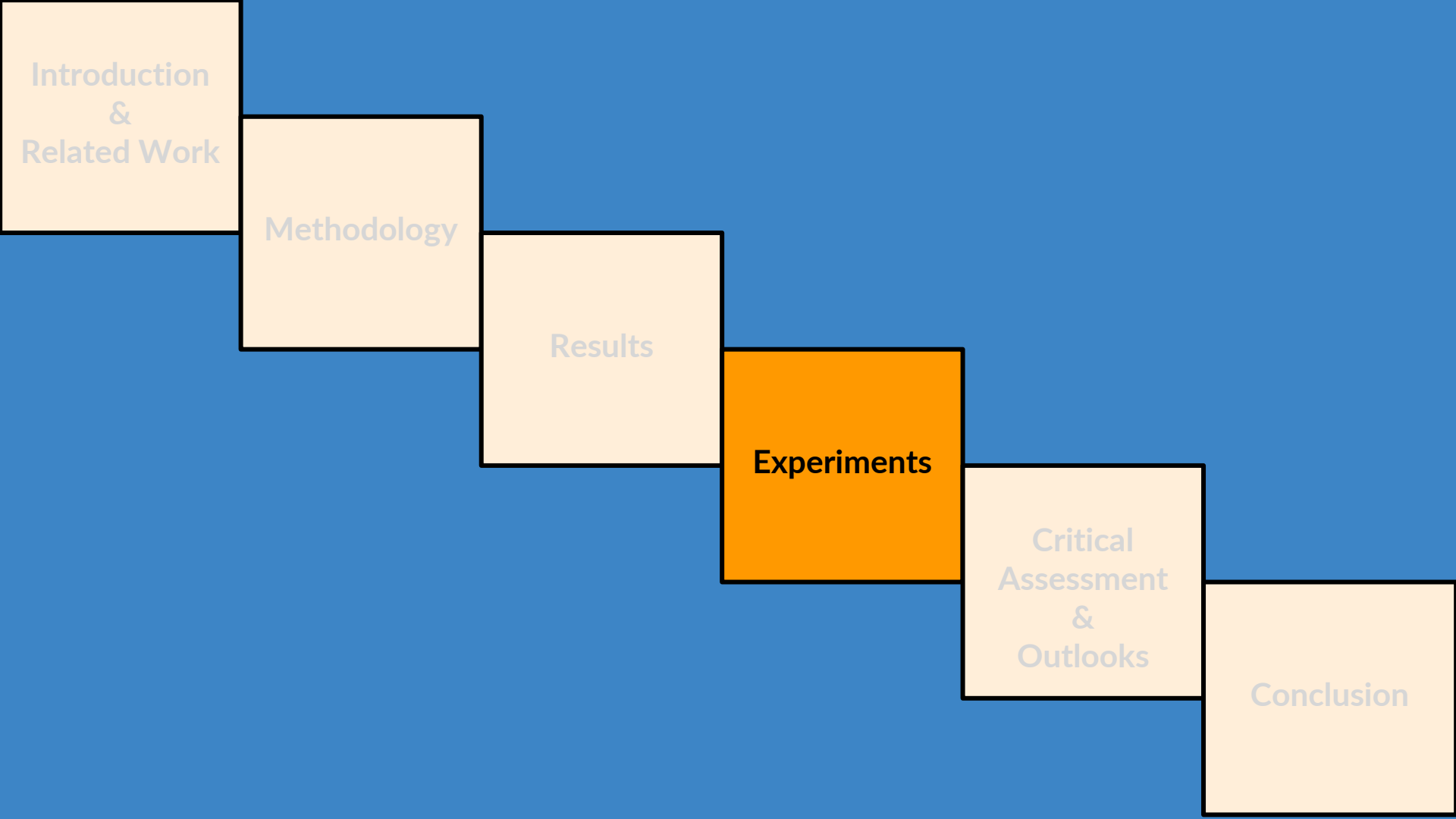
- **WaveNet autoencoder:** it exhibits more realistic and perceptually interesting blends
- **Baseline Model:** adds phase distortion



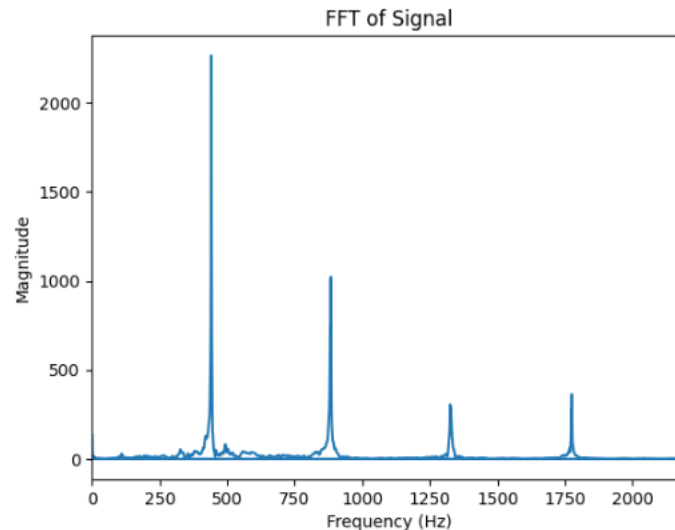
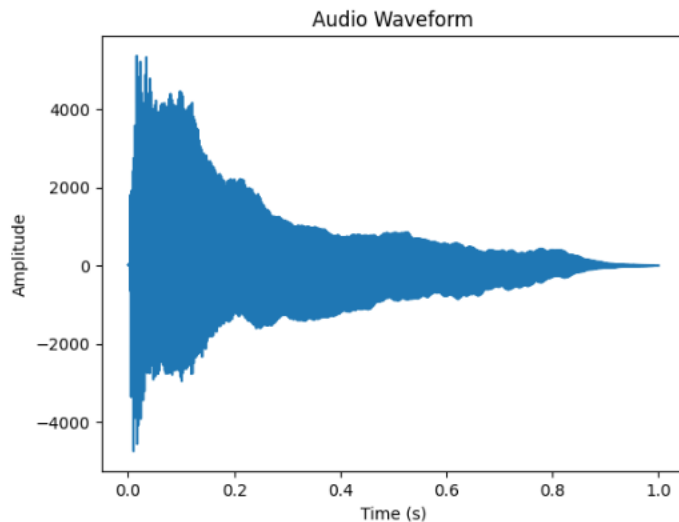
Entanglement of Pitch and Timbre

- Decrease in pitch classification accuracy with conditioning.
- The effect is more pronounced in models with smaller embedding sizes.

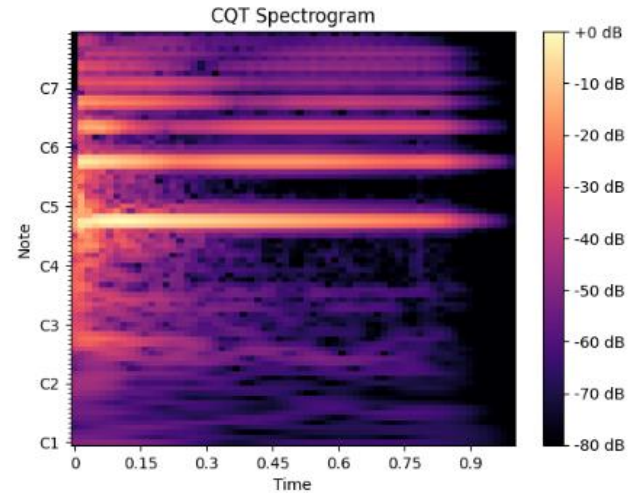
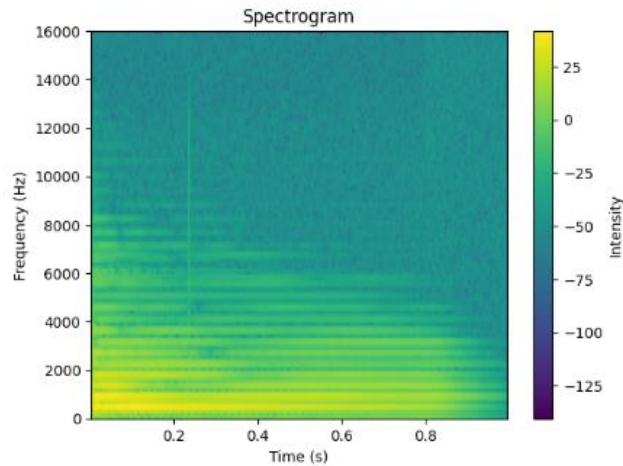




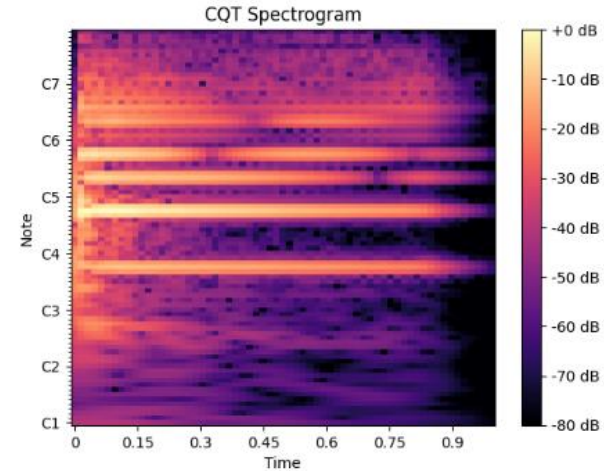
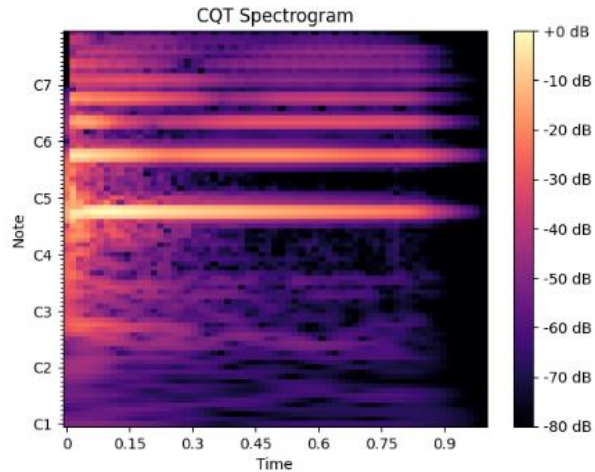
Study a simple note – A4



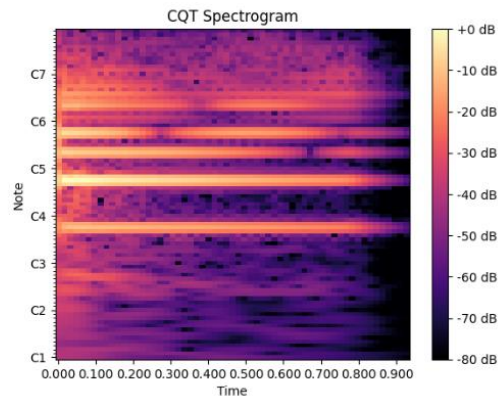
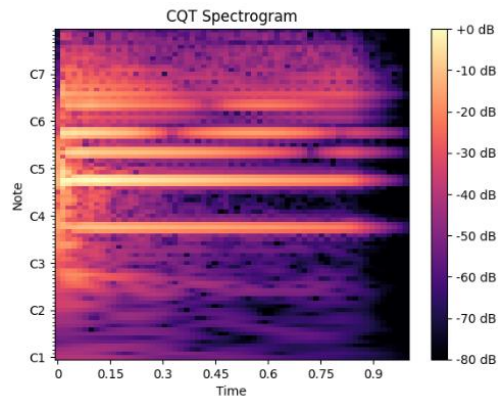
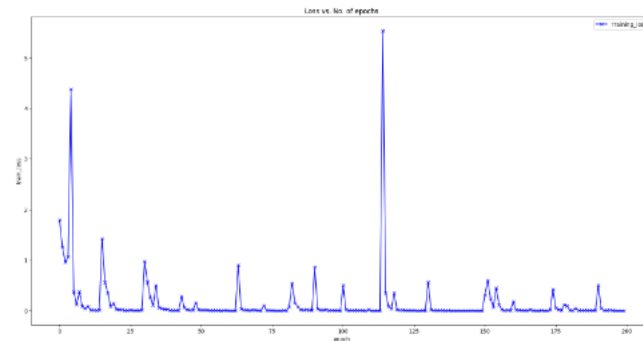
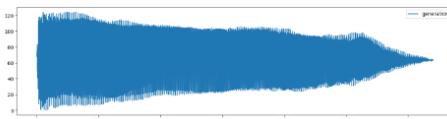
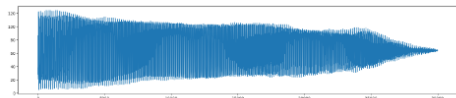
Comparison of the spectrograms

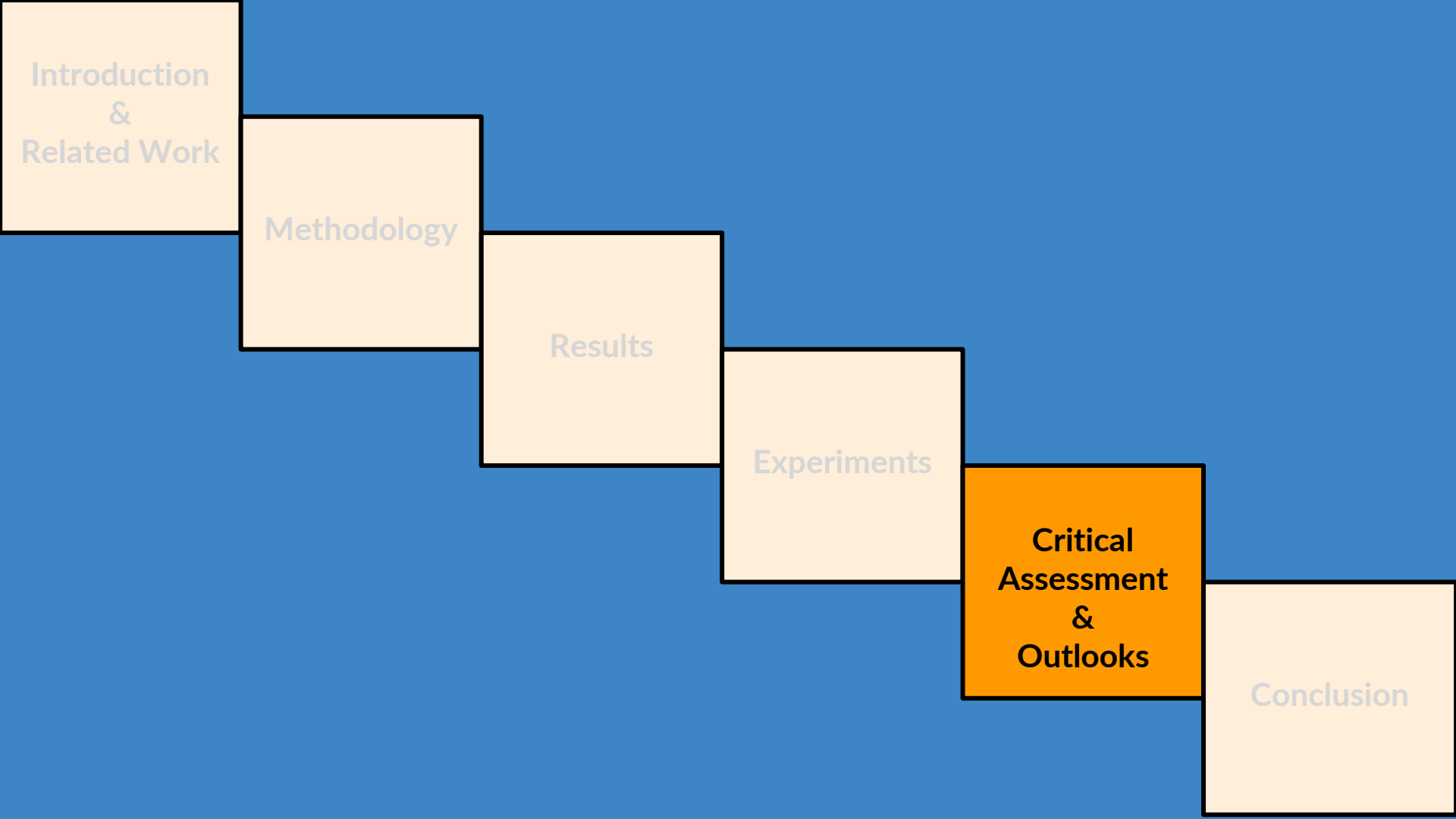


Comparison of A4 spectrogram with A3A4 spectrogram



WaveNet architecture





Introduction
&
Related Work

Methodology

Results

Experiments

**Critical
Assessment
&
Outlooks**

Conclusion

Strengths & Weaknesses of the paper

- Methodology is **thorough** and **well structured**.
- It enables reproducibility and further research (e.g they give the hyperparameters)
- **Qualitative** and **quantitative** measures.
- They conducted diverse and comprehensive experiments.

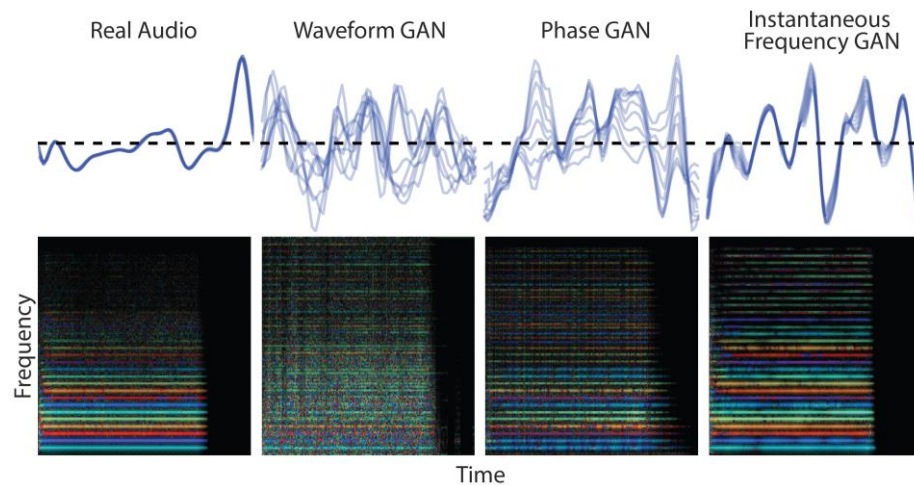
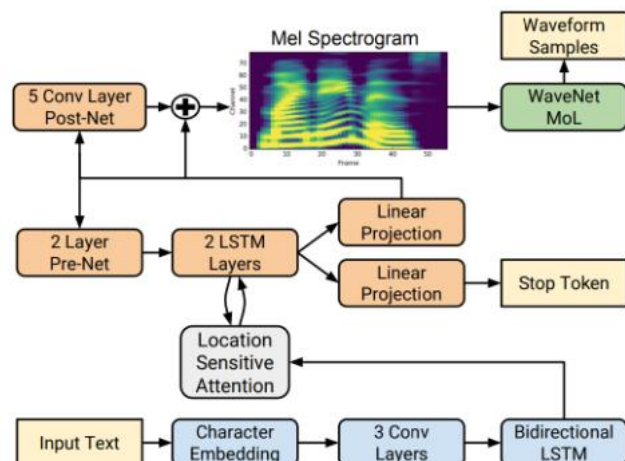
- They only compare the WaveNet autoencoder with a baseline autoencoder.
- The Nsynth dataset is not **balanced** within classes → Is it an issue?
- No study on **generalization** or overfitting.
- No training on other datasets.
- No discussion on the **computational part**.

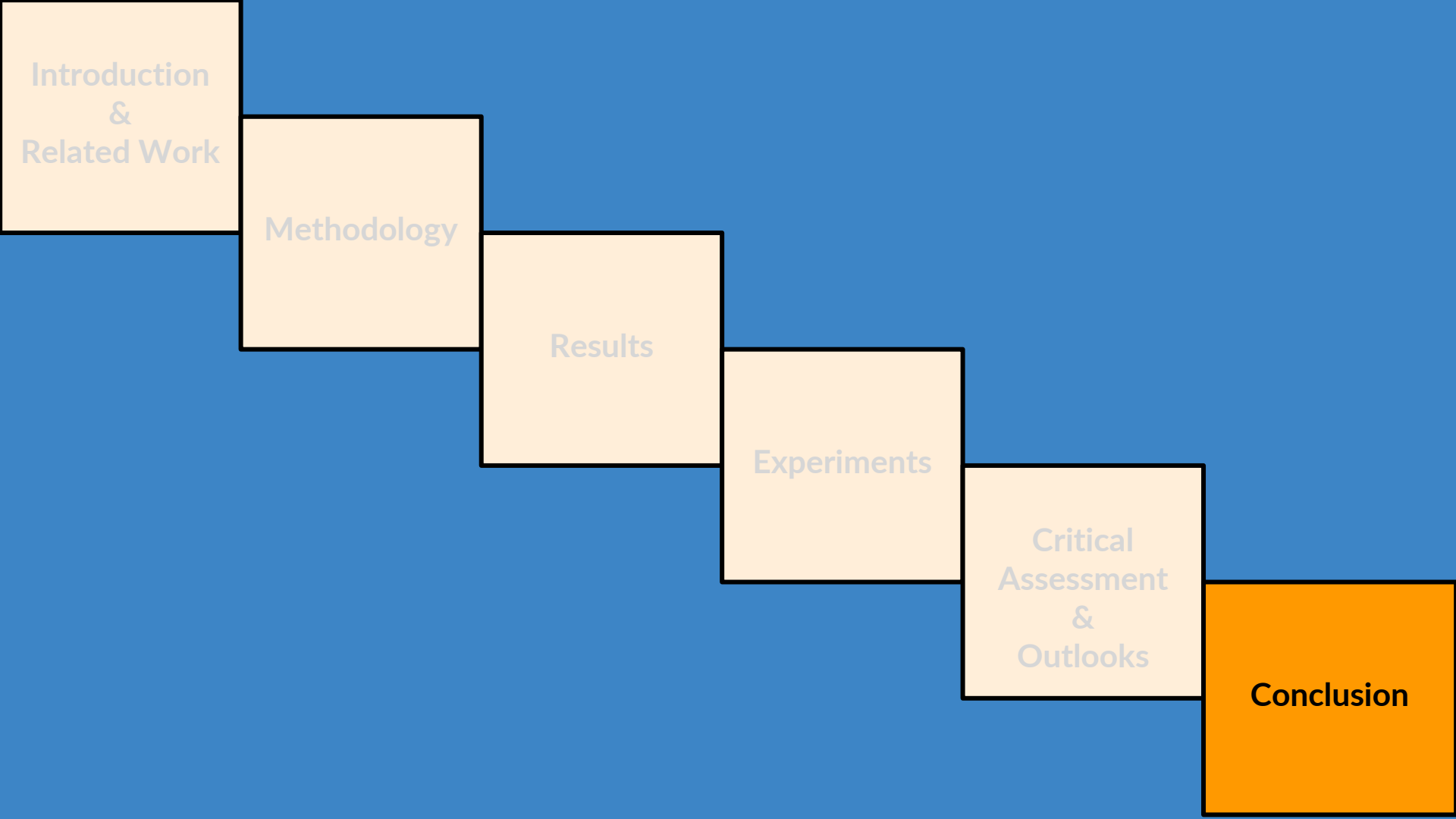


- Explore more memory-efficient neural network architectures (LSTM, Transformer)
- Improve the model's ability to separate pitch and timbre (e.g alternative conditioning strategies)
- More extensive validation on external datasets, add regularization techniques.

Outlooks

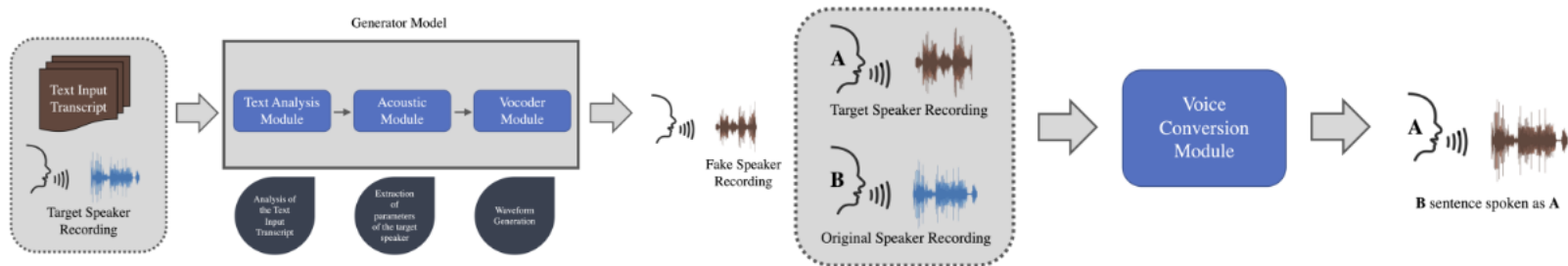
- JukeBox & MuseNet models by OpenAI demonstrates the capability of AI to compose music in various styles and genres.
- Google TacoTron or Facebook MelGAN





Conclusion

- The paper creates a large and diverse audio dataset.
- The field has used it for benchmarking and training (e.g **GanSynth**)
- The WaveNet autoencoder can do audio synthesis without external conditioning.
- But we have to be careful!
- One drawback can be the surge of **deep fakes**
- Cautious progress to avoid the ethical pitfalls.





THANK YOU!