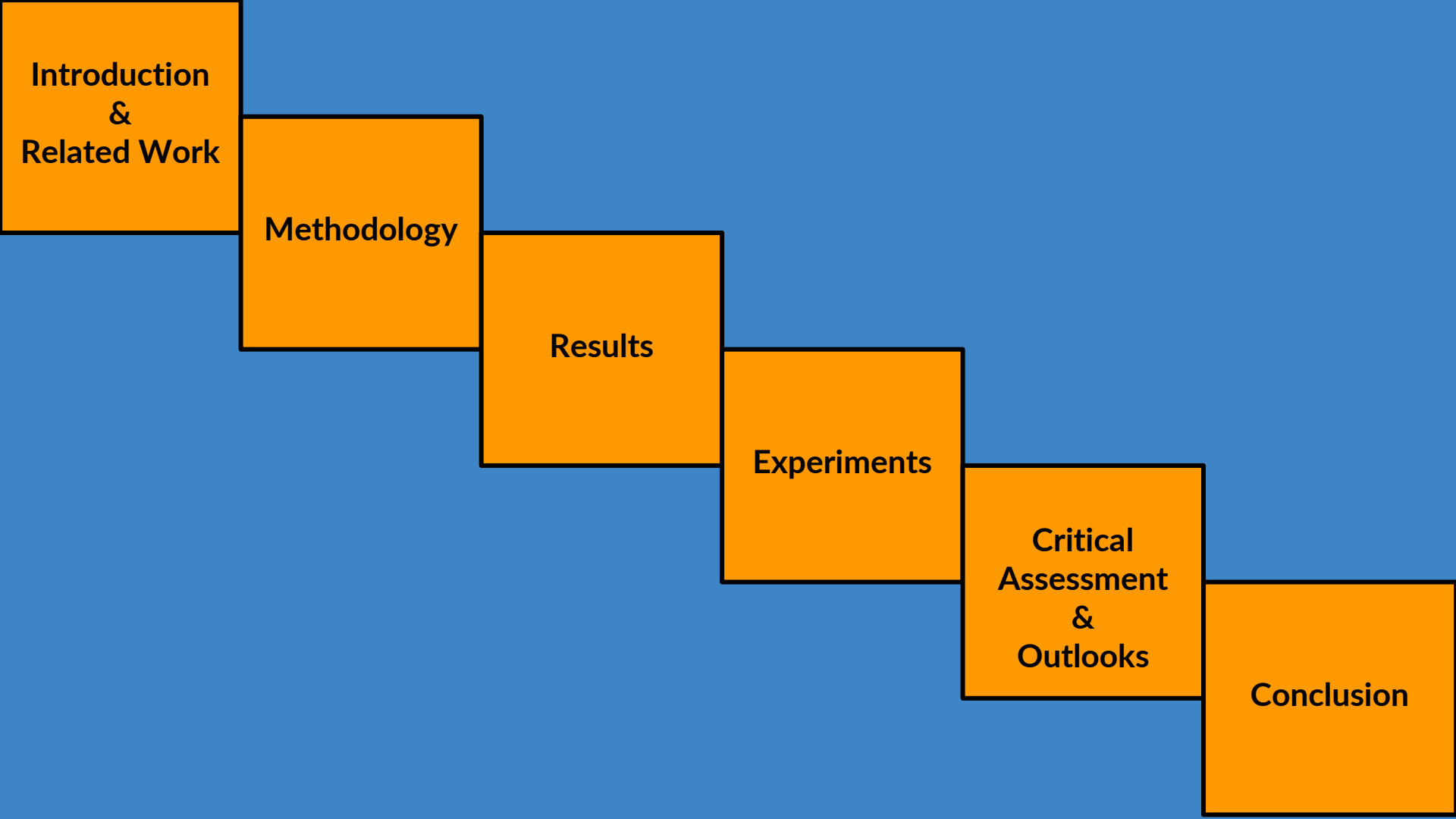
Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders

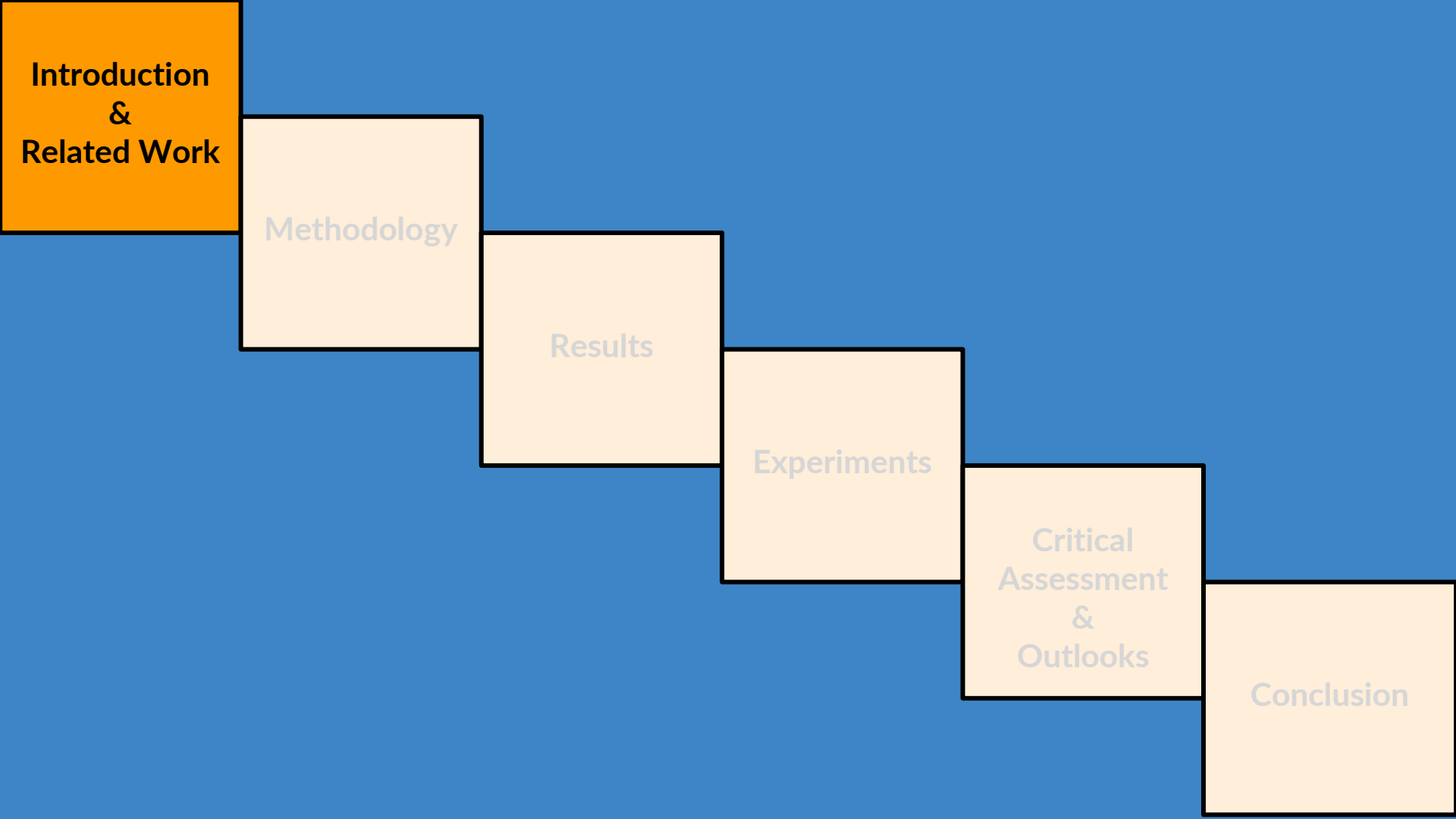
Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck et al. 2017

Student - Matteo MARENGO [1]
Teacher – Emmanuel BACRY [2]

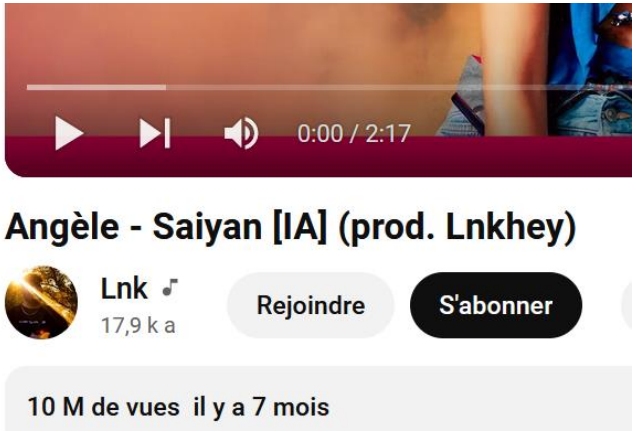
27/03/2024

[1] ENS Paris-Saclay, Msc MVA (*Mathématiques Vision Appentissage)*
[2] CEREMADE, Université Paris-Dauphine, PSL


Introduction
&
Related Work
Methodology
Results
Experiments
Critical
Assessment
&
Outlooks
Conclusion


Introduction
&
Related Work
Methodology
Results
Experiments
Critical
Assessment
&
Outlooks
Conclusion

Audio Synthesis: Recent progresses



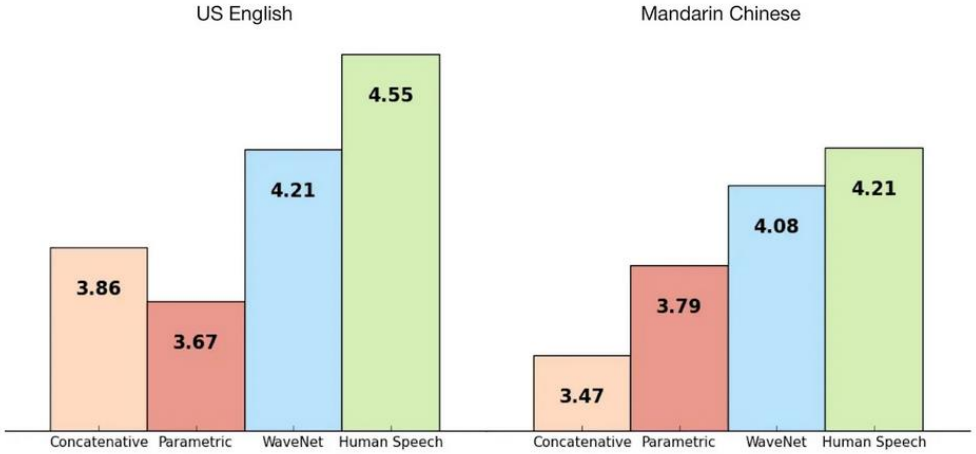
Générique Pokémon - Johnny Hallyday (AI Cover)

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

Introduction & Related Work

Methodology

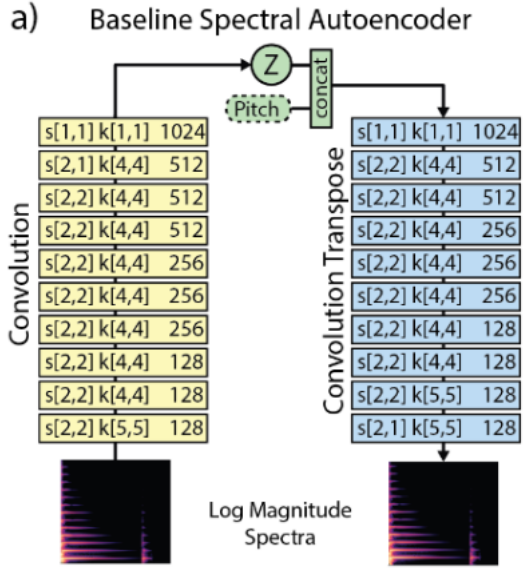
Results

Experiments

Critical Assessment & Outlooks

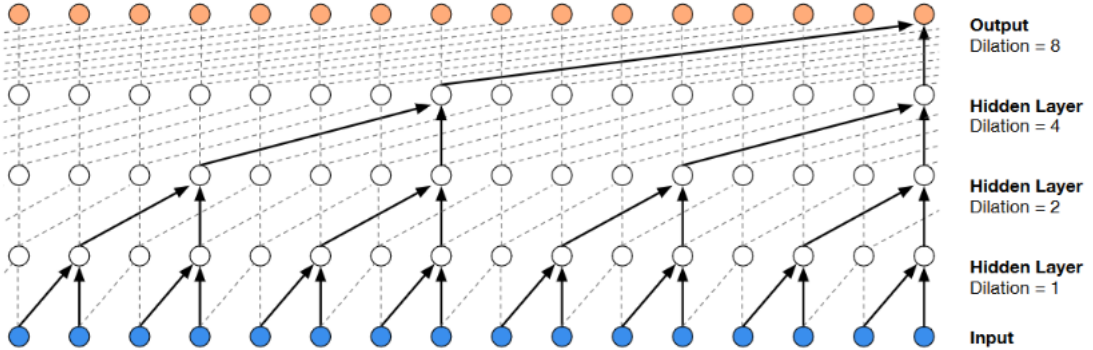
Conclusion

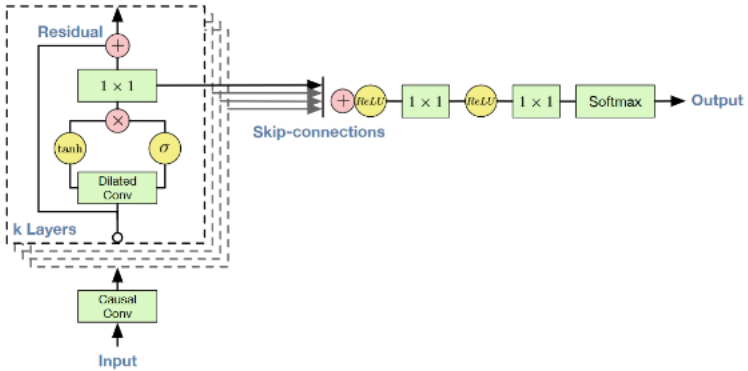
Spectral Autoencoder



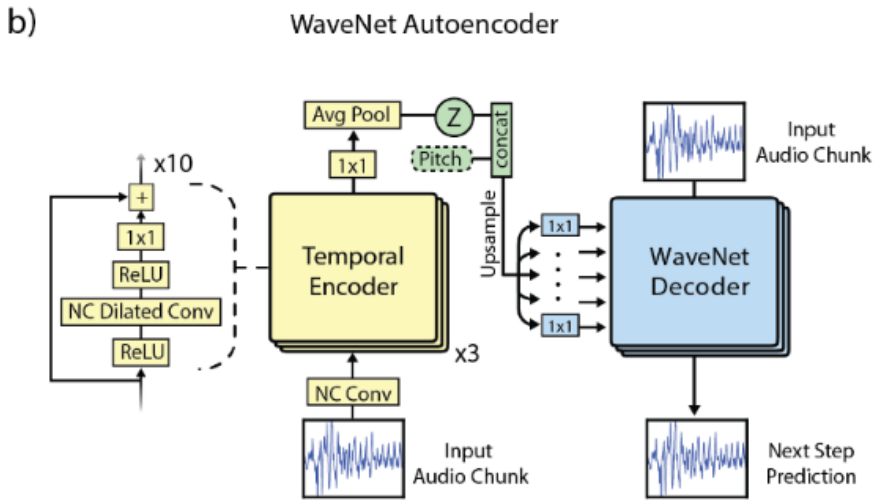
- Baseline autoencoder composed of convolutional structures
- Inspired by models in Computer Vision

WaveNet architecture





AutoEncoder WaveNet architecture

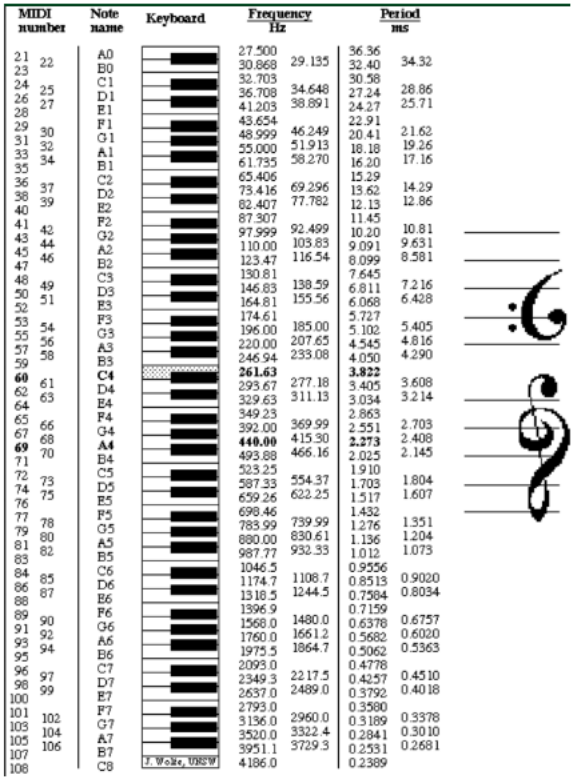


- Temporal encoder that captures hidden embeddings distributed in time.
- 30-layer residual network of dilated convolutions → generates a sequence of hidden codes that represent the audio temporally.

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.



Introduction & Related Work

Methodology

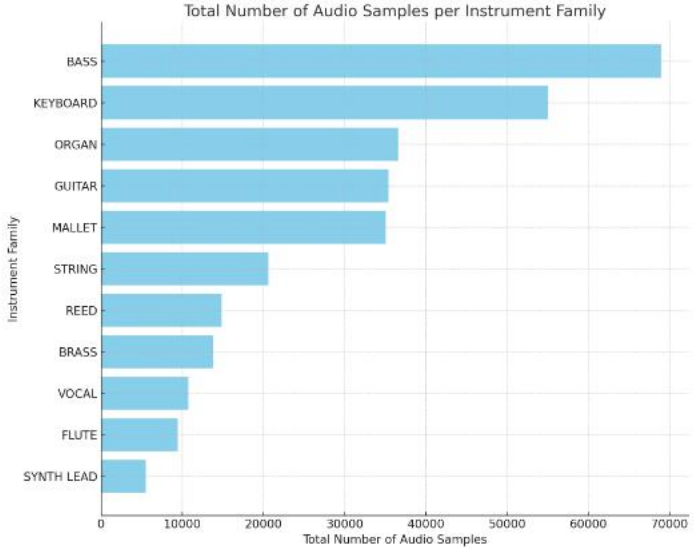
Results

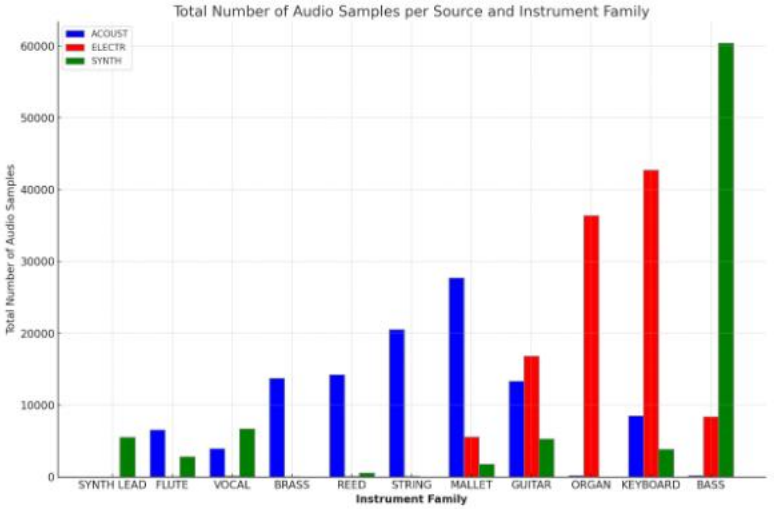
Experiments

Critical Assessment & Outlooks

Conclusion

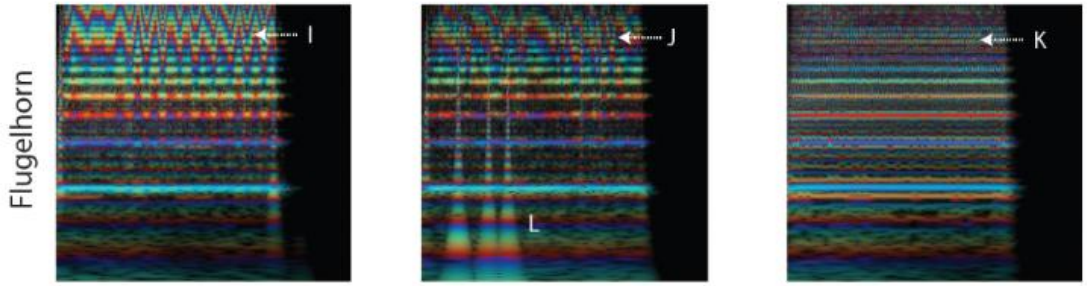
Nsynth dataset





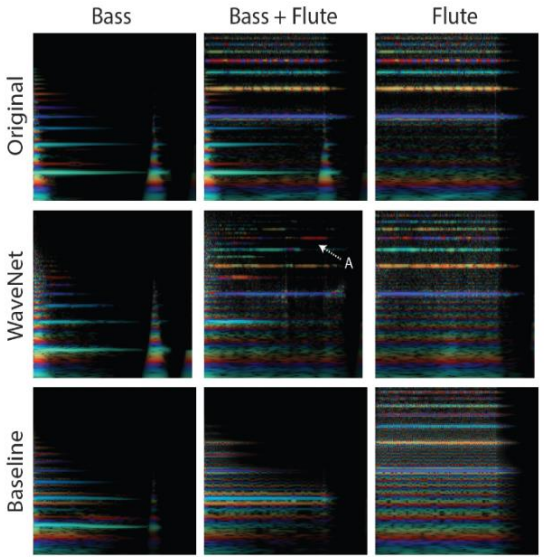
Reconstruction

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



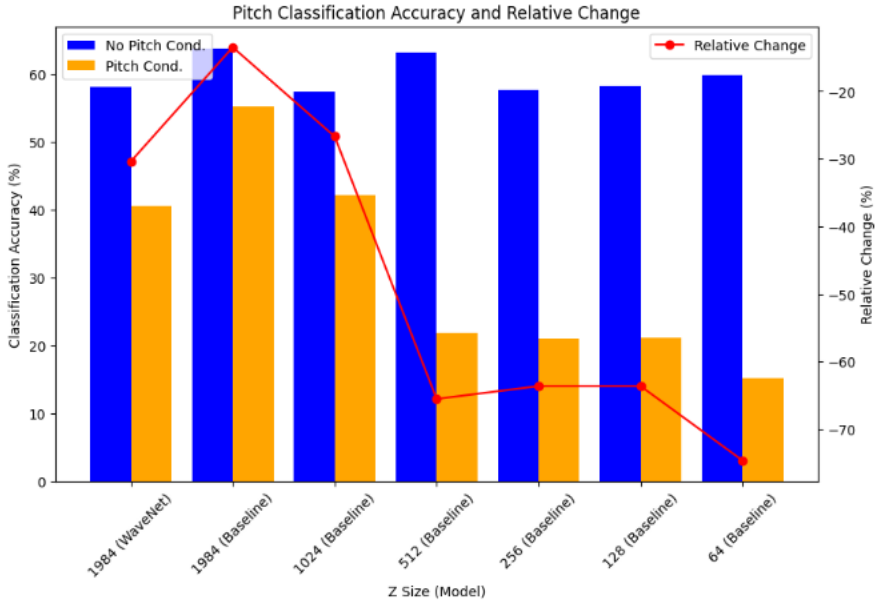
Interpolation in Timbre and Dynamics

- **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.



Introduction & Related Work

Methodology

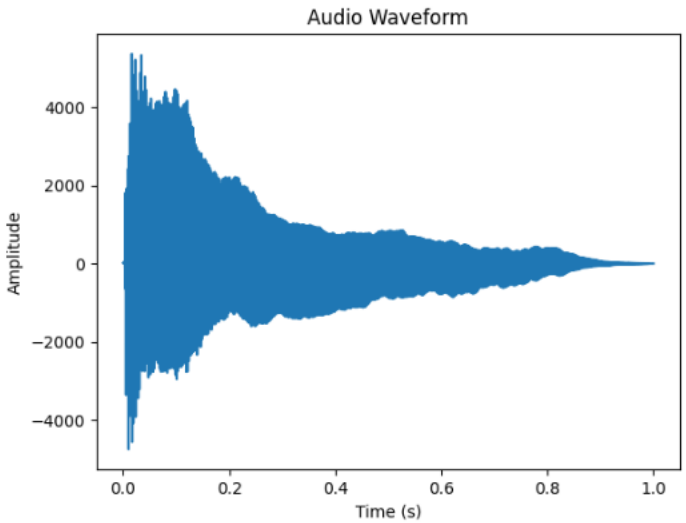
Results

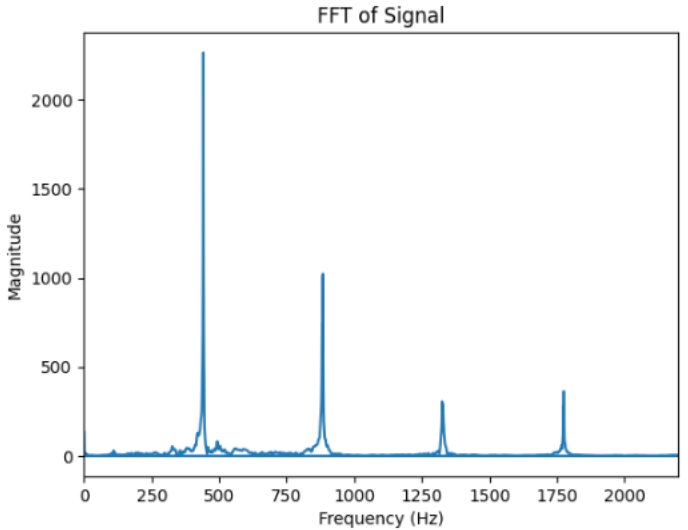
Experiments

Critical Assessment & Outlooks

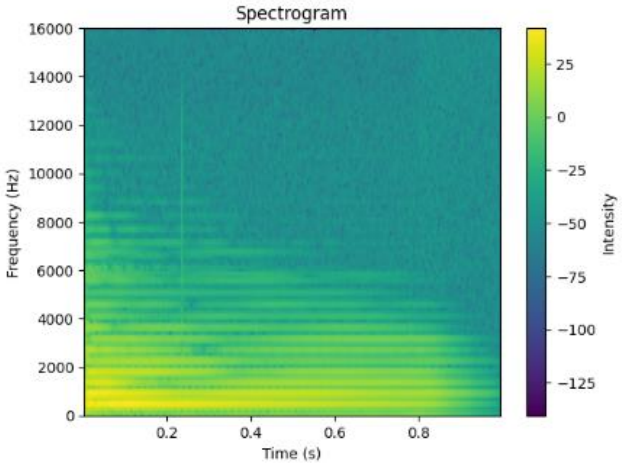
Conclusion

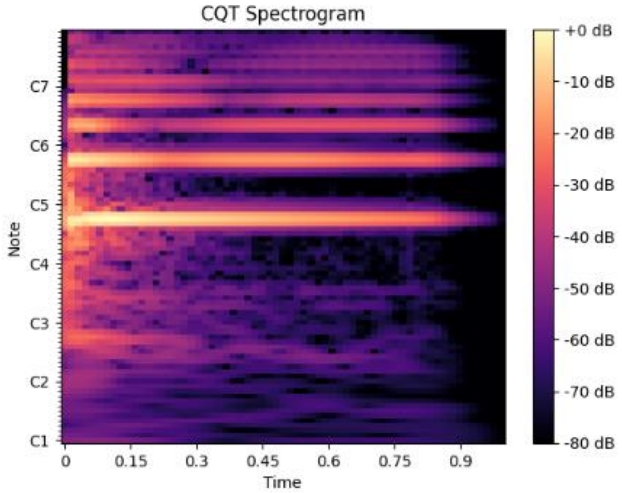
Study a simple note – A4



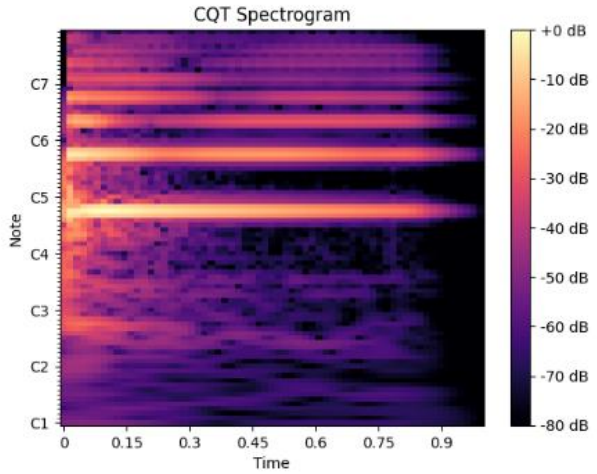


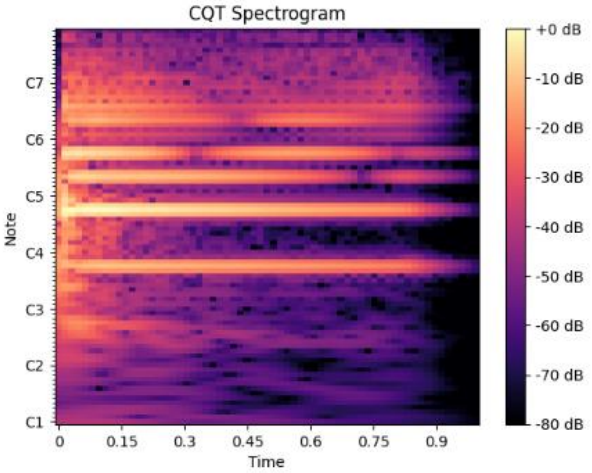
Comparison of the spectrograms



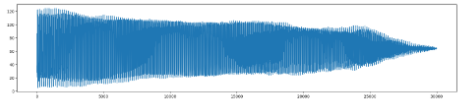


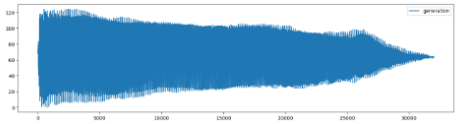
Comparison of A4 spectrogram with A3A4 spectrogram

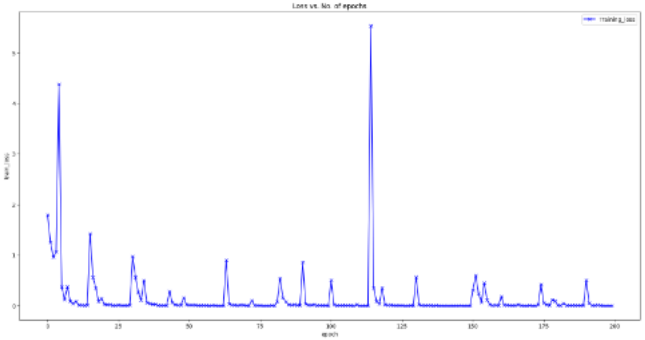


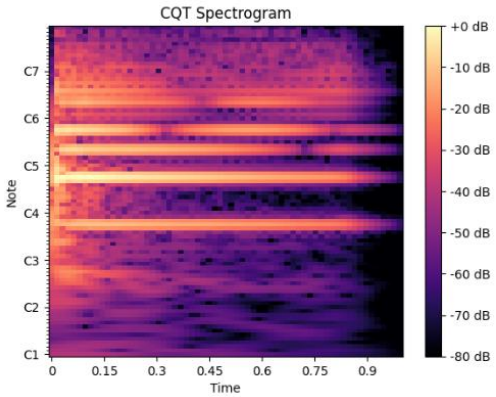


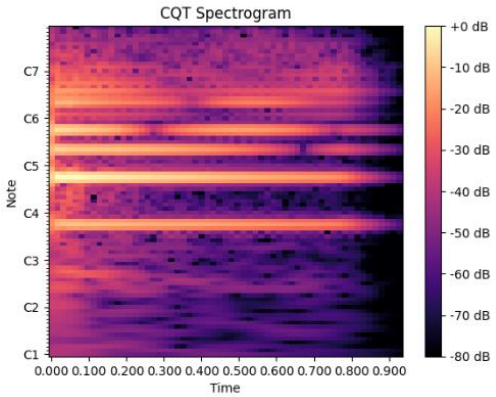
WaveNet architecture

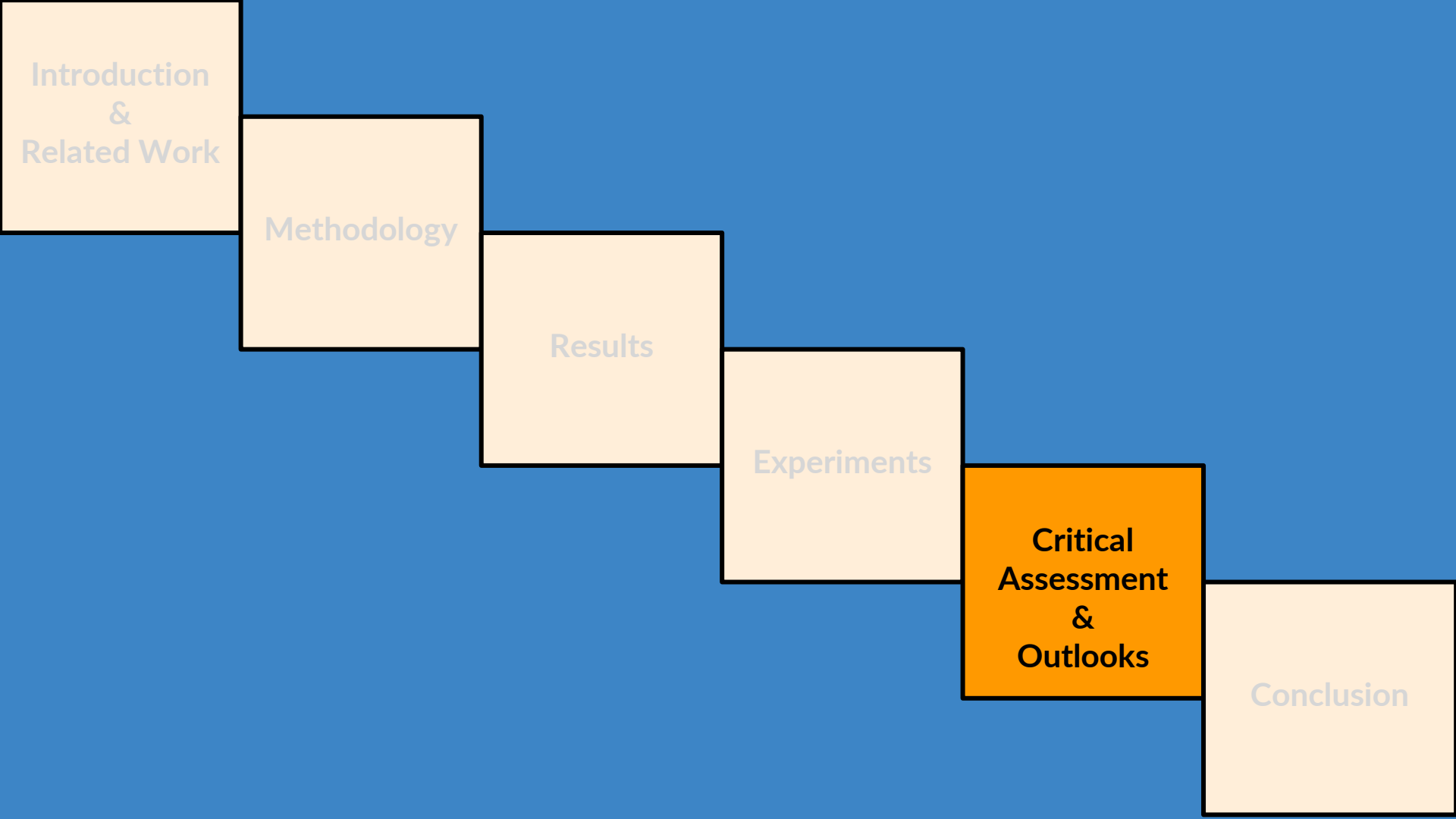











Introduction
&
Related Work
Methodology
Results
Experiments
Critical
Assessment
&
Outlooks
Conclusion

Strengths & Weaknesses of the paper

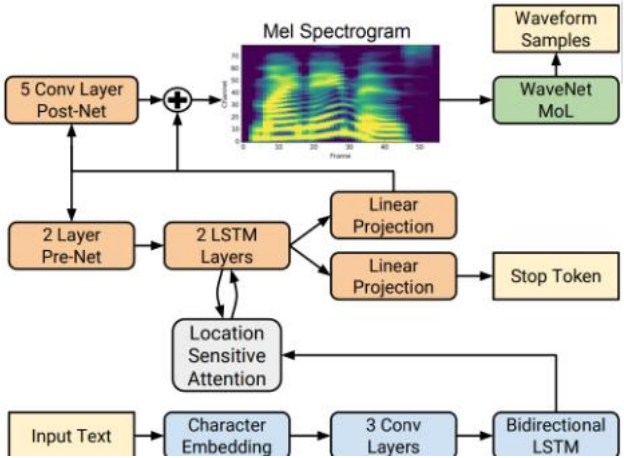
- Methodology is **thorough** and **well structured**.
- It enables reproducibility and further research (e.g they give the hyperparameters)
- **Qualitative** and **quantitative** measures.
- They conducted diverse and comprehensive experiments.

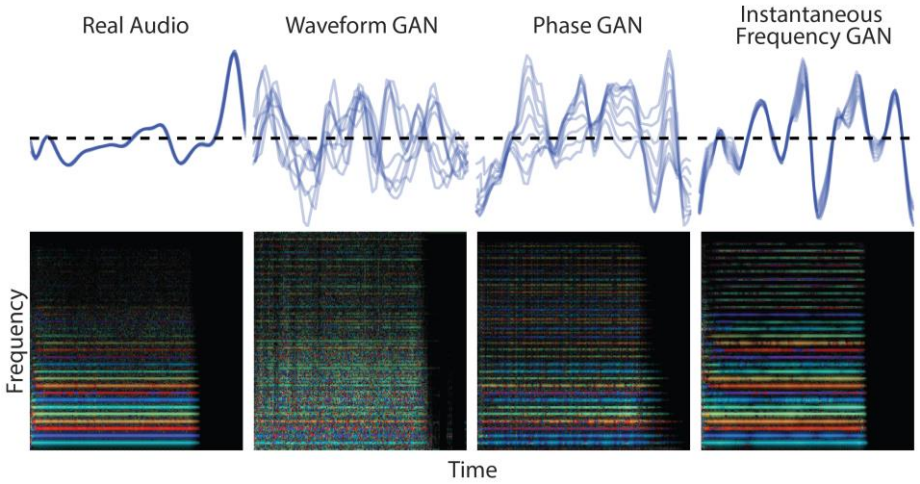
- They only compare the WaveNet autoencoder with a baseline autoencoder.
- The Nsynth dataset is not **balanced** within classes → Is it an issue?
- No study on **generalization** or overfitting.
- No training on other datasets.
- No discussion on the **computational part**.

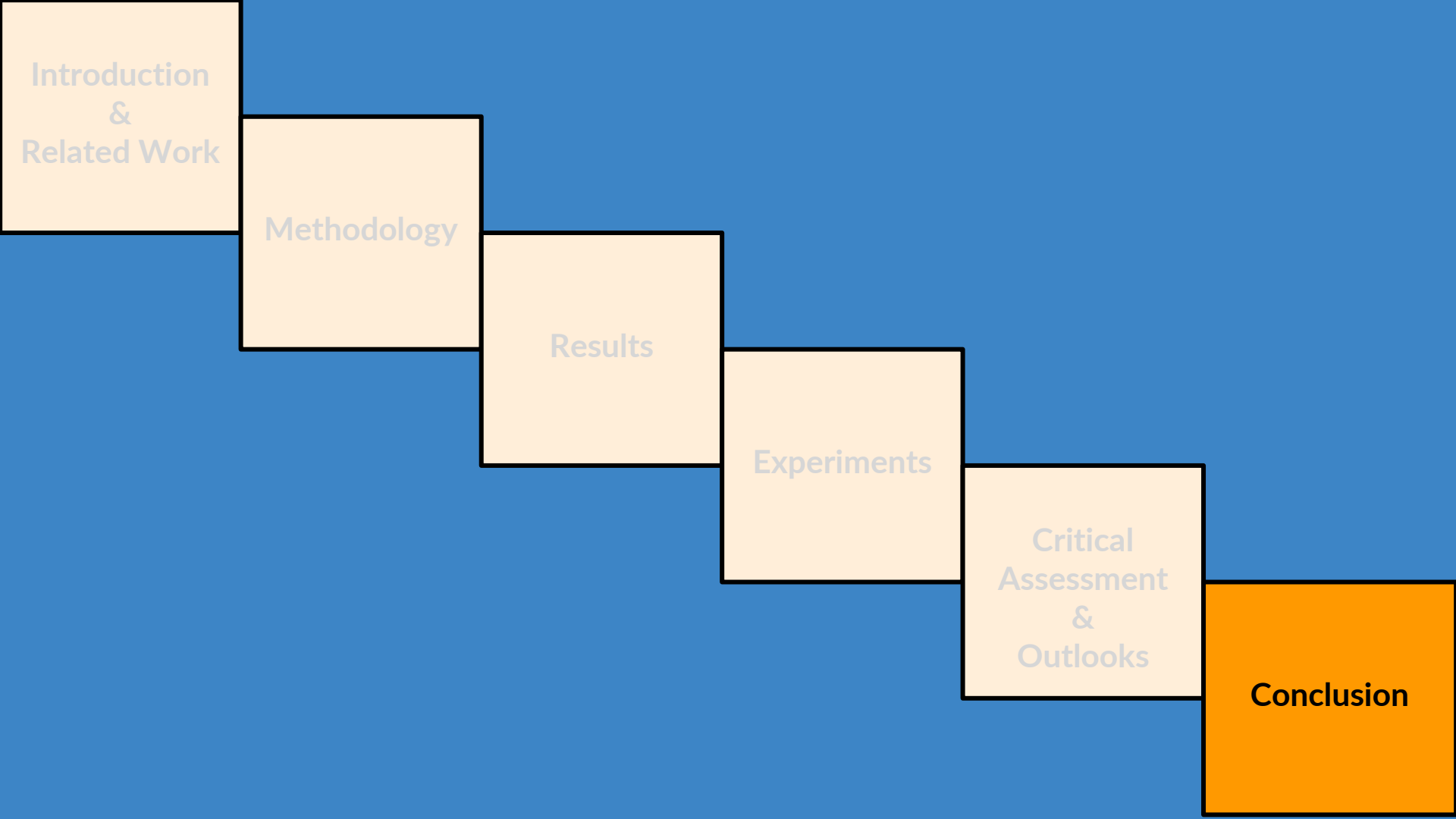
- Explore more memory-efficient neural network architectures (LSTM, Transformer)
- Improve the model's ability to separate pitch and timbre (e.g alternative conditioning strategies)
- More extensive validation on external datasets, add regularization techniques.

Outlooks

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



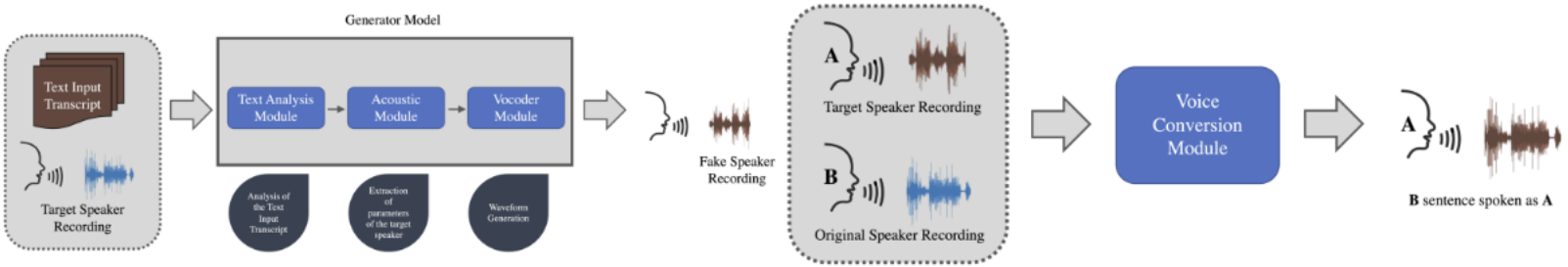



Introduction
&
Related Work
Methodology
Results
Experiments
Critical
Assessment
&
Outlooks
Conclusion

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!