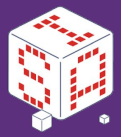


Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

Variational Autoencoders for Polyphonic Music Interpolation

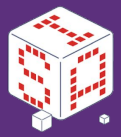
Pablo López Diéguez and Soo Von-Wun

TAAI 2020 – Technologies and Applications of Artificial Intelligence

Taipei, Taiwan

國立清華大學
NATIONAL TSING HUA UNIVERSITY

Outline



Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

01 Context

VAE

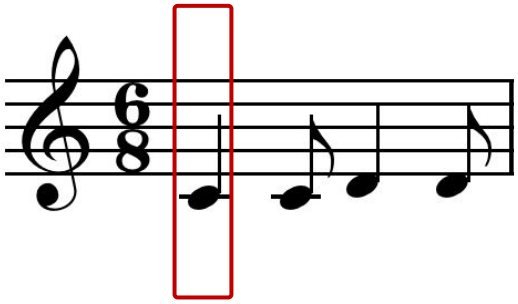
Variational Autoencoders for Polyphonic Music Interpolation

Variational Autoencoders for Polyphonic Music Interpolation

Polyphonic Music

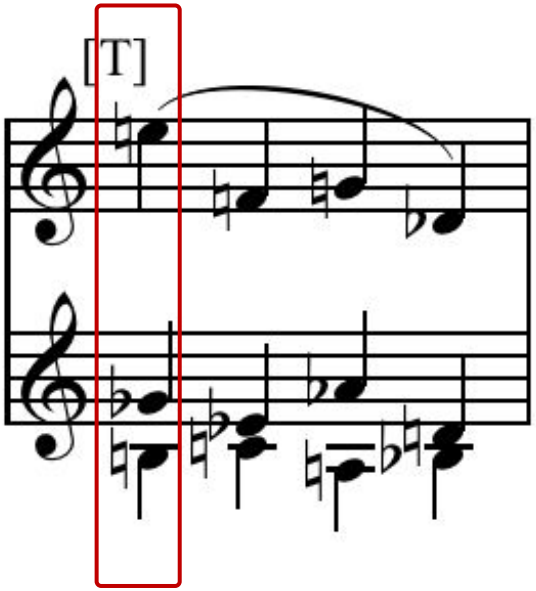
Texture (music)
- Monophonic music
- Polyphonic music

Monophonic



single note

Polyphonic



multiple notes

**timestep = time unit*

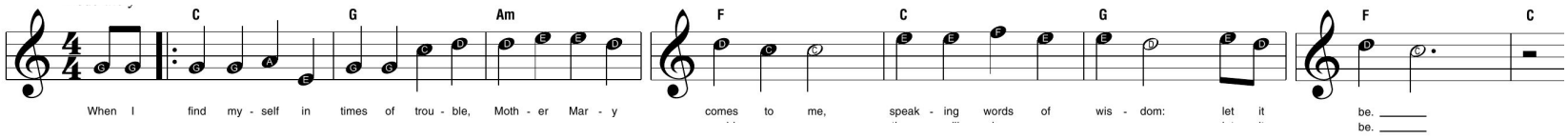
Variational Autoencoders for Polyphonic Music Interpolation

Interpolation

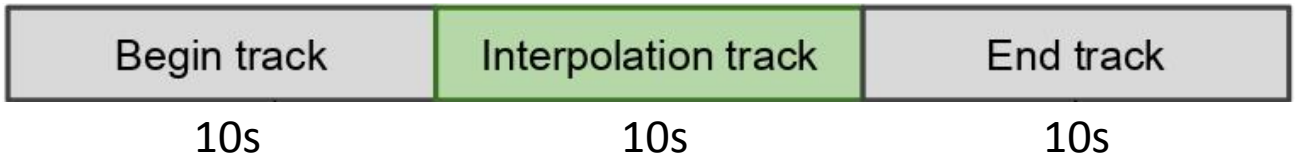
"Interpolation is a type of **estimation**, a method of constructing new **data points** within the range of a **discrete set** of known data points"
(Fleetwood, 1991)

Interpolation

In traditional Machine Learning, the generation of music is conditioned on the past events.

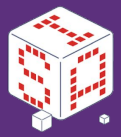


Past events *New prediction* *Future events*



Applications

- Music composition for scenes transitioning in cinema.
- Music composition between an advertisement and the TV show song.
- Real-time music interpolation for musicians.



Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

02 Related work

References

Google Brain magenta

"A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music"
(Roberts *et. al*, 2018)

- VAE for modelling long-term sequences.
- Tested performance for interpolation.

- Not focusing on interpolation
- Low level heuristic
- Monophonic
- Data not available
- High training cost

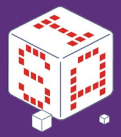
Linear sampling of latent-space: state-of-the-art method for interpolation

ETH zürich

"Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer"
(Brunner *et. al*, 2018)

- VAE for style transfer.
- Tested performance for interpolation.
- Training: 48h on a single GPU.

- Not focusing on interpolation
- Low level heuristic
- Style classifier required
- High training cost



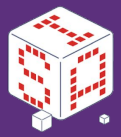
Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

03 Objectives

國立清華大學
NATIONAL TSING HUA UNIVERSITY

Objectives

01 Interpolation	Solving the interpolation problem.
02 Efficiency	Reducing the computational and time requirements of the state-of-the-art interpolation approaches.
03 Outperformance	Developing an interpolation technique that outperforms the state-of-the-art interpolation approaches.
04 Metrics	Proving the superiority of the models proposed with a set of objective metrics.



Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

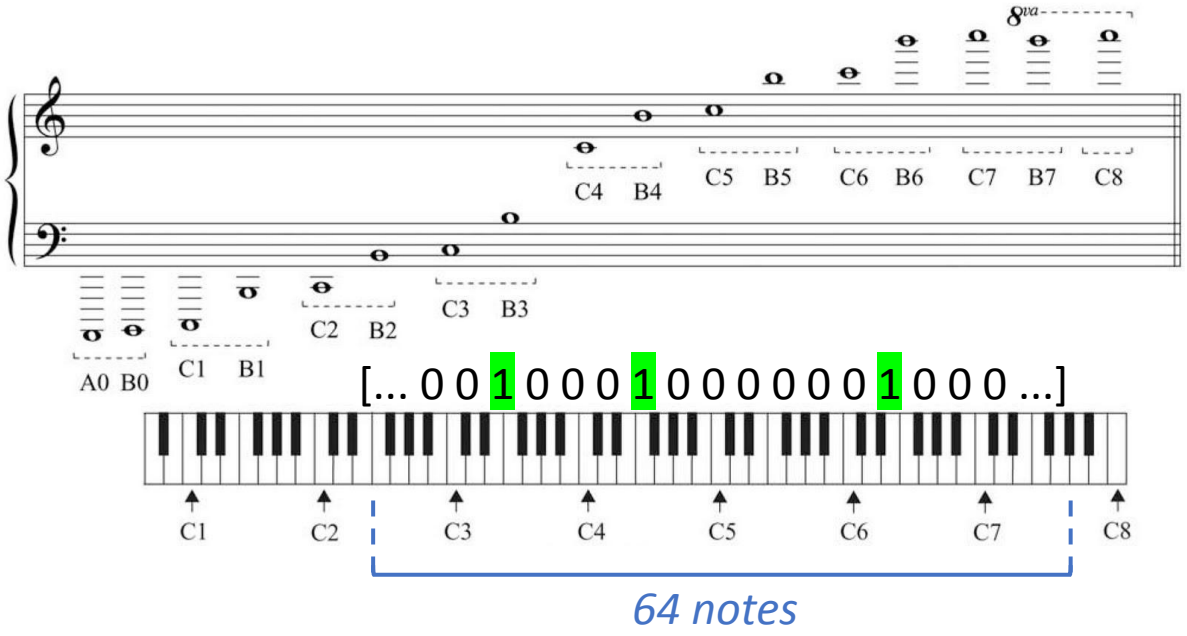
04 Theory

國立清華大學
NATIONAL TSING HUA UNIVERSITY

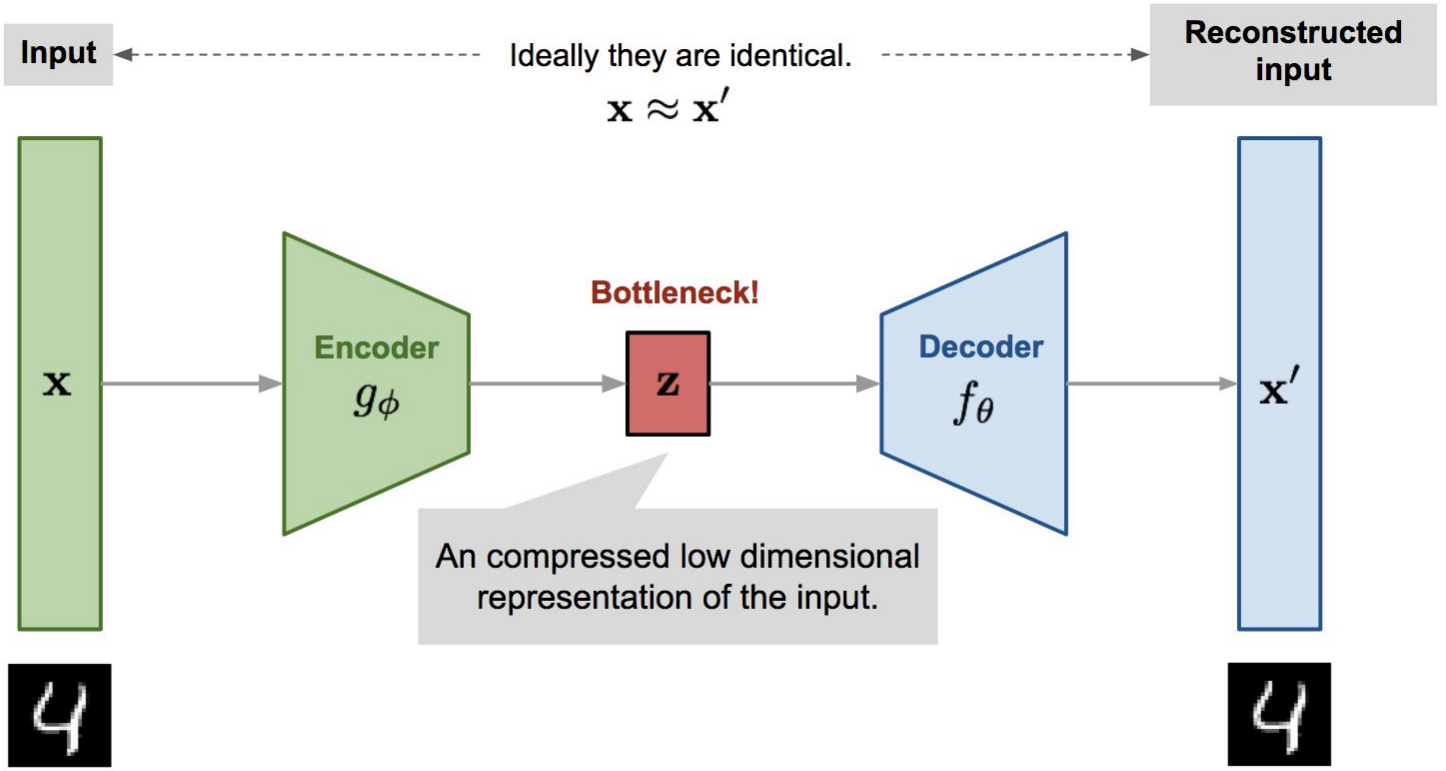
Music modelling

MIDI (Musical Instrument Digital Interface)

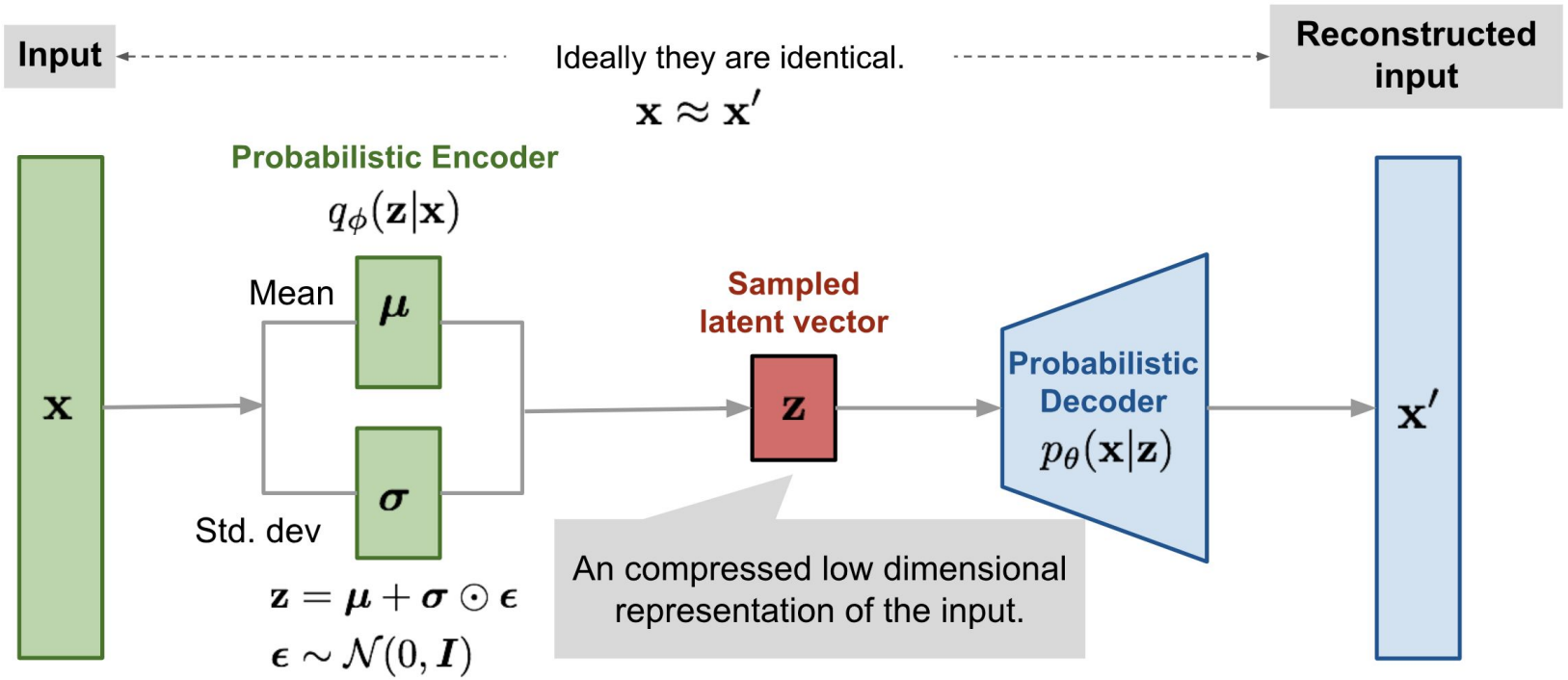
- Protocol to store musical information in a symbolic way.
- 128 values for pitches (notes).

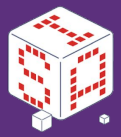


Autoencoder



Variational Autoencoder (VAE)





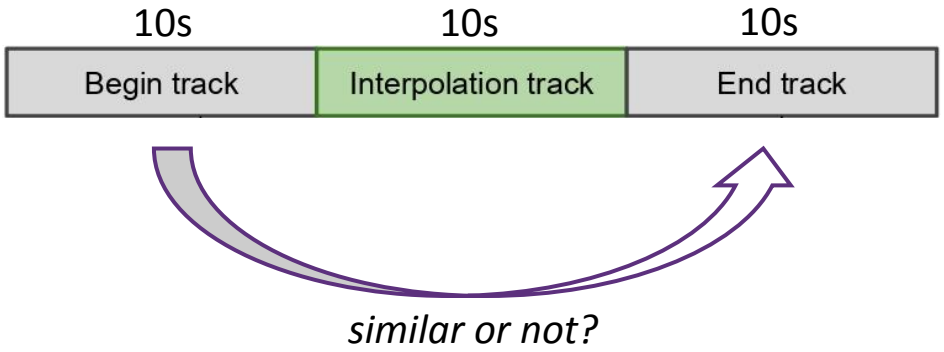
Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

05 Experiments

國立清華大學
NATIONAL TSING HUA UNIVERSITY

Dataset

- Lack of musical data for interpolation:



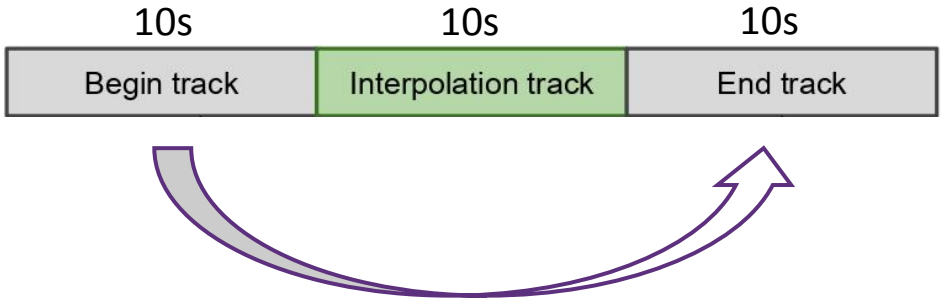
each training example

- Constructed **Hsinchu Interpolation Dataset** based on the Lahk MIDI Dataset:

  1) Extracted all the piano tracks.
  2) Divided them in segments of 30s.
  3) Evaluated similarity.

Dataset

- **Neural Network** to evaluate similarity:



- After evaluating all piano tracks:

Hsinchu Interpolation Dataset
containing 30,830 tracks of 30s each.

Experiments

01 Random Data

02 Bi-LSTM

03 VAE

04 VAE+NN

Experiments

01 Random Data

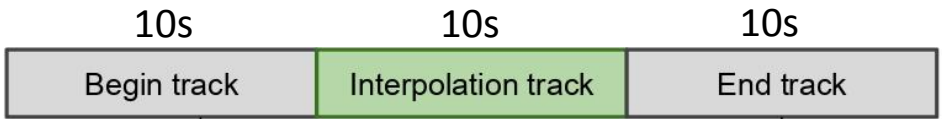
02 Bi-LSTM

03 VAE

04 VAE+NN

01 Random Data

- Naive and **non-Machine-Learning** approach to the interpolation problem.
- Use a Bernoulli random variable to choose the new element of the interpolation track.

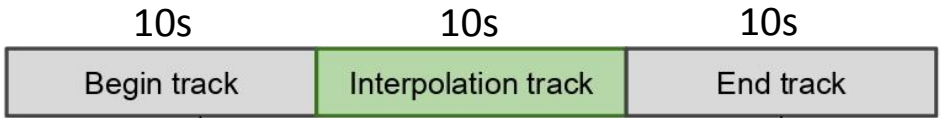


- Every timestep (time unit) we randomly choose an element from the begin track or from the end track.

Baseline model 01

02 Bi-LSTM

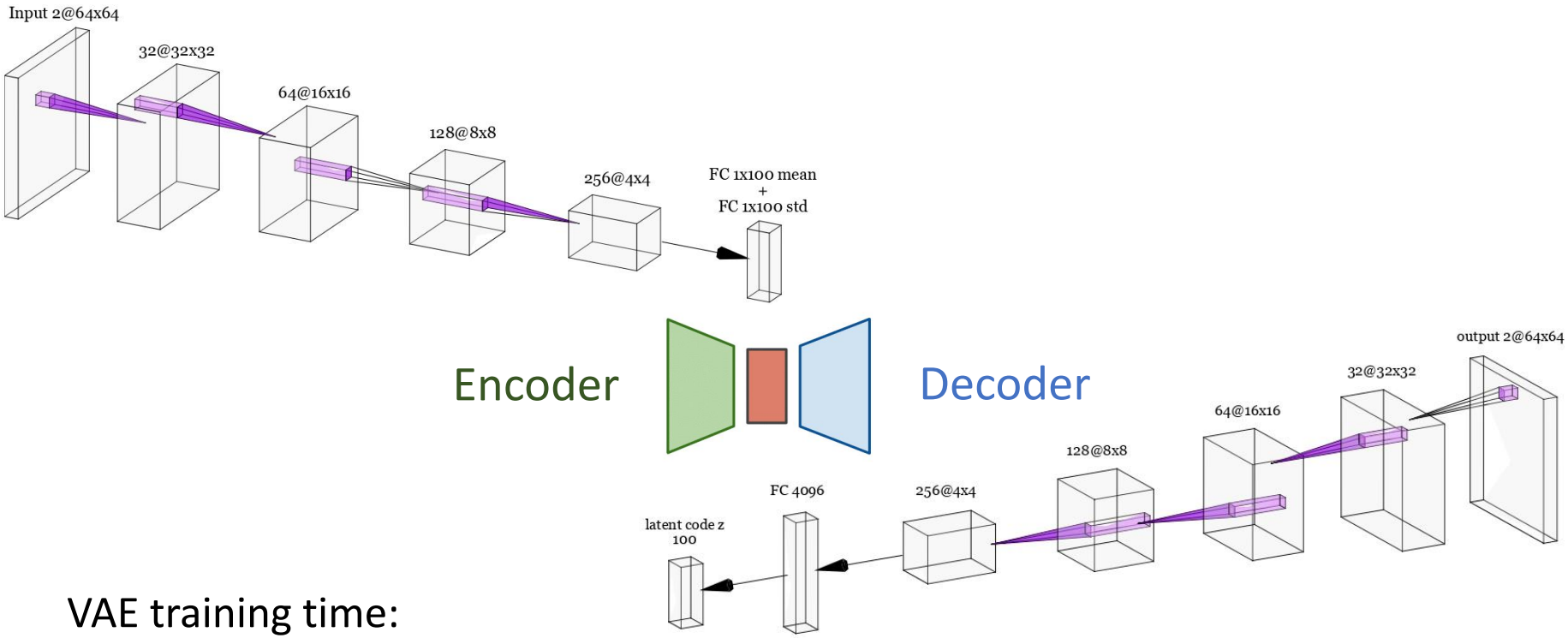
- Bidirectional Long Short-Term Memory.
- LSTM can capture local structure of **melody** and long-term structure of **musical style**.
- Bidirectional models can learn from begin and end tracks.



Baseline model 02

03 VAE

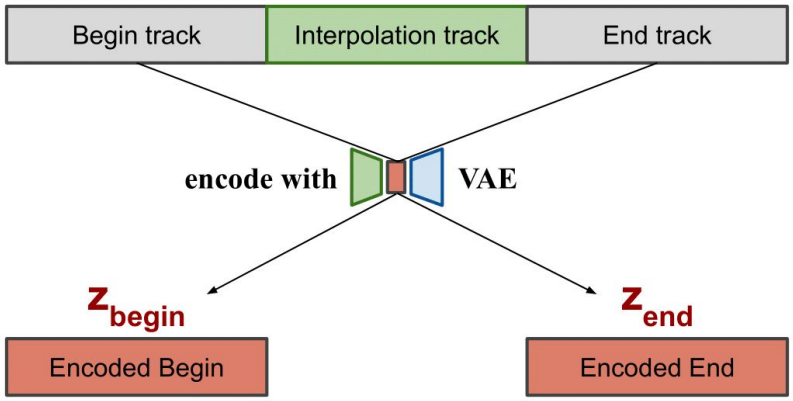
Encoder and decoder are based on convolutional neural networks.



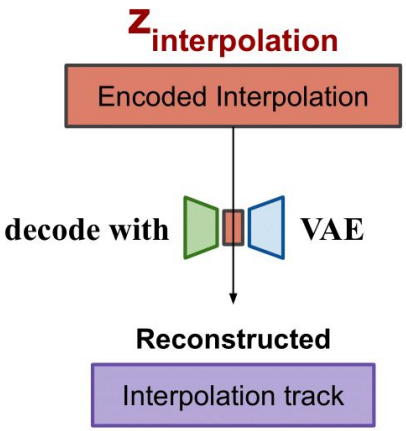
VAE training time:
4h on a single GPU

03 VAE

Interpolation method: linear sampling of latent space



$$\frac{z_{begin} + z_{end}}{2} = z_{interpolation}$$



Ideally they are identical.

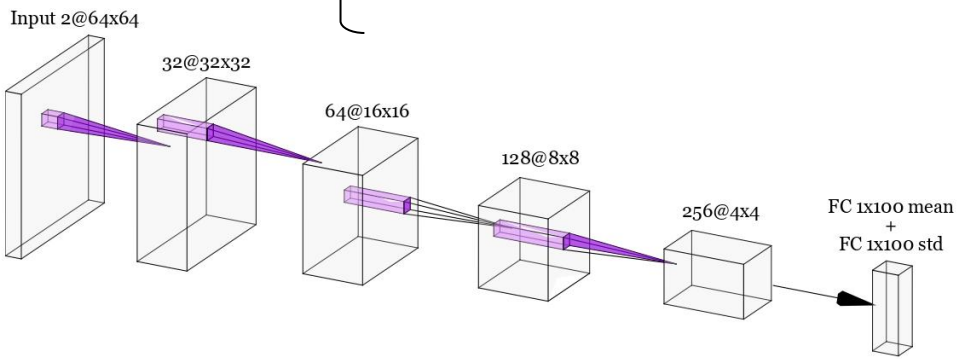
Interpolation track ≈ Interpolation track

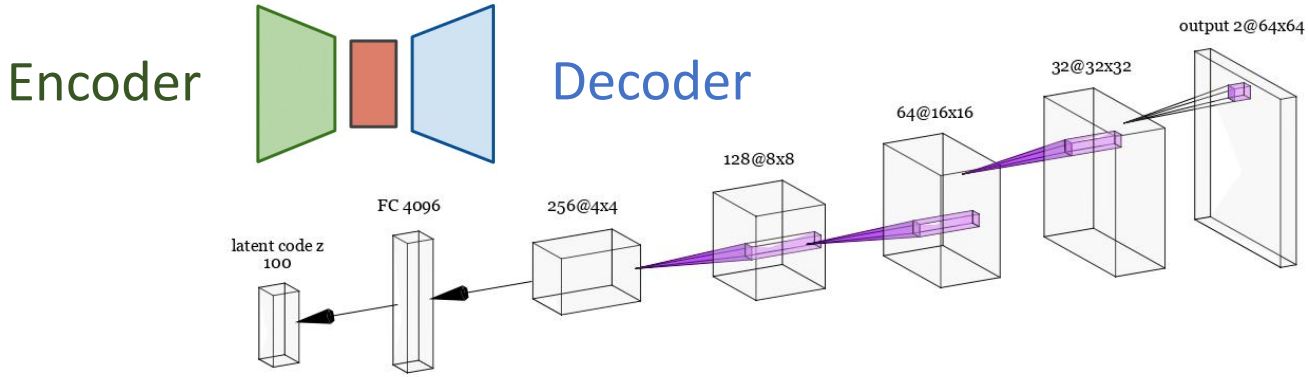
04 VAE + NN

VAE + NN
- VAE
- NN

04 VAE + NN

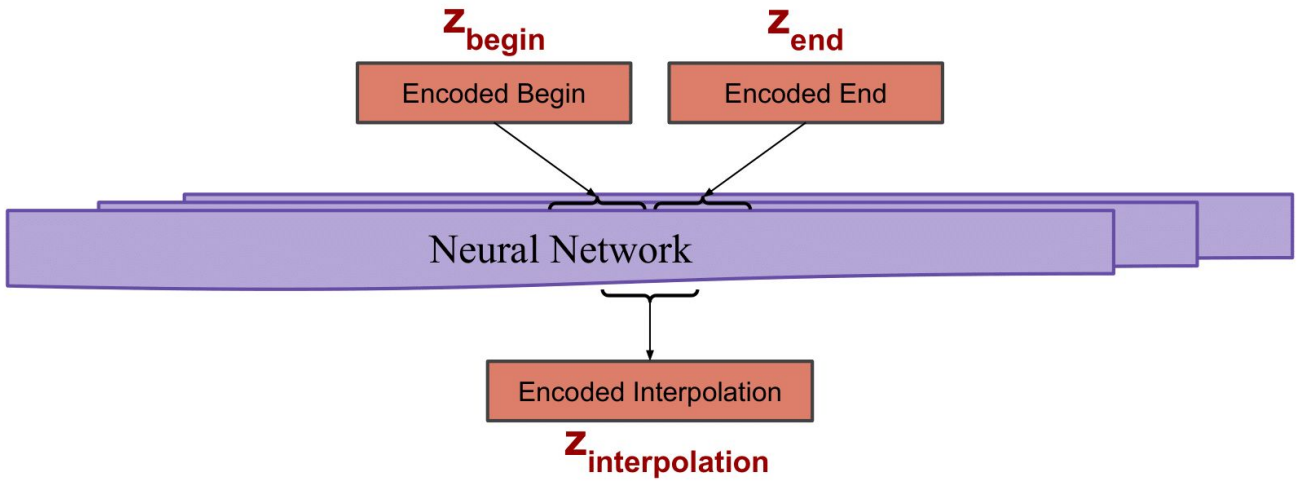
VAE + NN — VAE





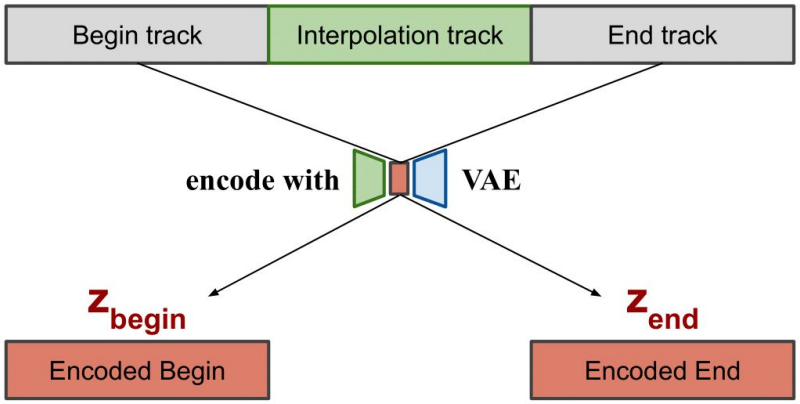
04 VAE + NN

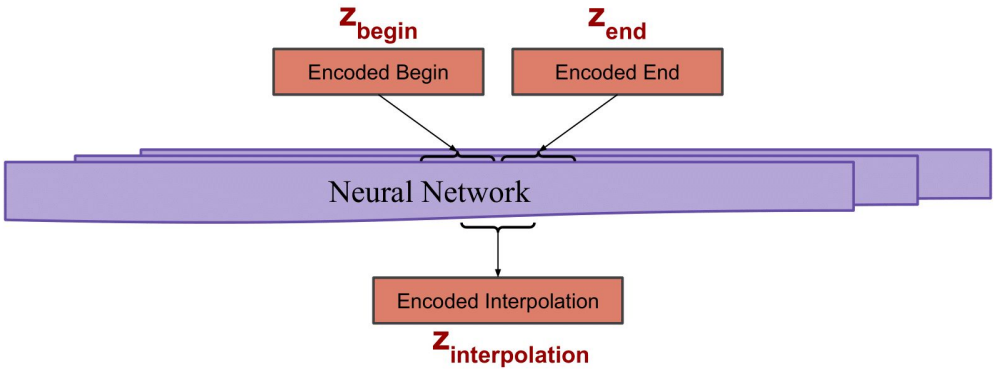
VAE + NN — NN

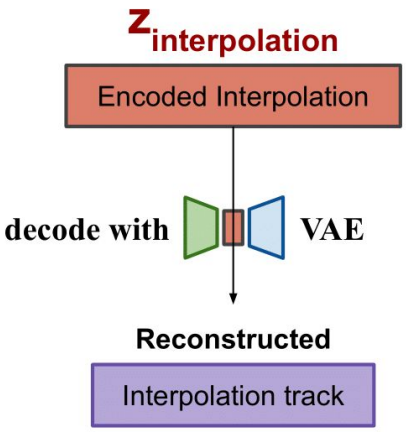


04 VAE + NN

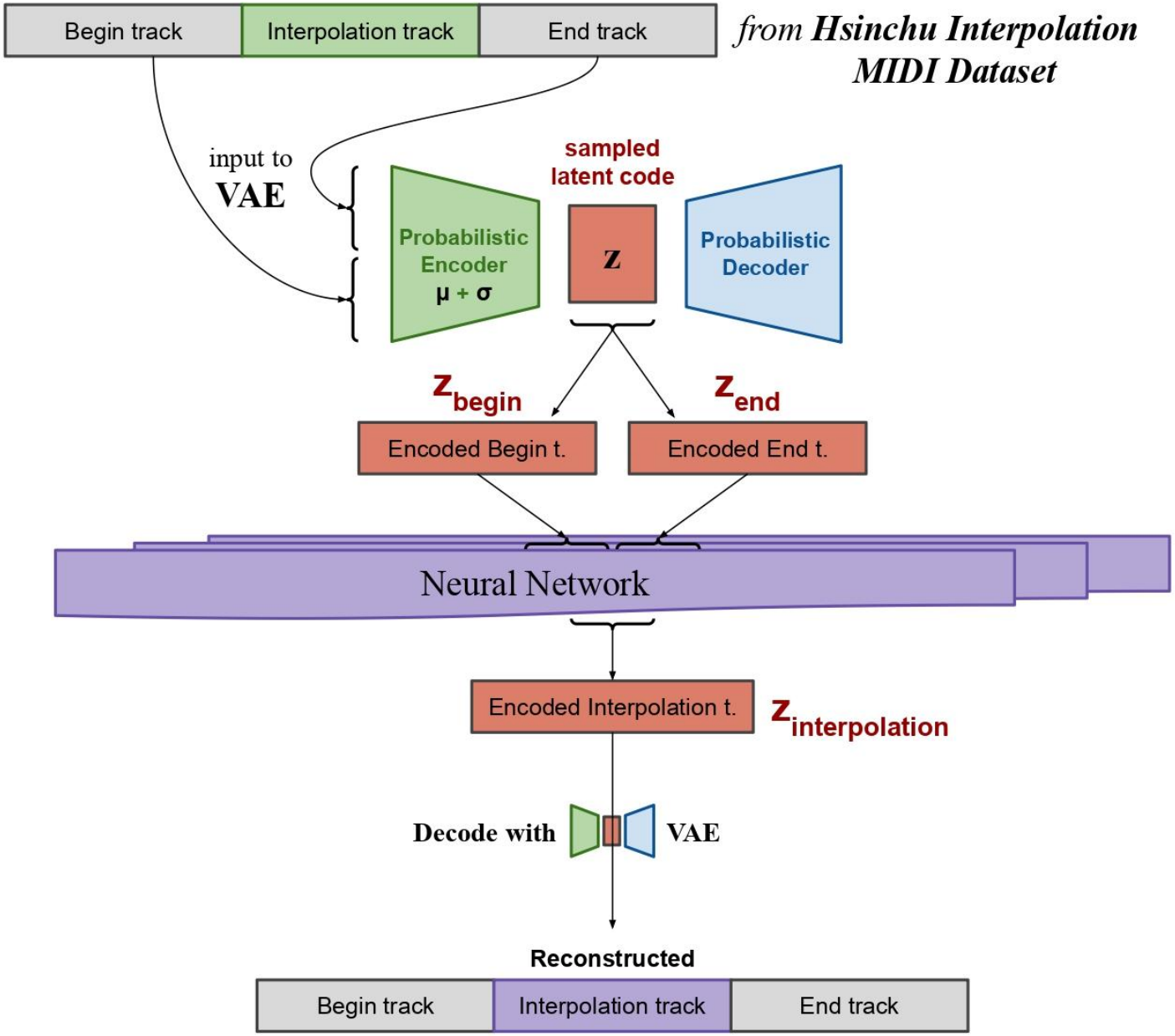
Interpolation method: direct estimation of the interpolation encoded vector

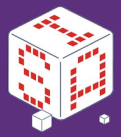






Ideally they are identical.


Begin track
Interpolation track
End track
from Hsinchu Interpolation MIDI Dataset
input to VAE
sampled latent code
Probabilistic Encoder
μ + σ
z
Probabilistic Decoder
z begin
z end
Encoded Begin t.
Encoded End t.
Neural Network
Encoded Interpolation t.
z interpolation
Decode with
VAE
Reconstructed
Begin track
Interpolation track
End track



Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

06 Results

Evaluation

- Objective
  - Compute error on the actual tracks
  - Compute error on the latent vectors
- Subjective

Evaluation

- Objective
  - Compute error on the actual tracks
  - Compute error on the latent vectors

  Error: Mean squared error (**MSE**) from previous papers.
  The **lower** the closer to the human compositions.

- Subjective

Objective evaluation

- Objective — Compute error on the actual tracks

Random (1)	Bi-LSM (2)	VAE (3)	VAE+NN (4)	Improv. (4) vs (1)	Improv. (4) vs (2)	Improv. (4) vs (3)
0.0581	0.0470	0.0494	**0.0431**	25.82%	8.30%	12.75%

- **VAE+NN** is better than Random Data baseline
- **VAE+NN** is better than Bi-LSTM baseline
- **VAE+NN** is better than **VAE**

Objective evaluation

- Objective — Compute error on the actual tracks

τ	Bi-LSTM (1)	VAE (2)	VAE+NN (3)	Improvement (3) vs (1)	Improvement (3) vs (2)
0.10	**0.0600**	0.0934	0.0840	- 40%	10.06%
0.20	0.0554	0.0673	**0.0464**	16.25%	31.05%
0.30	0.0528	0.0523	**0.0384**	27.27%	26.58%
0.40	0.0479	0.0443	**0.0371**	22.55%	16.03%
0.50	0.0447	0.0375	**0.0364**	18.57%	2.93%
0.60	0.0434	0.0375	**0.0365**	15.90%	2.67%
0.70	0.0401	0.0374	**0.0364**	9.23%	2.67%
0.80	0.0398	0.0374	**0.0364**	8.54%	2.67%
0.90	0.0392	0.0373	**0.0364**	7.14%	2.41%

- **VAE** is better than Bi-LSTM baseline in 7 out of 9 thresholds
- **VAE+NN** is better than Bi-LSTM baseline in 8 out of 9 thresholds.
- **VAE+NN** is better than **VAE** in all thresholds

Objective evaluation

- Objective — Compute error on the latent vectors

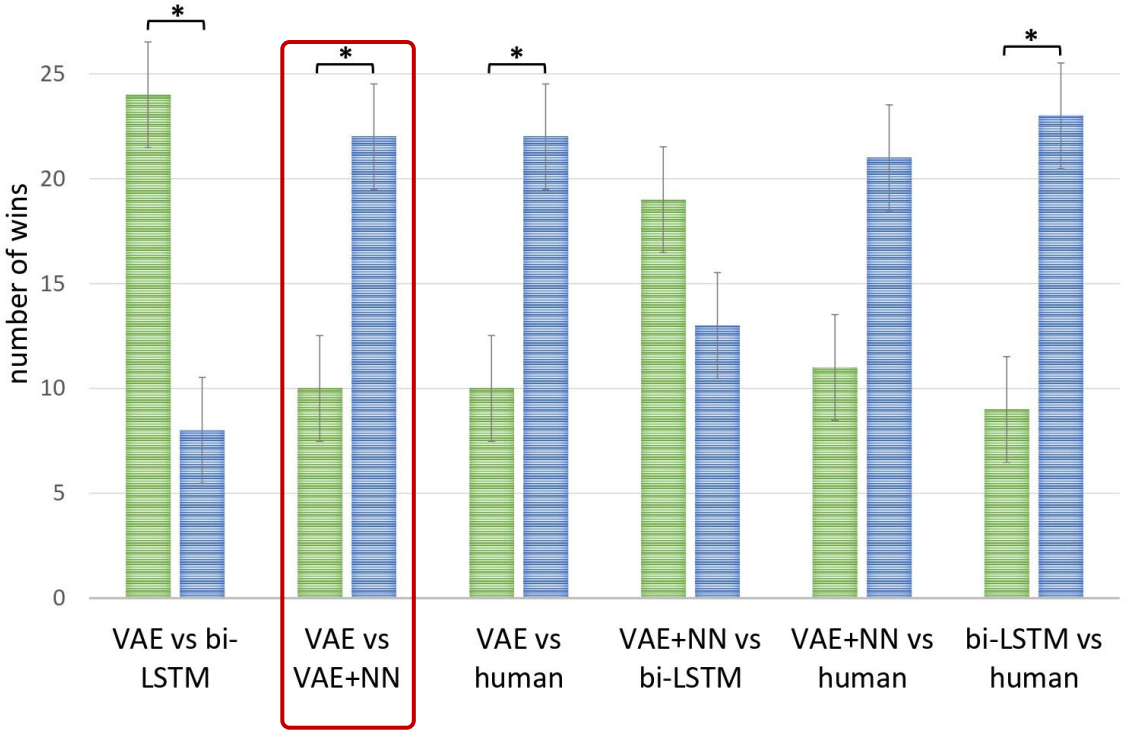
VAE (1)	VAE+NN (2)	Improvement (2) vs (1)
1.3726	**1.0611**	22.69%

- **VAE+NN** is better than **VAE** by 22.69%

Subjective evaluation

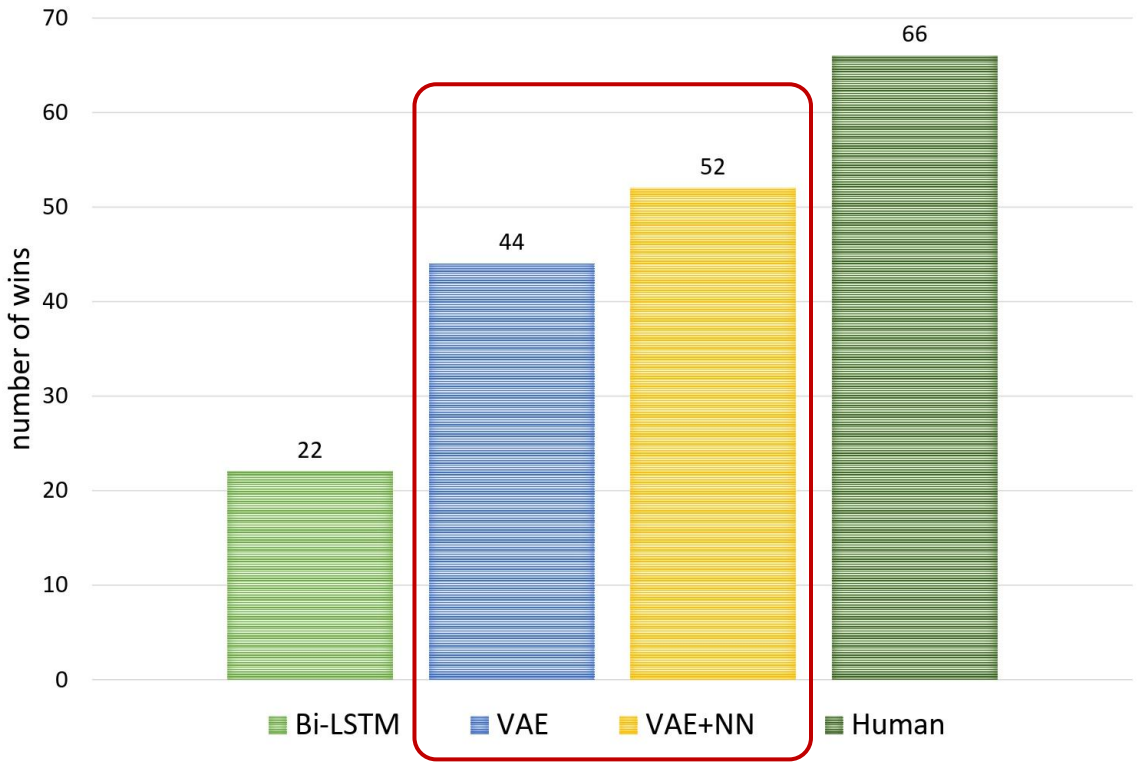
- **32** participants from **8** different countries in Africa, America, Asia and Europe.
- **65.6%** had a musical background.
- Pair-wise comparisons of tracks from all the models.
- All possible combinations were compared.

Subjective evaluation



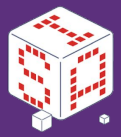
- VAE and VAE+NN models were preferred over the bi-LSTM baseline.
- Human interpolations were preferred in general.
- VAE+NN was preferred over VAE.
- Our interpolation approach was preferred over the state-of-the-art approach.

Subjective evaluation



Subjective evaluation

10s	10s	10s
Begin track	Interpolation track	End track
human 01	**VAE**	**human 02**



Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

07 Conclusion

Conclusion

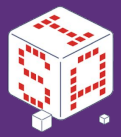
1. Used **VAE** to solve the problem of polyphonic music interpolation.
2. Reduced VAE training due to the **Hsinchu Interpolation Dataset** created.
3. Our proposed models (VAE and VAE+NN) outperform the baselines.
4. Our interpolation approach (VAE+NN) outperforms the state-of-the-art interpolation approaches (VAE):
    - Objective evaluation
    - Subjective evaluation

Future work

1. Explore other **architectures** for VAE.

2. **Expand** the Hsinchu Interpolation Dataset.

3. Explore other Machine Learning **models** to directly estimate the encoded **interpolation** vector.



Institute of Information Systems and Applications
College of Electrical Engineering and Computer Science

Variational Autoencoders for Polyphonic Music Interpolation

Pablo López Diéguez and Soo Von-Wun

TAAI 2020 – Technologies and Applications of Artificial Intelligence

Taipei, Taiwan

國立清華大學
NATIONAL TSING HUA UNIVERSITY