Generation of Symbolic Music Based on MusicVAE

Jakob Lerch

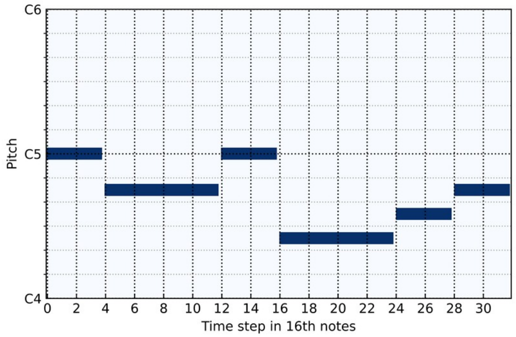
Supervisors:
Prof. Dr. Shardt
Dr. Andrew McLeod
Dr. Jakob Abeßer

July 27, 2023

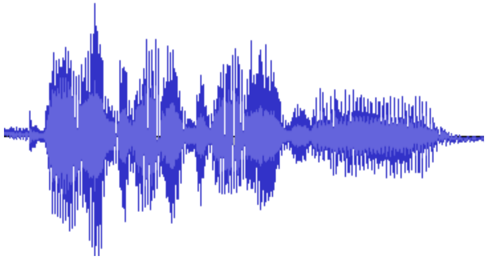
Music generation could be **supportive for composition or live performances**.

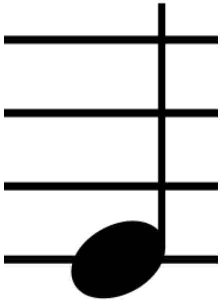
Picture retrieved July 16, 2023 from
https://pixabay.com/photos/music-producer-studio-actor-audio-4507819/

Introduction

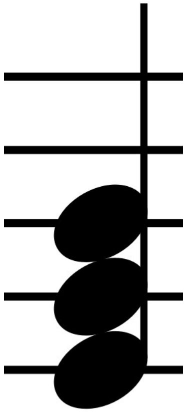


or





or



unconditioned or conditioned

Introduction: Objectives

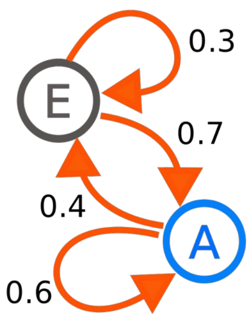
(1) review **state of the art**

(2) **re-implement** MusicVAE

(3) **evaluate quality** of generated excerpts

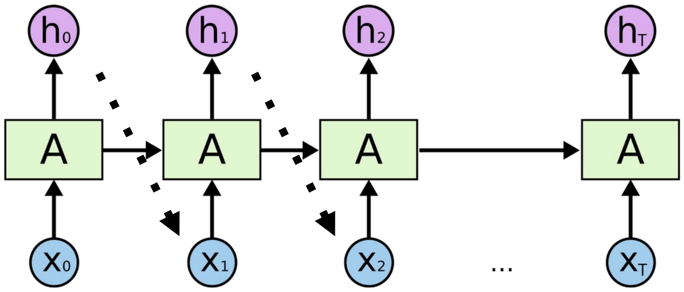
State of the Art

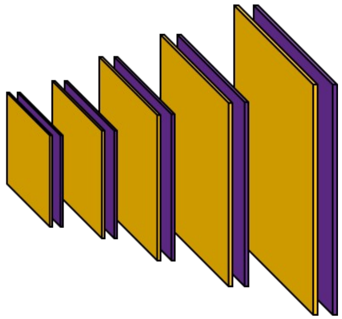
State of the Art



(1)

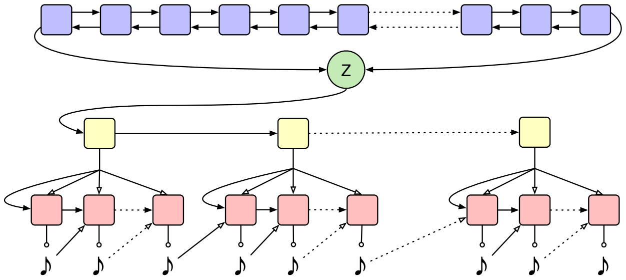
(2)





(3)

(4)



(1) Picture retrieved July 17, 2023 from https://en.wikipedia.org/wiki/Markov_chain#/media/File:Markovkate_01.svg

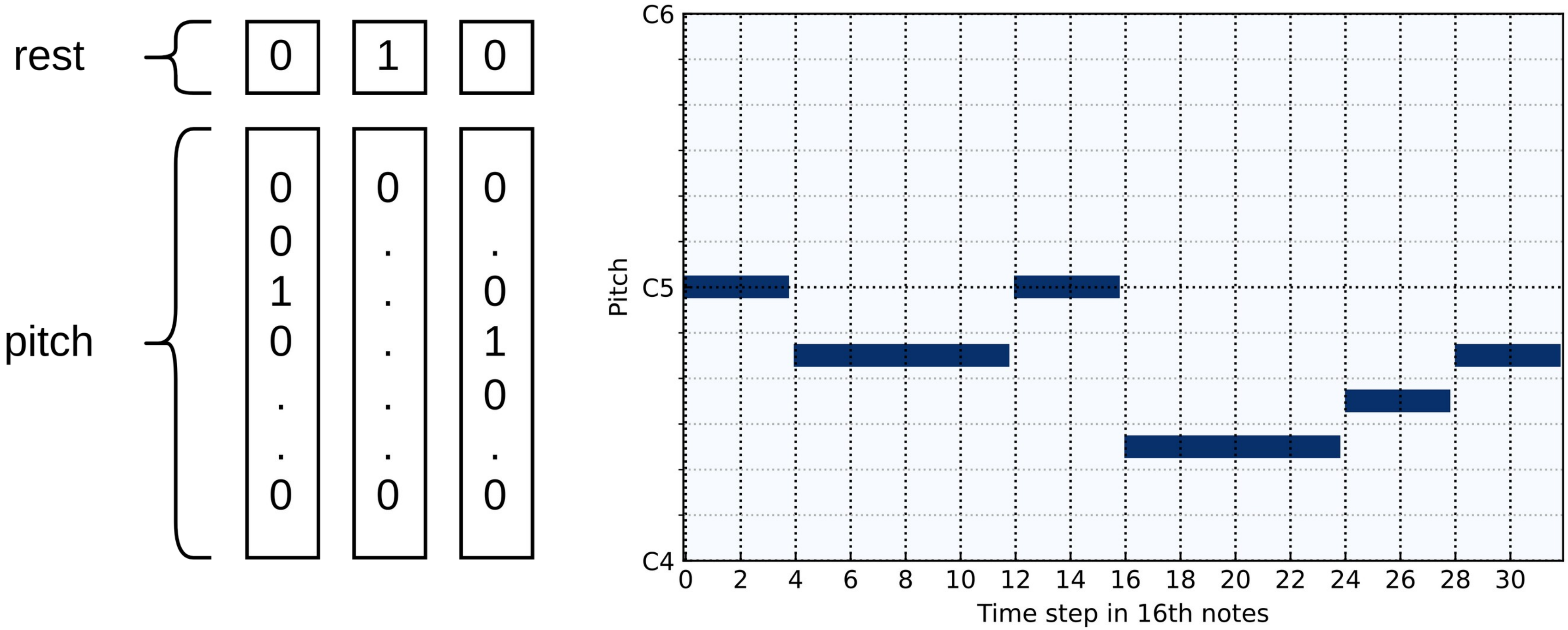
(2) Picture retrieved and changed July 17, 2023 from https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-unrolled.png

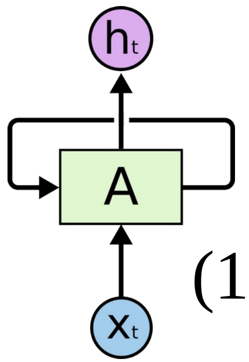
(3) Picture retrieved and changed July 17, 2023 from https://clinicadl.readthedocs.io/en/latest/images/transfer_learning.png

(4) Picture retrieved and changed July 17, 2023 from https://magenta.tensorflow.org/assets/music_vae/architecture.png

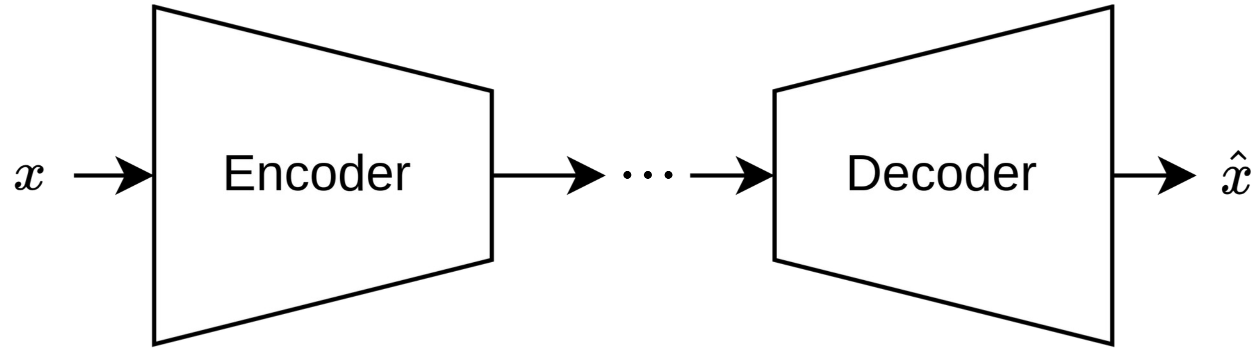
MusicVAE

Representation

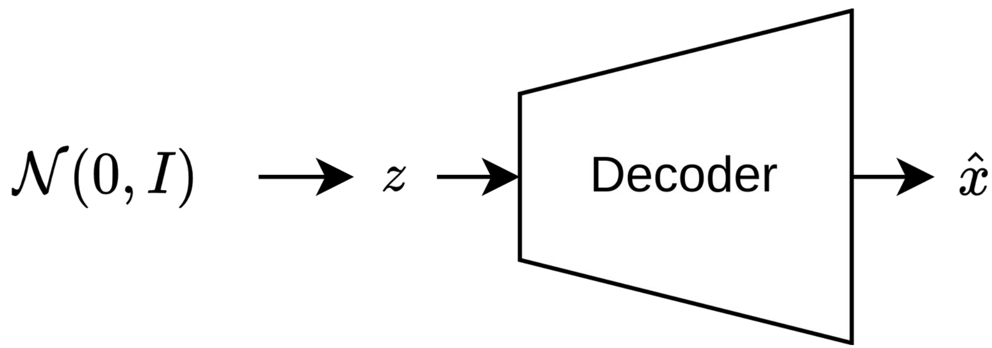




(1)

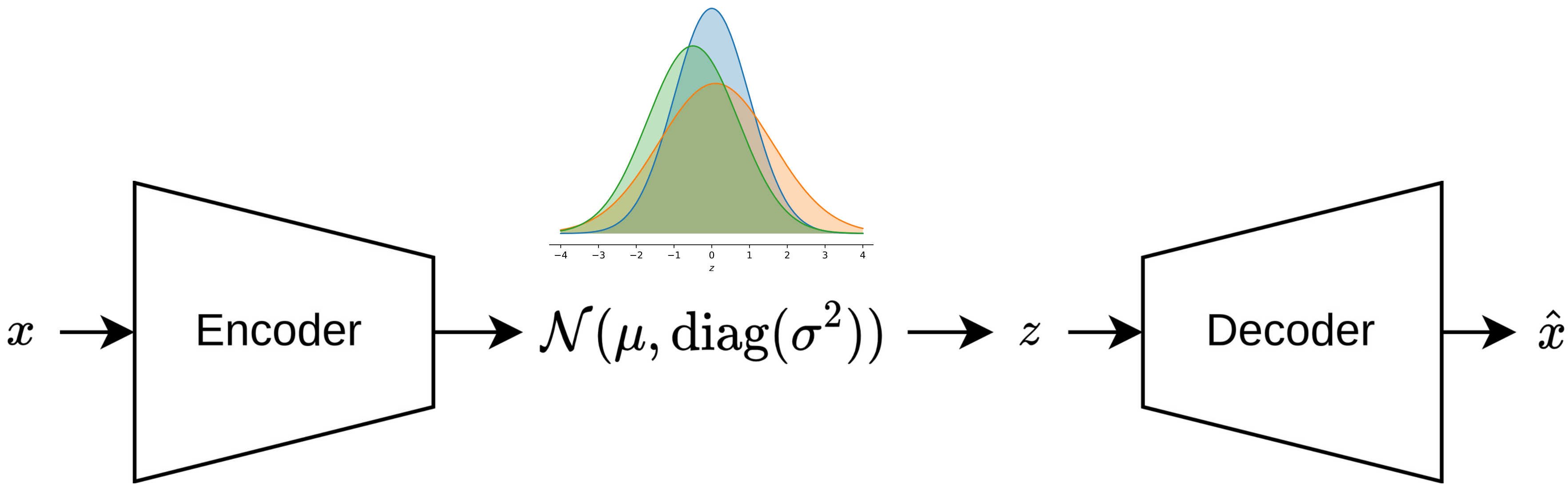


Train a model to generate music **from a random vector**.



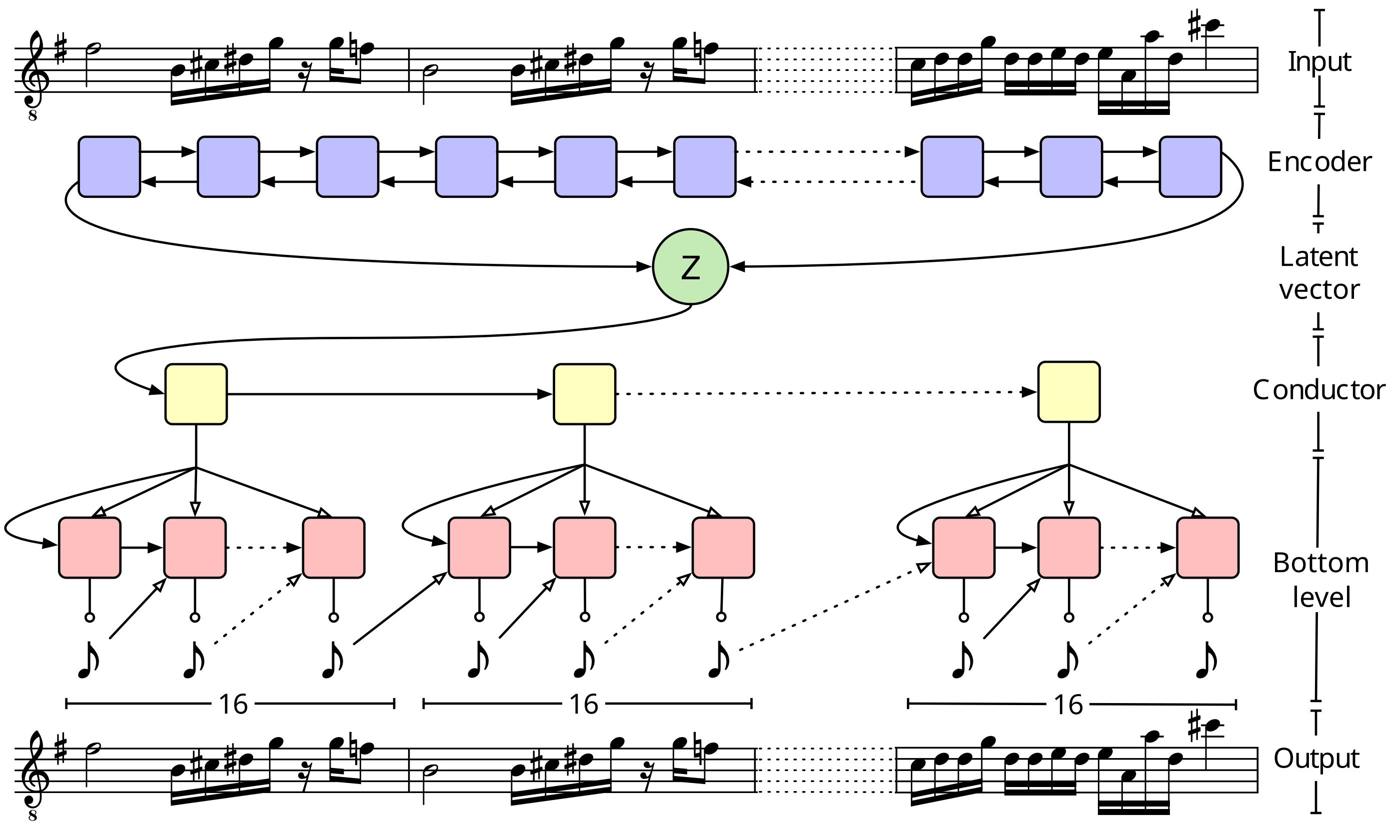
(1) Picture retrieved July 8, 2023 from https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-rolled.png

MusicVAE: VAEs



$$L(x) = \text{rec. loss} + D_{KL}\left(\underline{\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))} \;||\; \underline{\mathcal{N}(0, I)} \right)$$

MusicVAE


Input
Encoder
z
Latent vector
Conductor
Bottom level
16
16
16
Output

Implementation: Training

TECHNISCHE UNIVERSITÄT ILMENAU

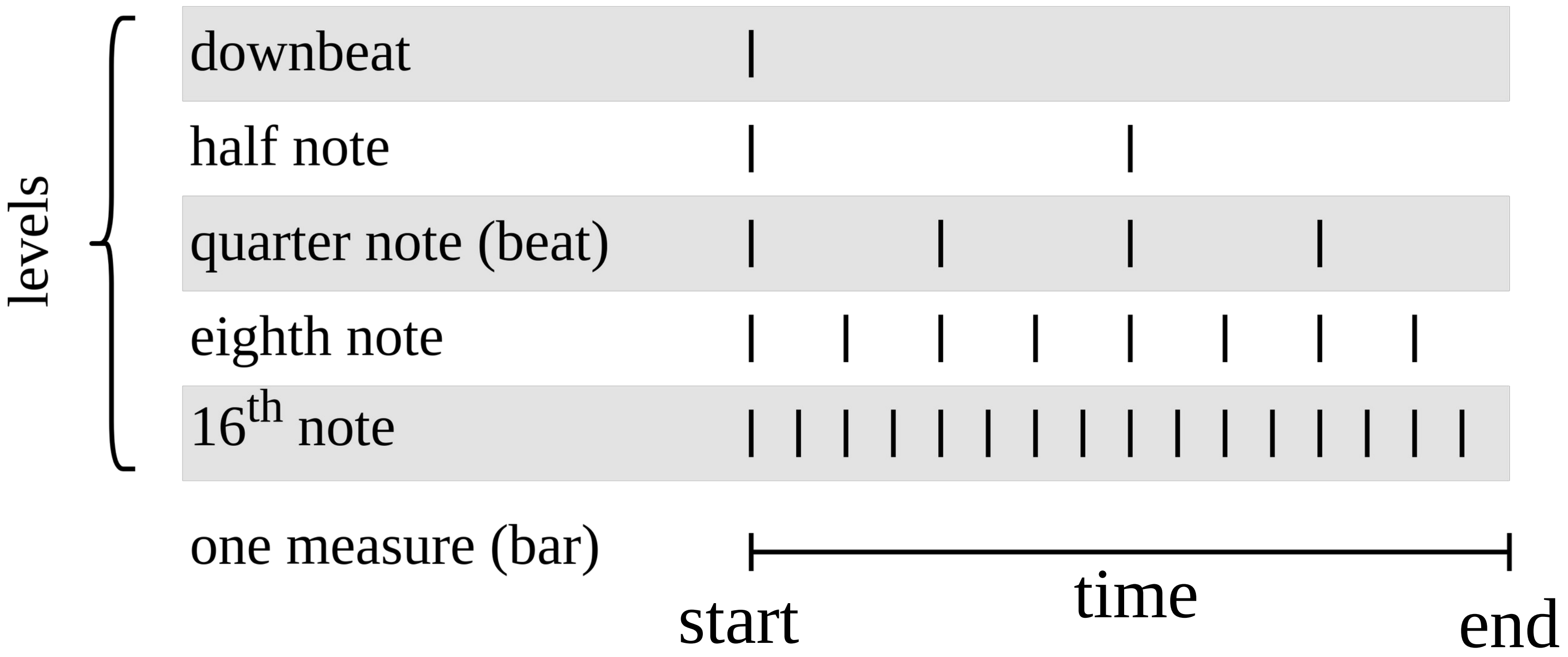
Fraunhofer IDMT

$$L(x) = \text{rec. loss} + \underline{\beta} \max[\, D_{KL}, \, \underline{\lambda} \,]$$

- **β and λ chosen after grid search**
  - $\beta = 1$; $\lambda = 33.3$
- optimizer
  - Adam
  - learning rate (LR) = 10^{-3}
- batch size = 64
- weight decay
  - L_2 regularization with weight 10^{-6}
- LR scheduling
  - customized variant of `ReduceLROnPlateau`
- early stopping was used

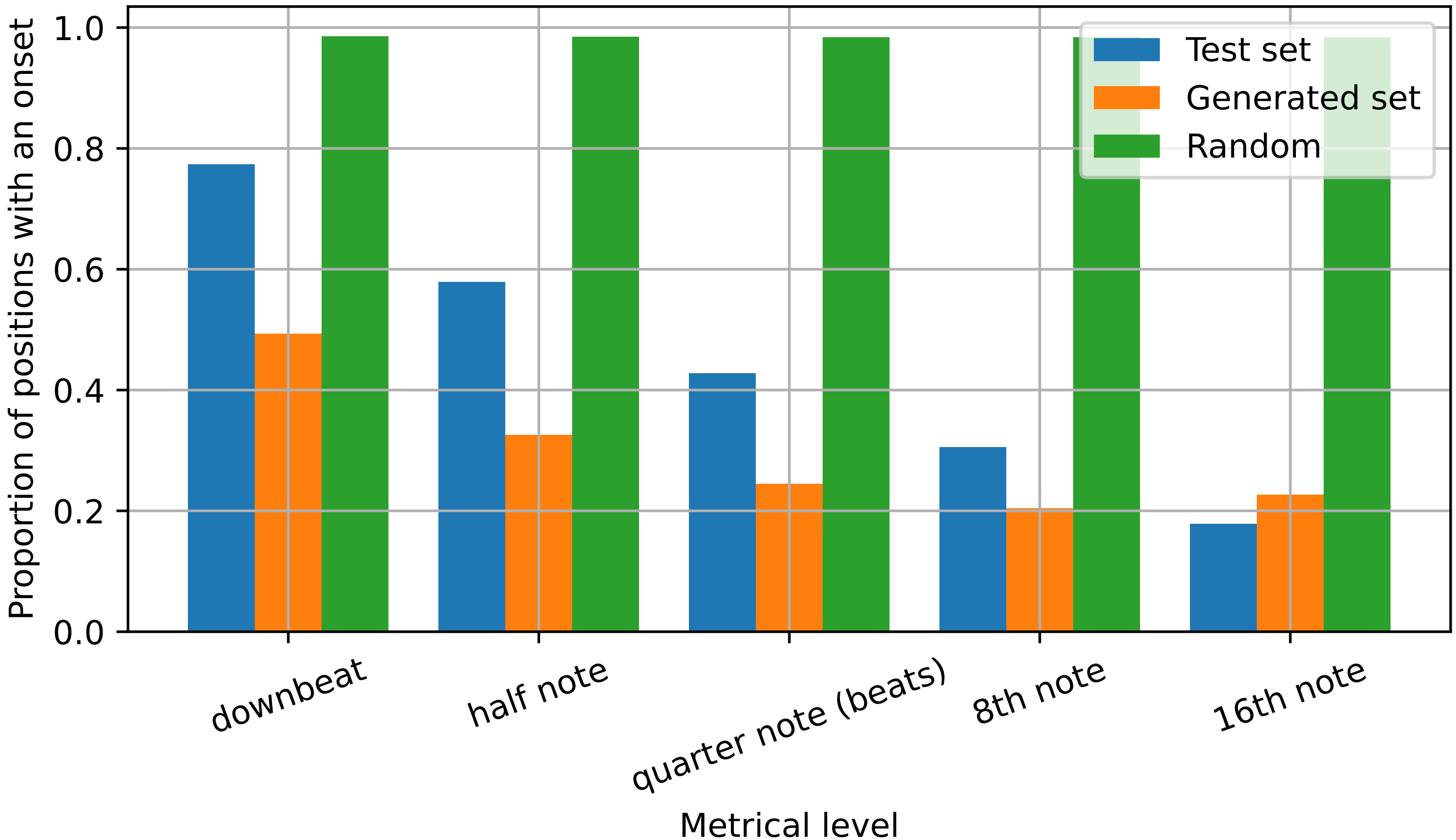
Results: Rhythmic Features

Results: Rhythmic Features



Note onsets can be assigned to a **metrical level**.

Results: Rhythmic Features

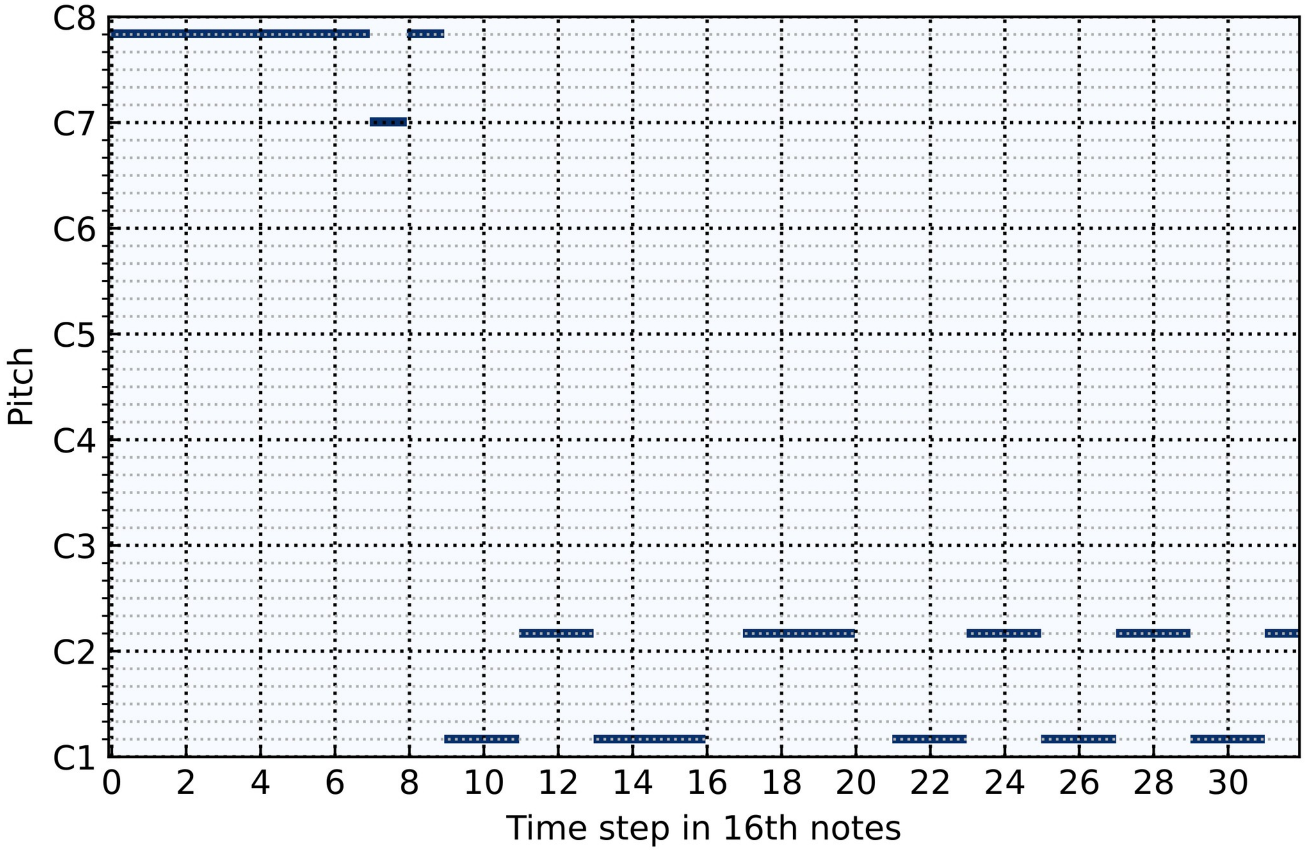


In the generated set there are **more onsets on uneven 16th notes** than on even ones.

Results: Rhythmic Features

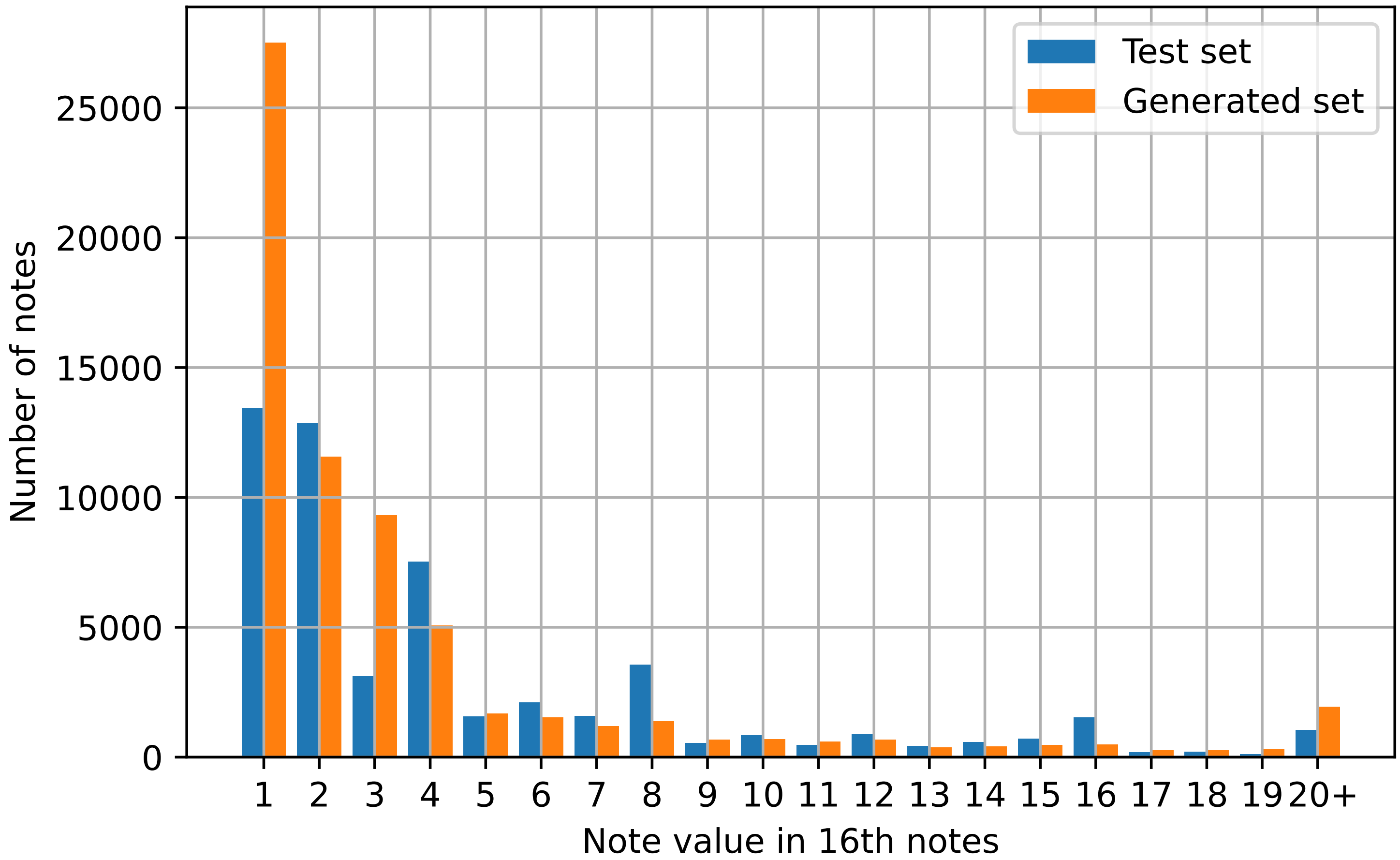
TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



Sequence Nr. 30 (Figure 6.4)

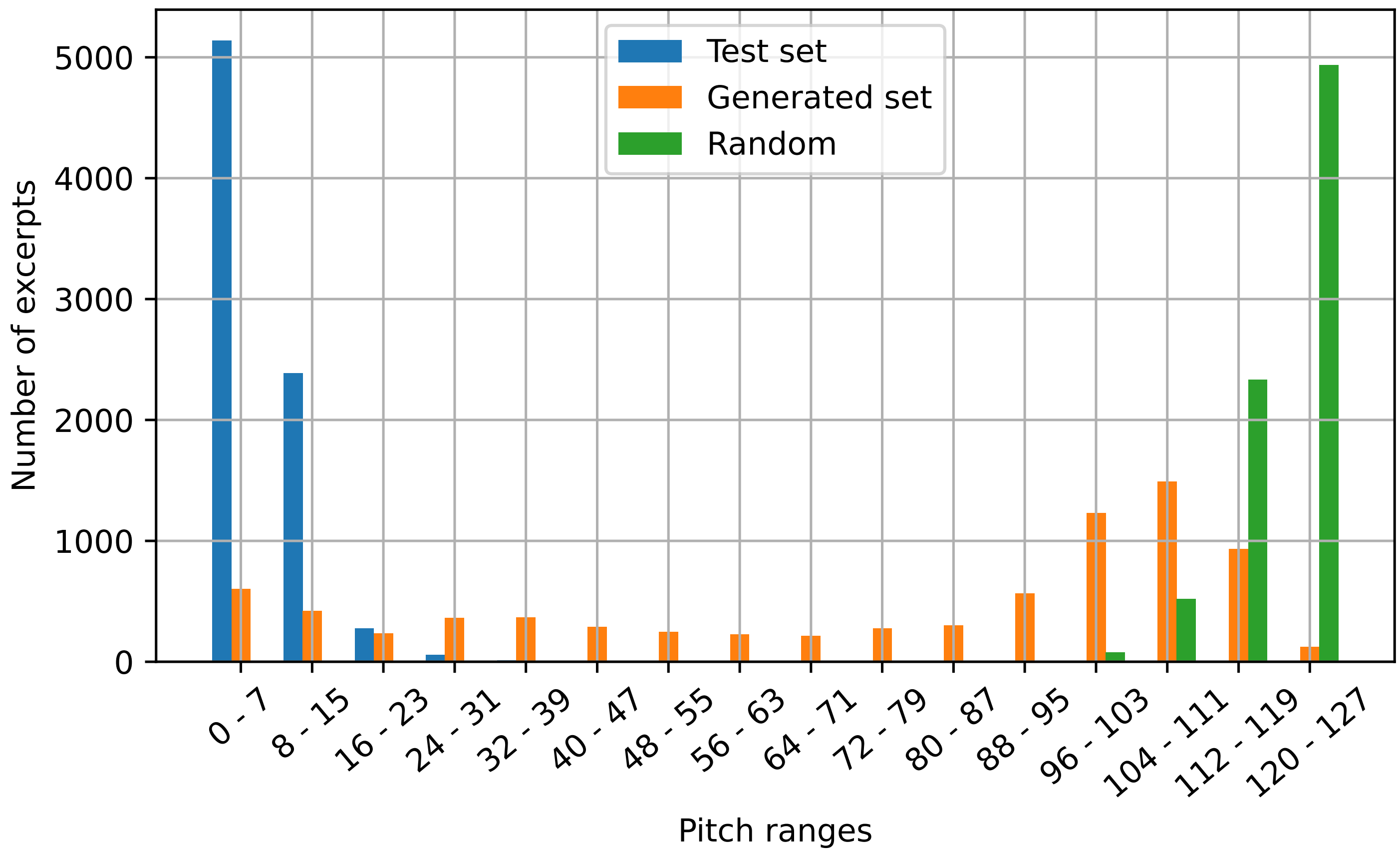
Results: Rhythmic Features



Peaks in the note length were not copied.

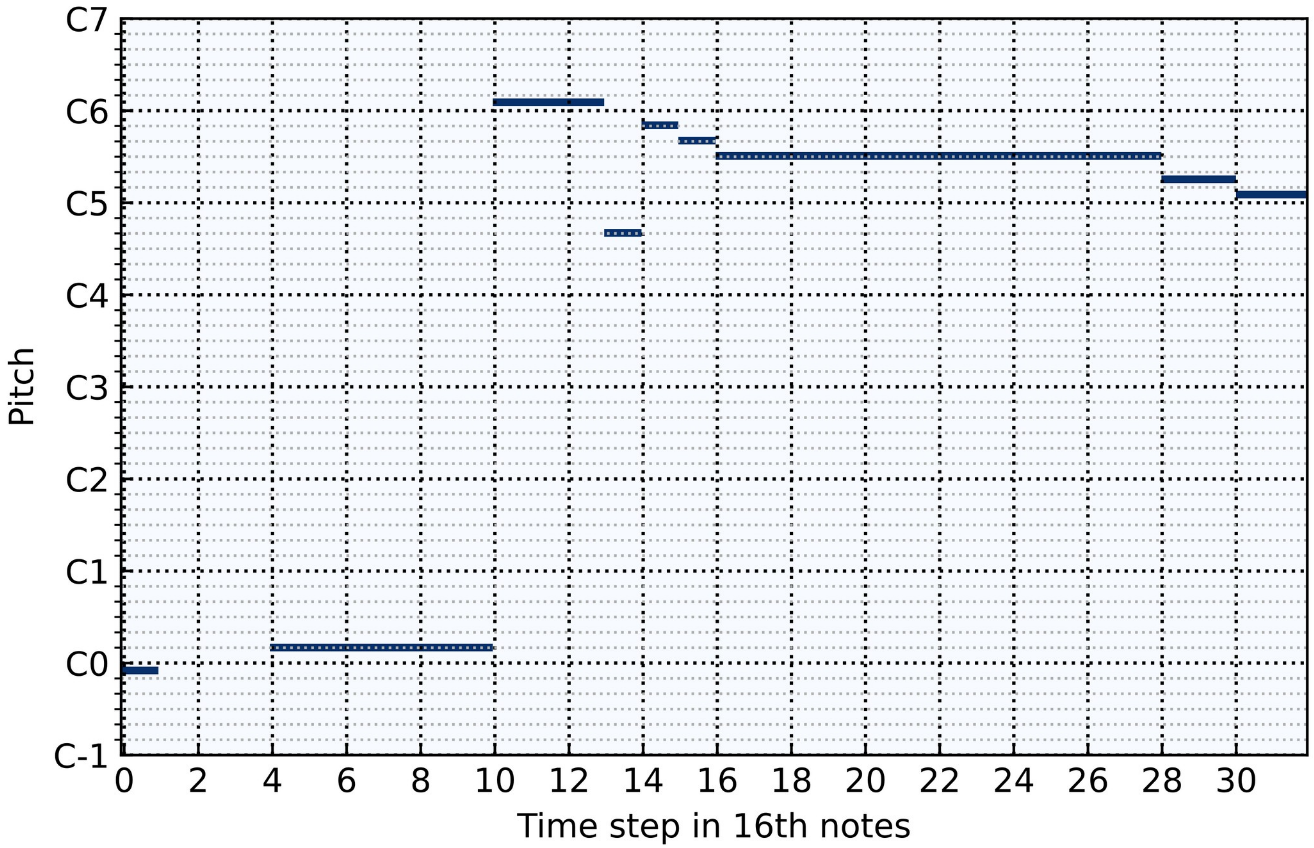
Results:
Melodic Features

Results: Melodic Features



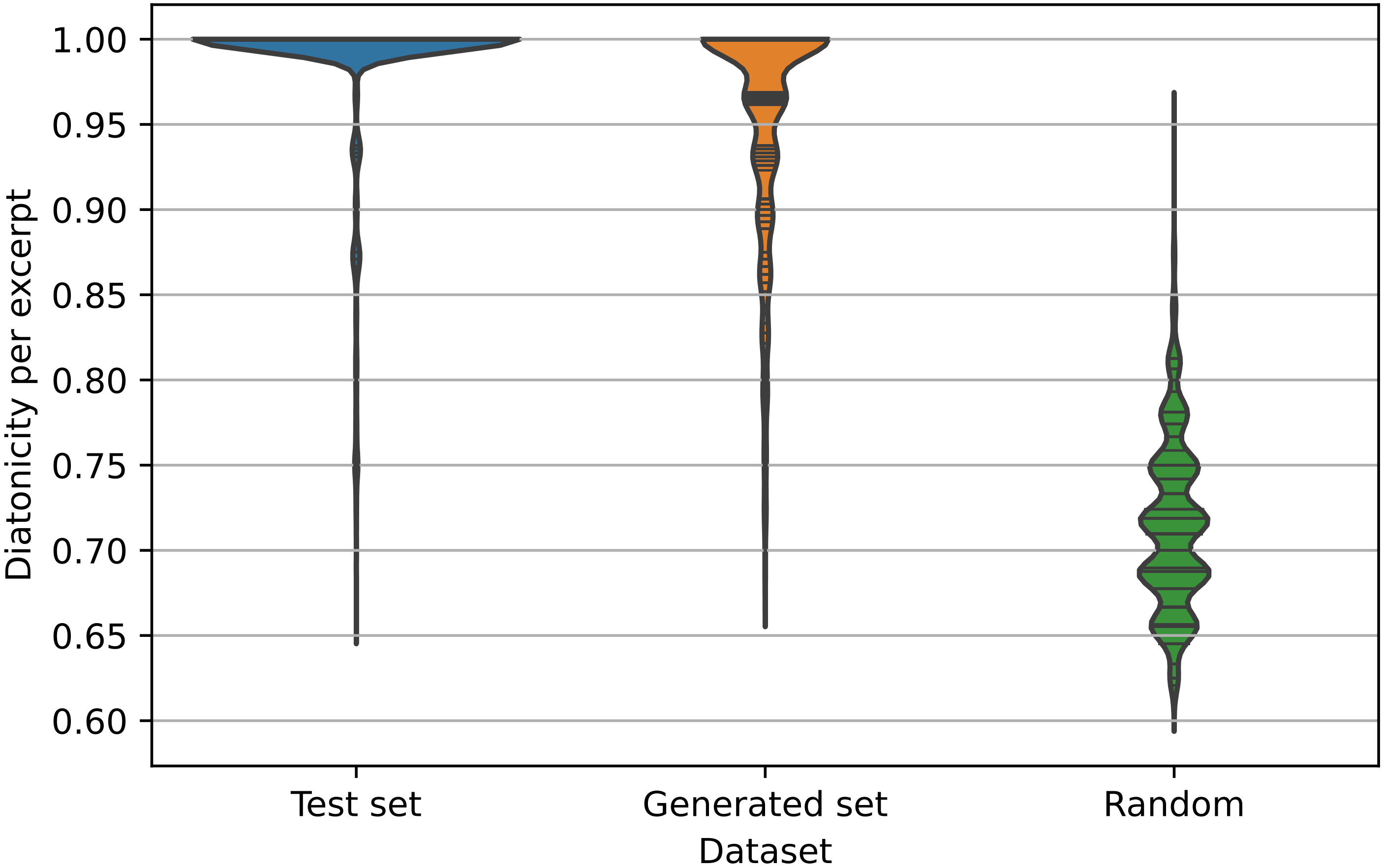
There are **high pitch jumps** in the generated excerpts.

Results: Melodic Features



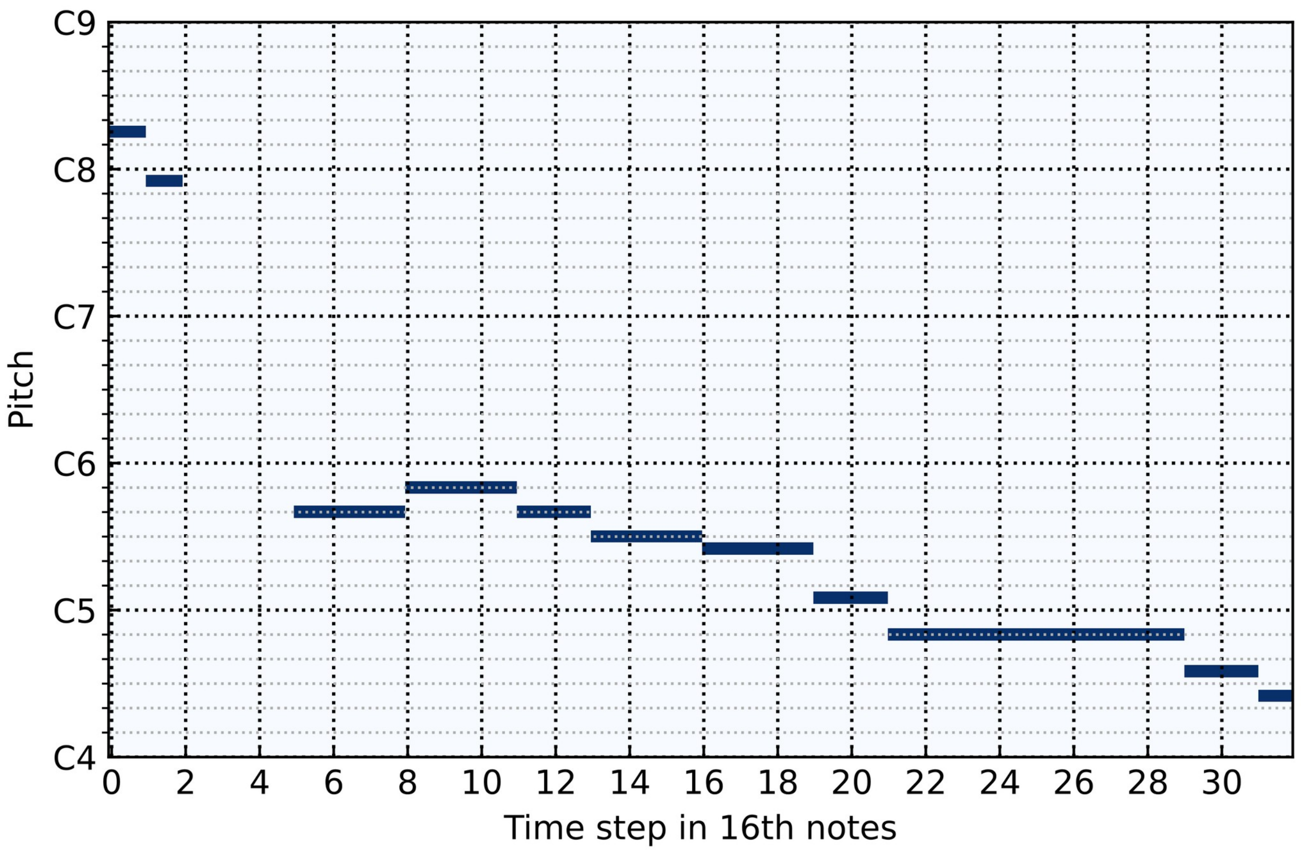
Sequence Nr. 24 (Figure 6.7)

Results: Melodic Features



Generated excerpts are **mostly diatonic**, but there are odd notes.

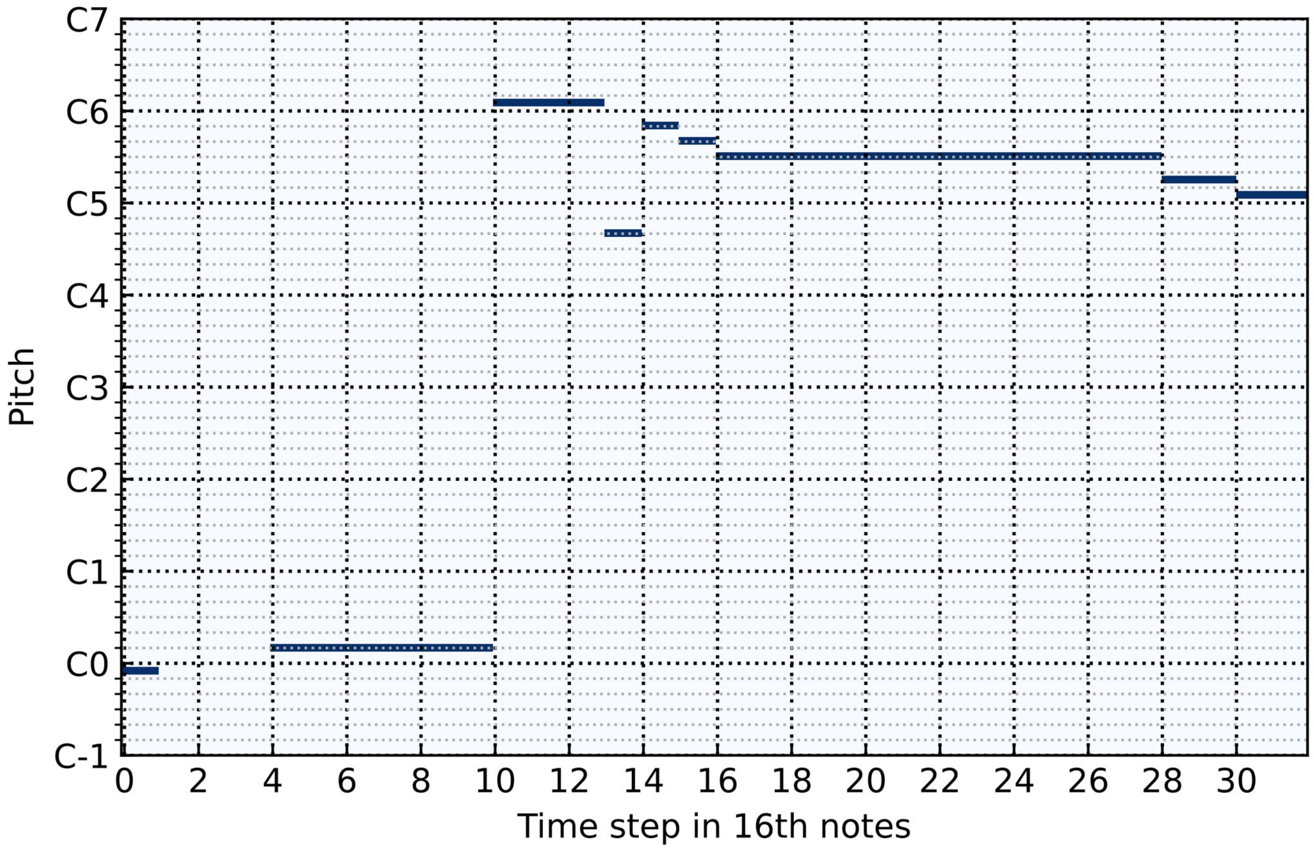
Results: Melodic Features



Sequence Nr. 13 (Figure 6.11)

Results:
Qualitative Evaluation

Results: Melodic Features

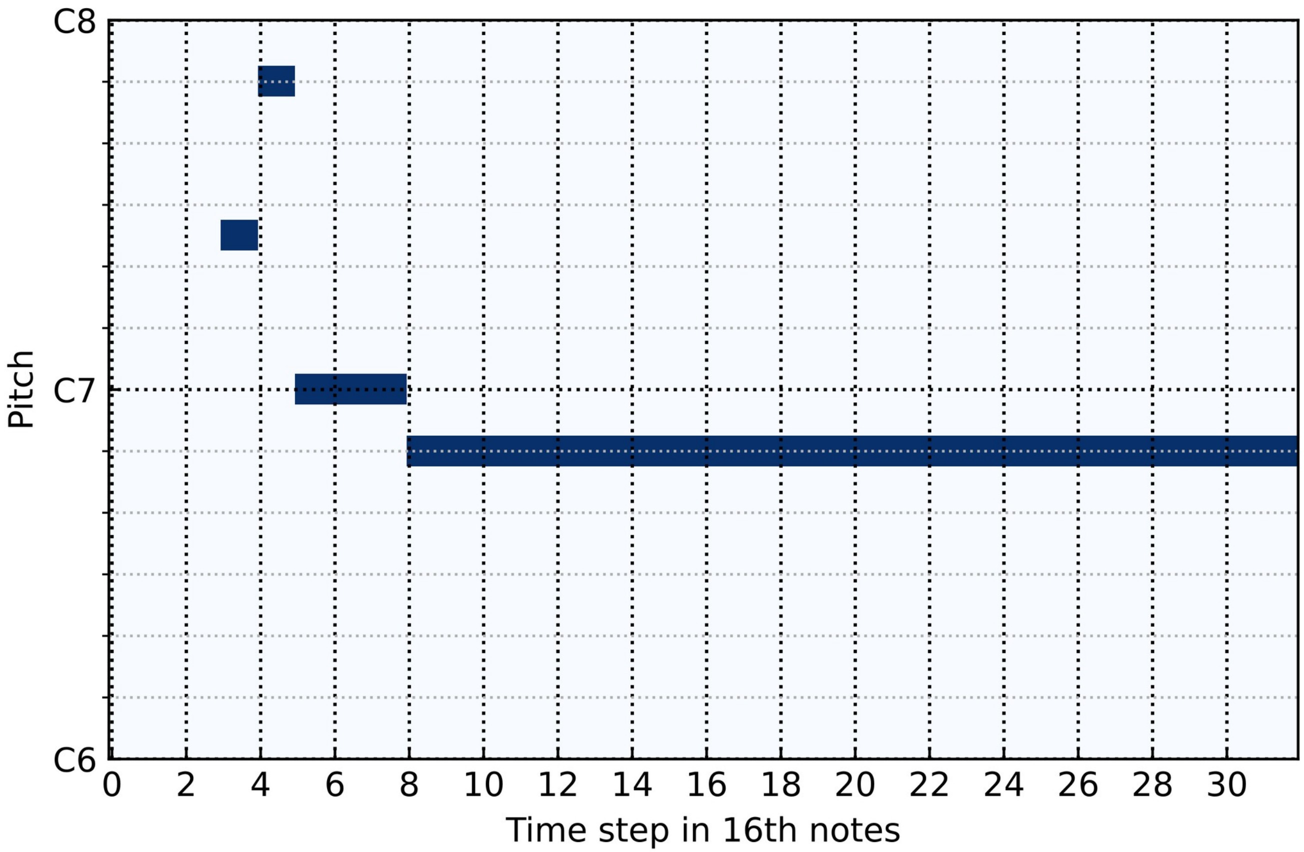


Sequence Nr. 24 (Figure 6.7)

Results: Qualitative Evaluation

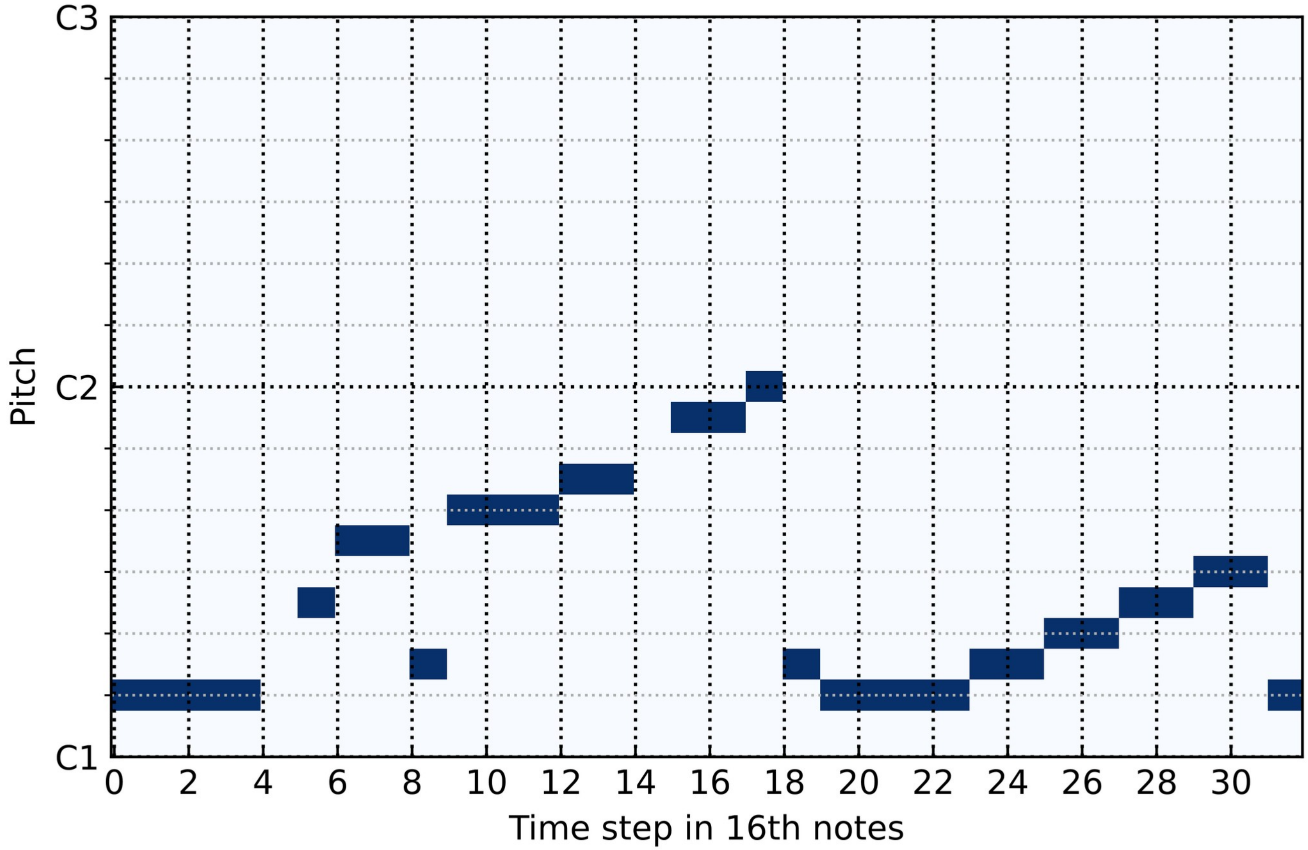
TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



Sequence Nr. 50 (Figure 6.12)

Results: Qualitative Evaluation



Sequence Nr. 34 (Figure 6.15)

Conclusion

Conclusion: Objectives

- state of the art has been reviewed
- re-implementation
  - MusicVAE's flat variant implemented
  - excerpts generated
- generated excerpts were evaluated

Conclusion: Findings

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

- more **onsets on uneven 16th notes**
- single **non-diatonic notes** & **pitch jumps**
- mostly musically coherent & **pleasant-sounding**

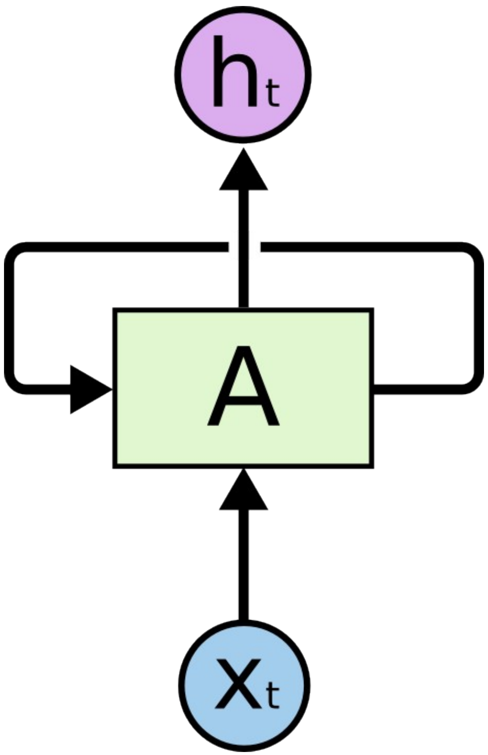
Conclusion: Future Work

- recreate training **dataset**
- adjust training **procedure**
- **extend** model
  - hierarcical, polyphonic, conditioned

Thank you!

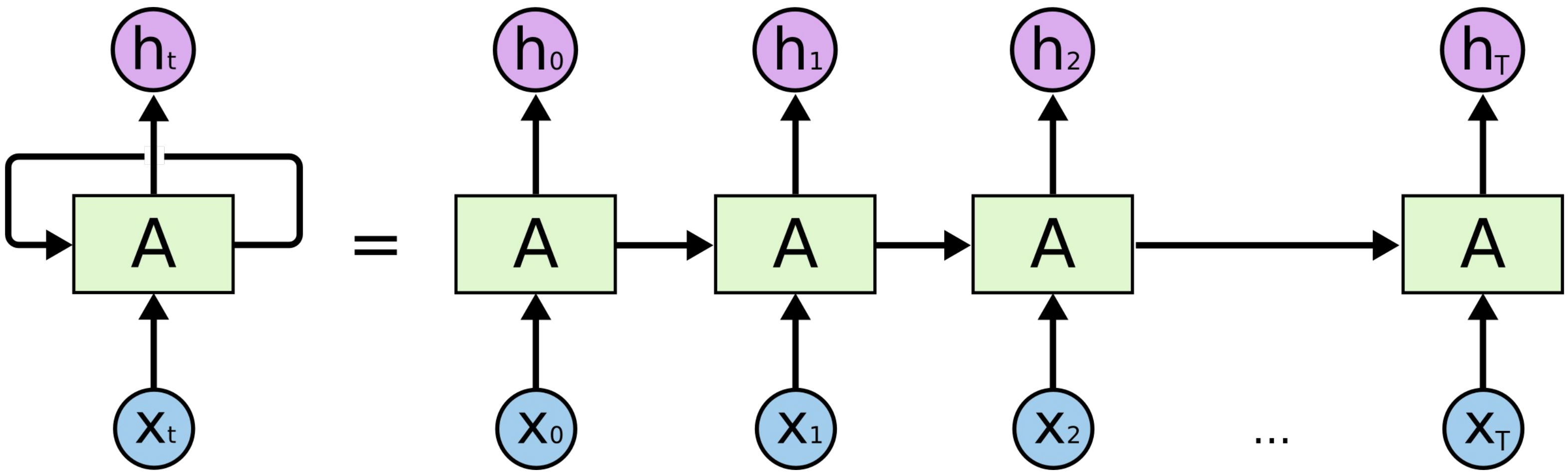
Recurrent Neural Networks

RNNs are neural networks with feedback.



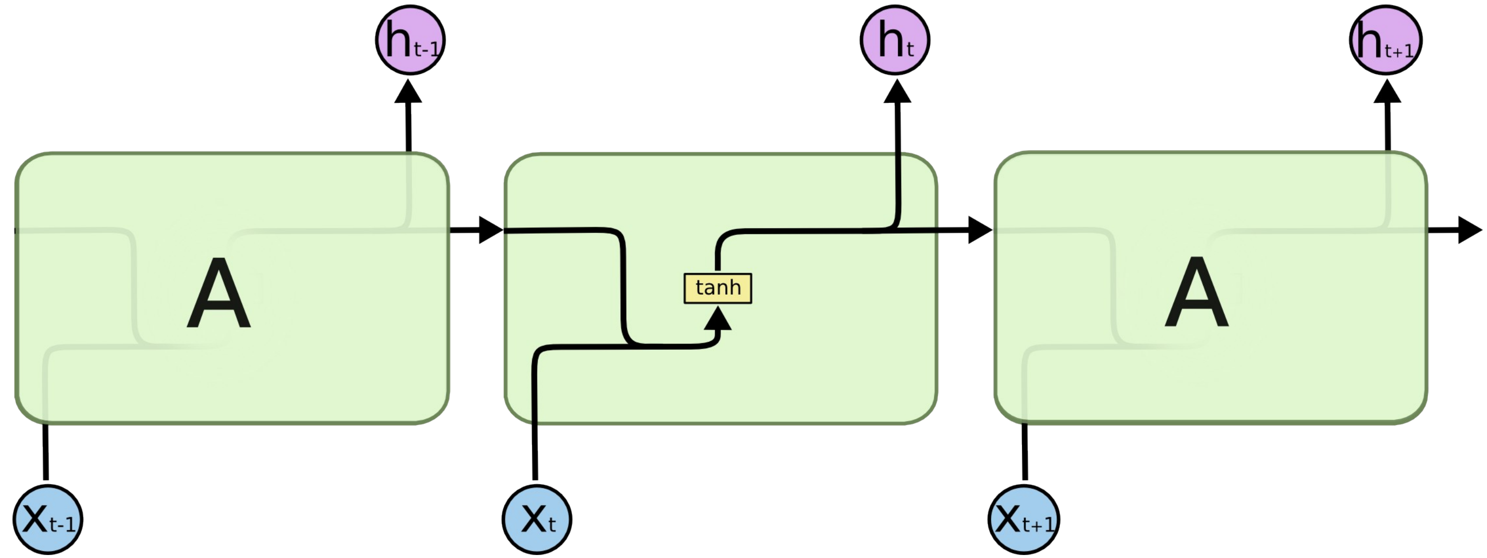
Picture retrieved July 8, 2023 from
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-rolled.png

RNNs can be represented unrolled.



Picture retrieved and changed July 8, 2023 from
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-unrolled.png

MusicVAE: RNNs

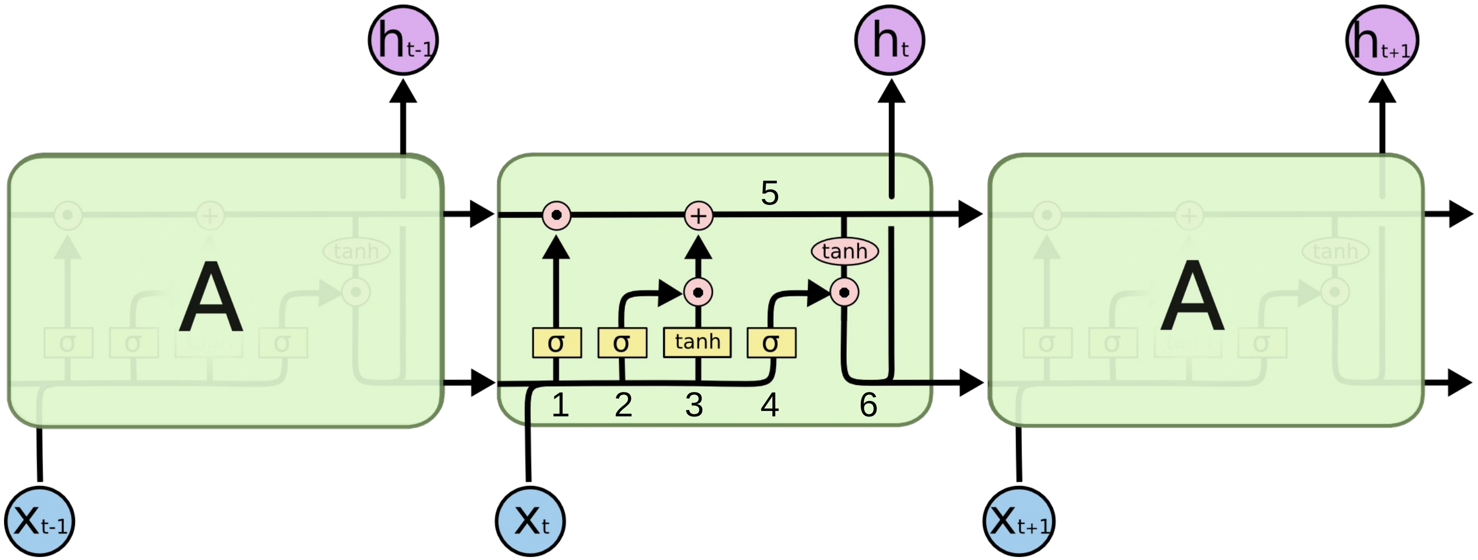


$$h_t = \tanh(W_i x_t + b_i + W_h h_{t-1} + b_h)$$

Standart RNNs have some drawbacks.

Picture retrieved July 8, 2023 from
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-SimpleRNN.png

TECHNISCHE UNIVERSITÄT ILMENAU Fraunhofer IDMT



$$1: \; i_t = \sigma(\underline{W_{ii}x_t + b_{ii}} + \underline{W_{hi}h_{t-1} + b_{hi}}) \qquad \sigma(x) = \mathrm{sigmoid}(x)$$

$$2: \; f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf})$$

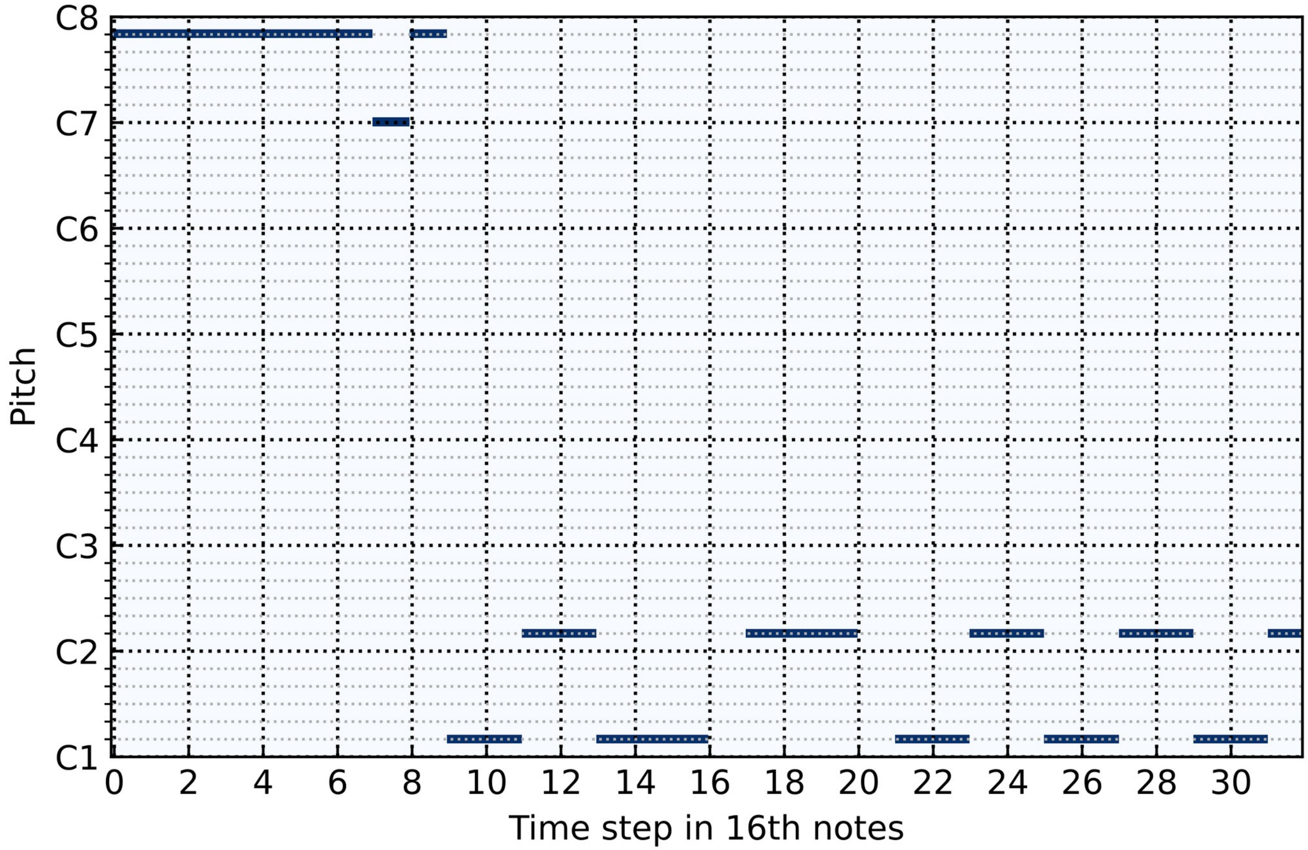
$$3: \; g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \qquad 5: \; c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$4: \; o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \qquad 6: \; h_t = o_t \odot \tanh(c_t)$$

Picture retrieved and changed July 8, 2023 from
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-chain.png

Generated Excerpts

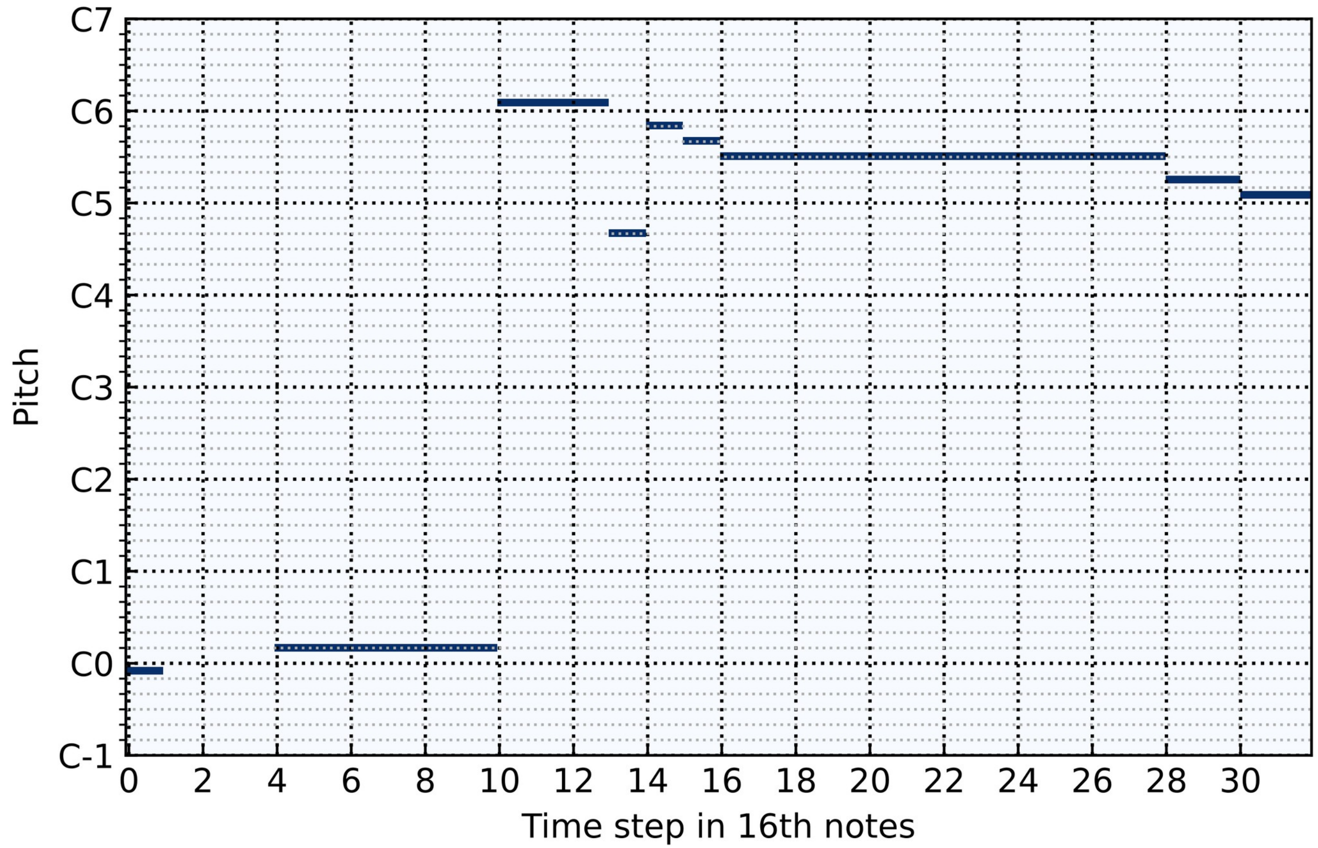
Generated Excerpts



Sequence Nr. 30 (Figure 6.4)

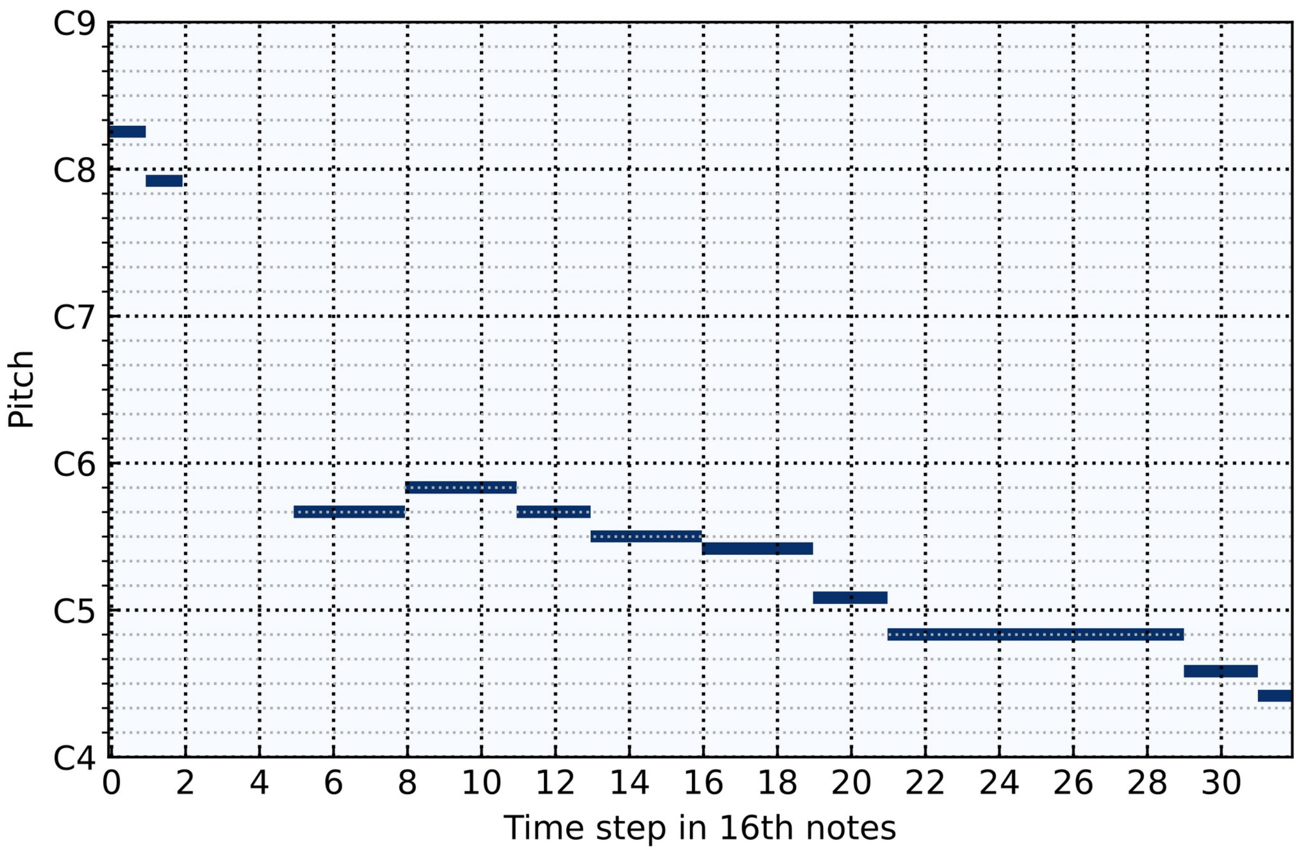
TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



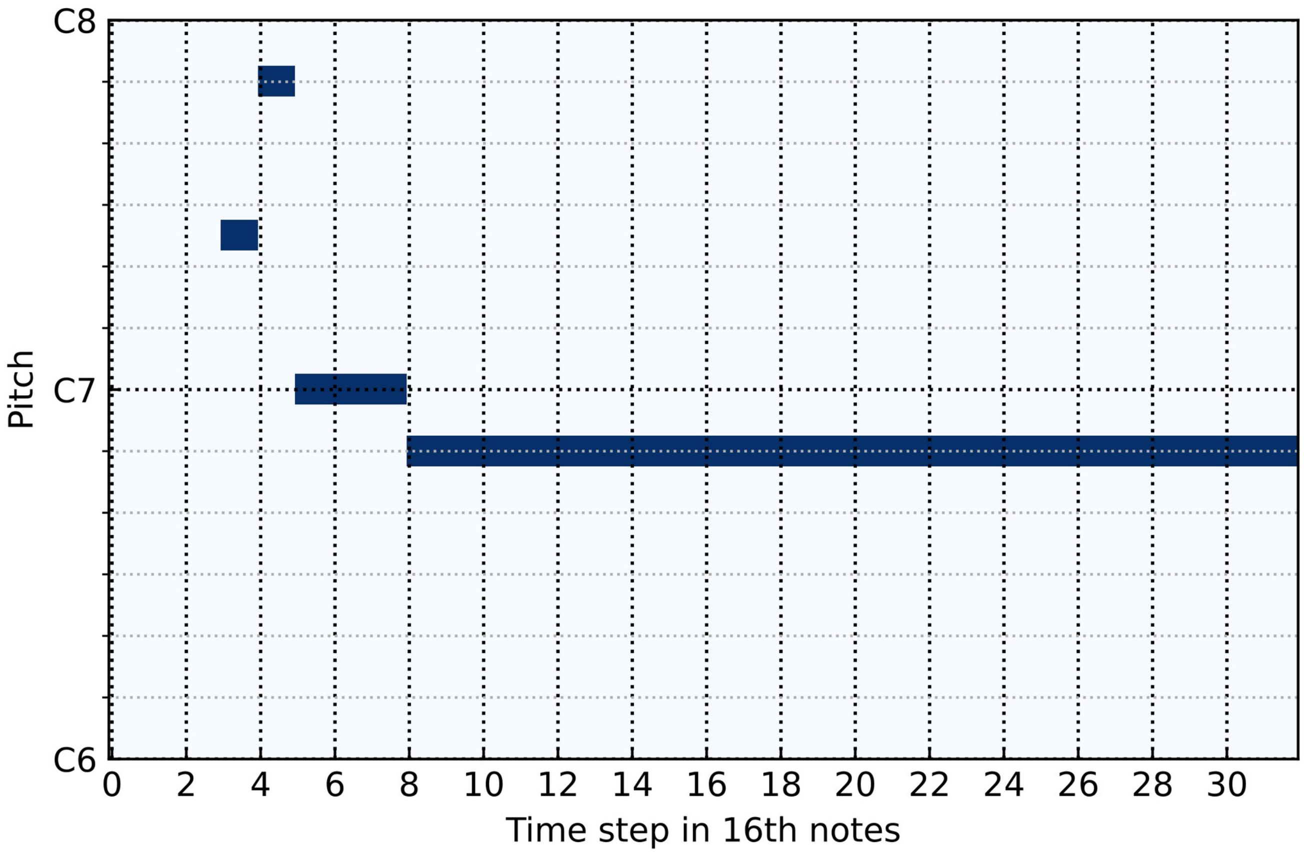
Sequence Nr. 24 (Figure 6.7)

Results: Melodic Features



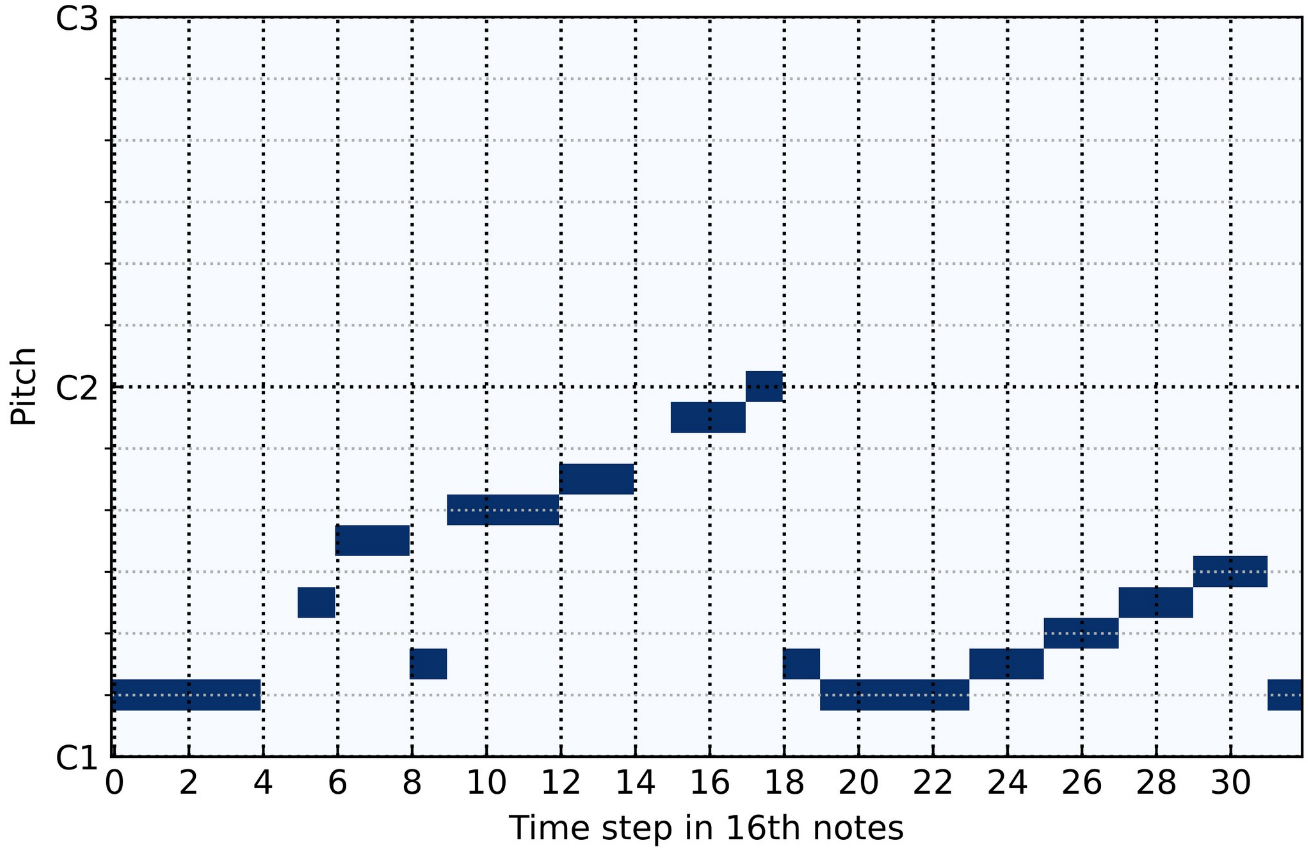
Sequence Nr. 13 (Figure 6.11)

Generated Excerpts



Sequence Nr. 50 (Figure 6.12)

Generated Excerpts



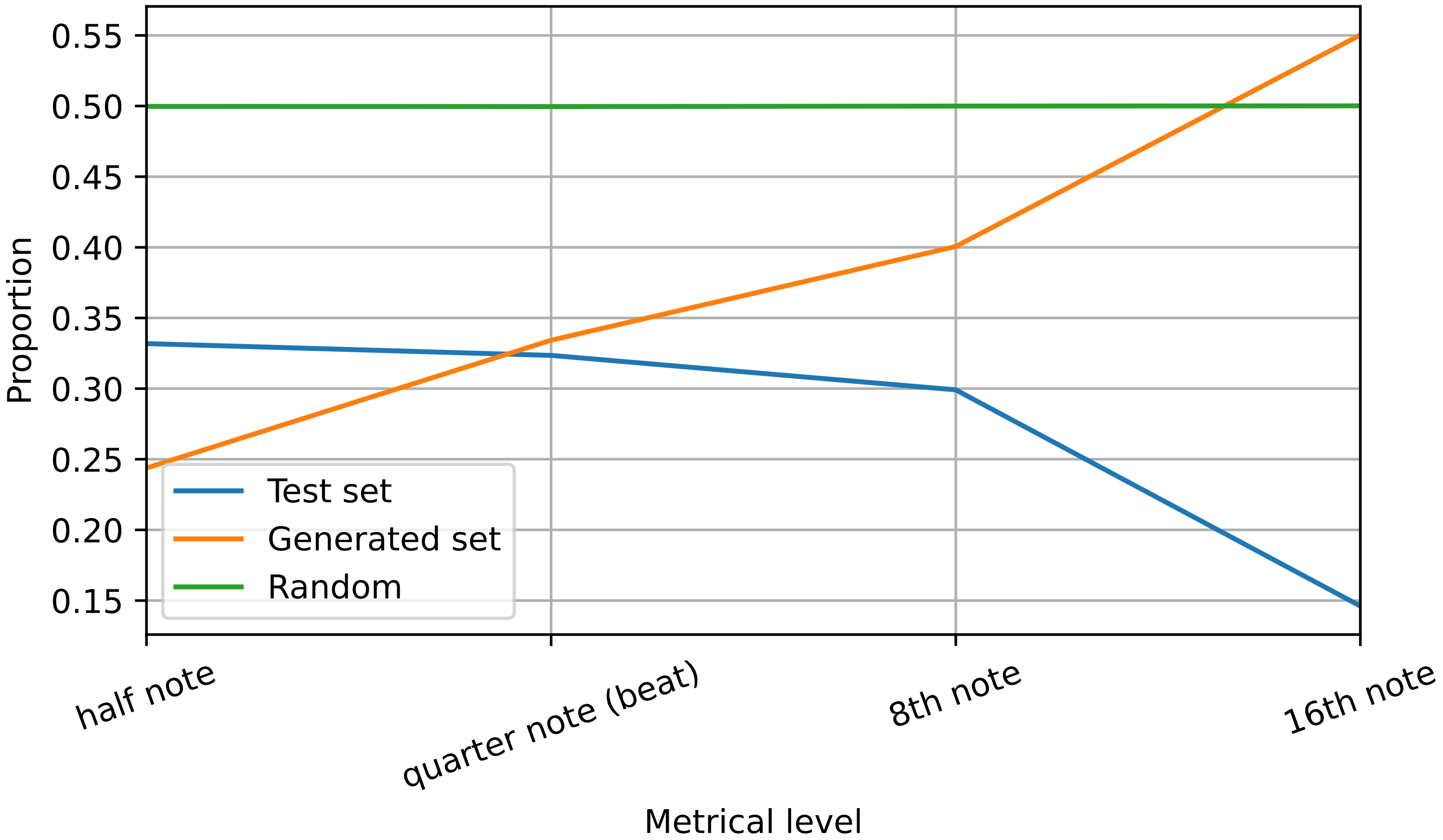
Sequence Nr. 34 (Figure 6.15)

Evaluation

Results: Rhythmic Features

TECHNISCHE UNIVERSITÄT ILMENAU

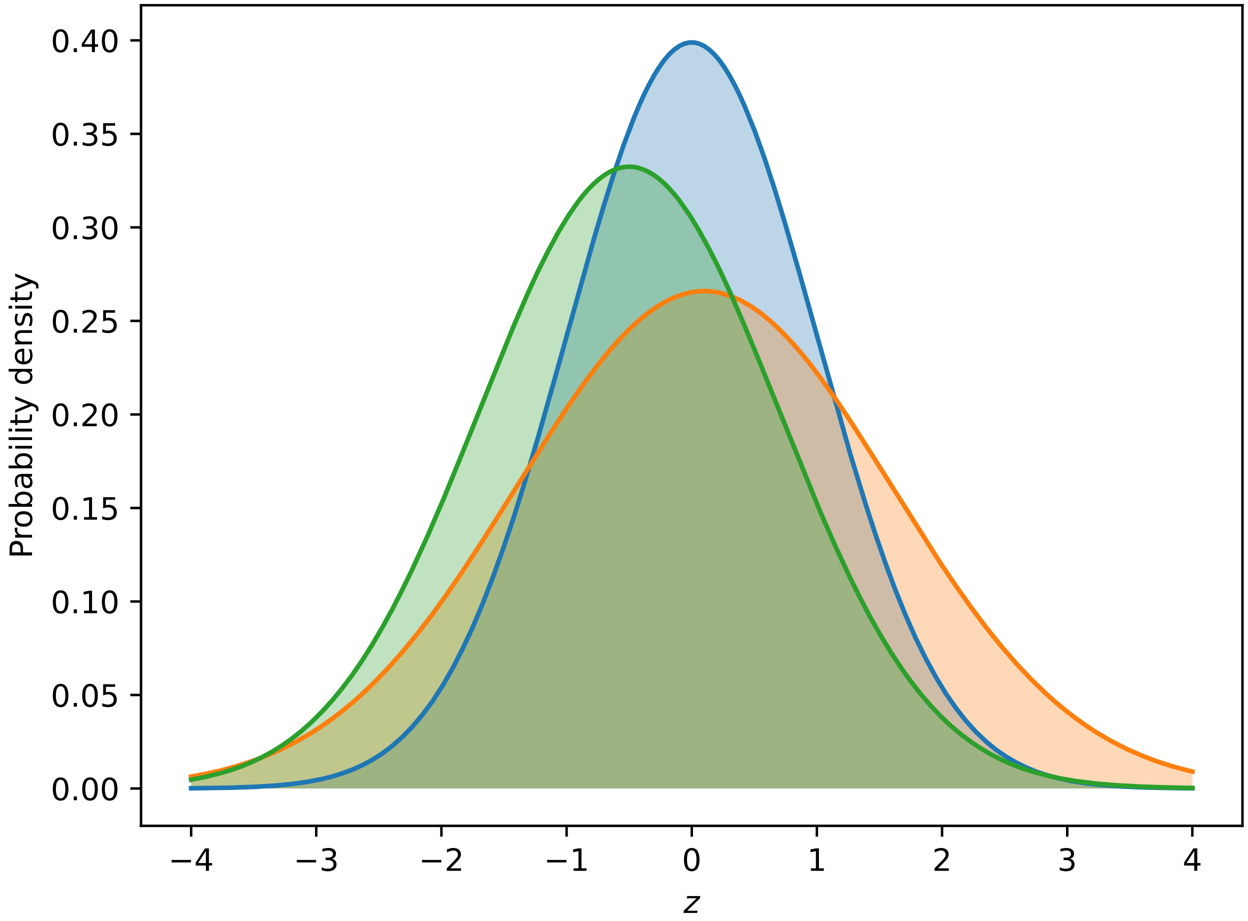
Fraunhofer IDMT



In the generated set there are **more onsets on uneven 16th notes** than on even ones.

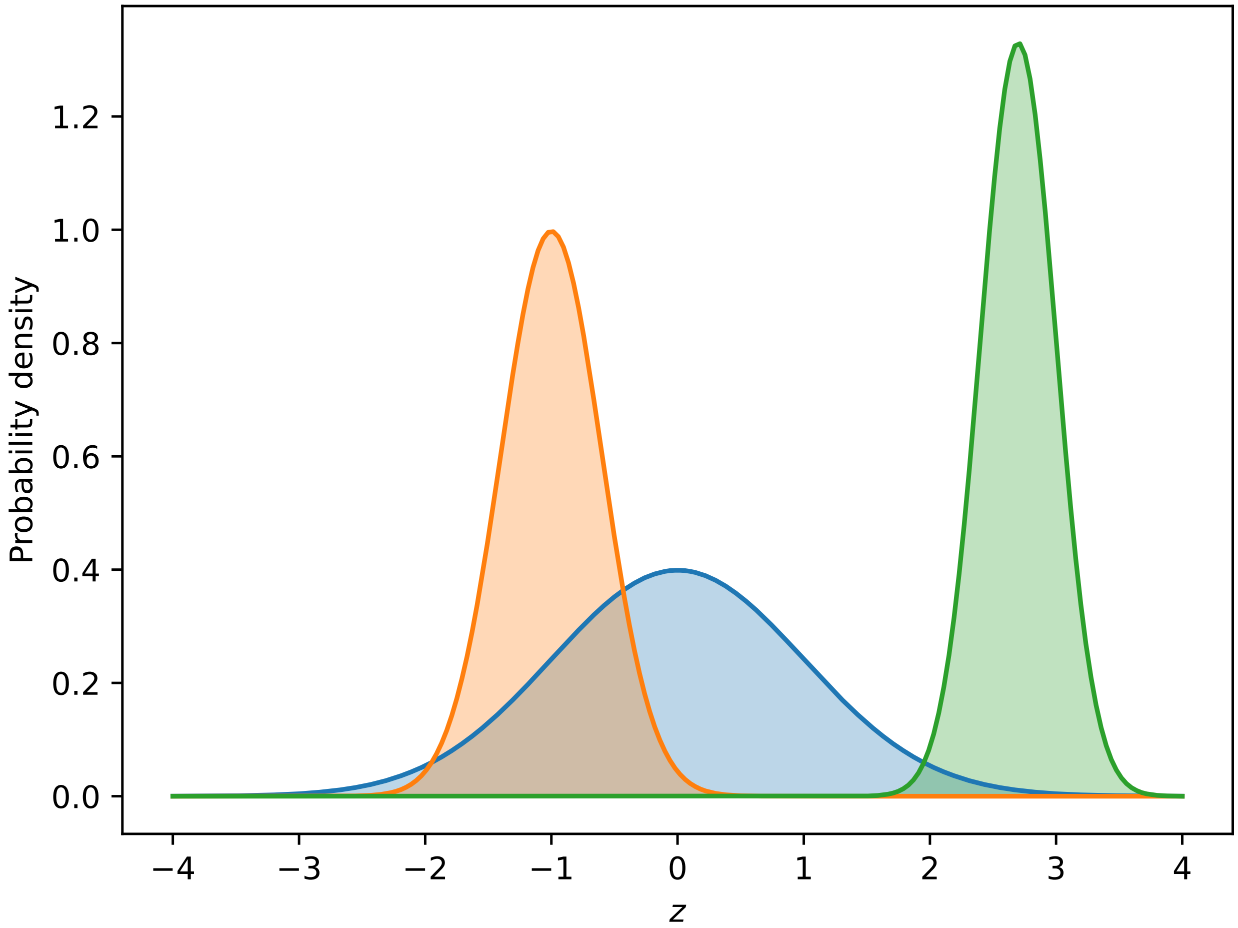
Latent Space

MusicVAE: VAEs



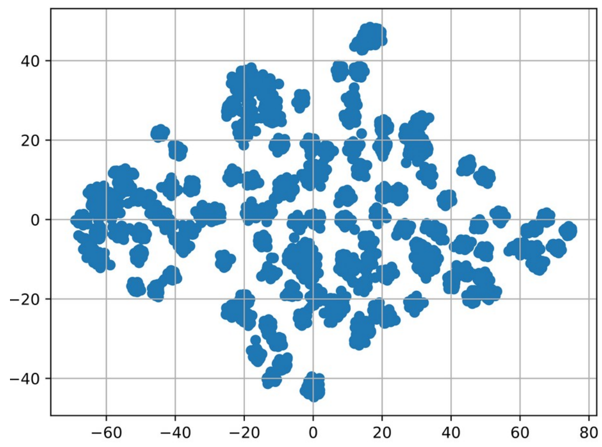
well-formed

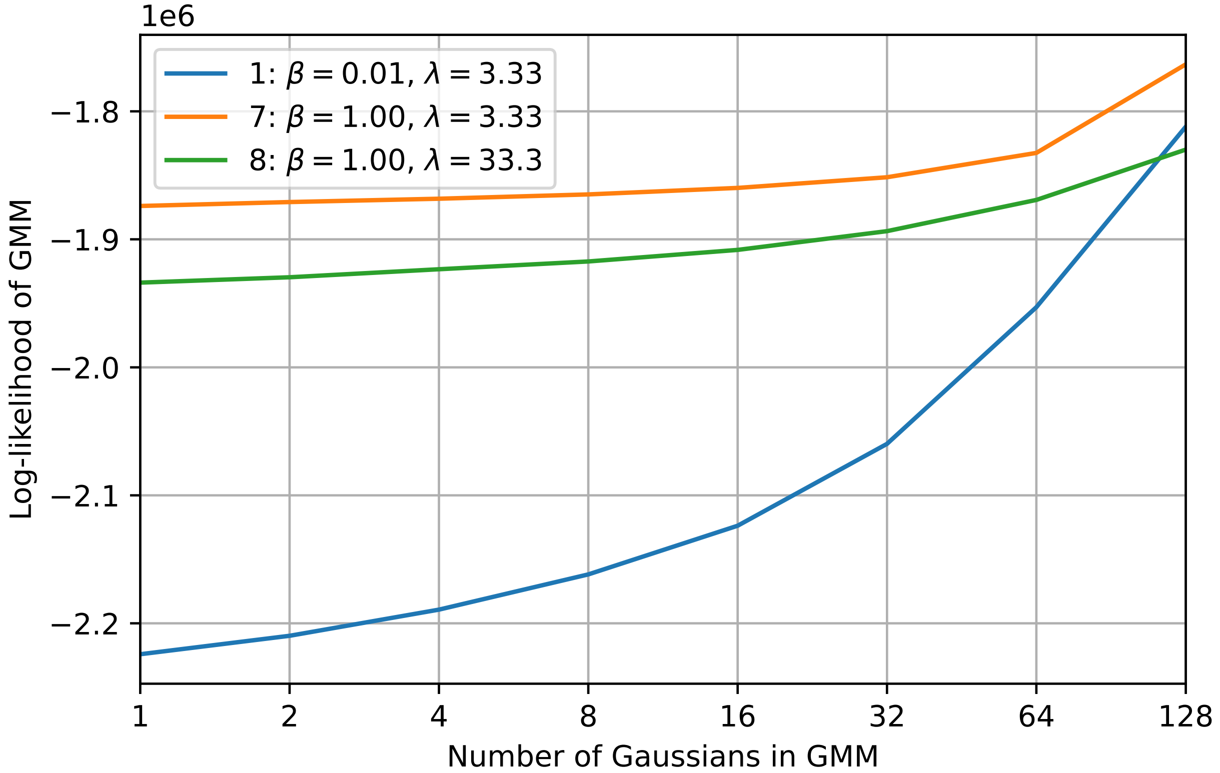
MusicVAE: VAEs



not well-formed

Results: Training





The structure of the latent spaces of Models 1, 7, and 8 was examined.

Generation of Symbolic Music Based on MusicVAE

Jakob Lerch

Supervisors:
Prof. Dr. Shardt
Dr. Andrew McLeod
Dr. Jakob Abeßer

July 27, 2023

Introduction

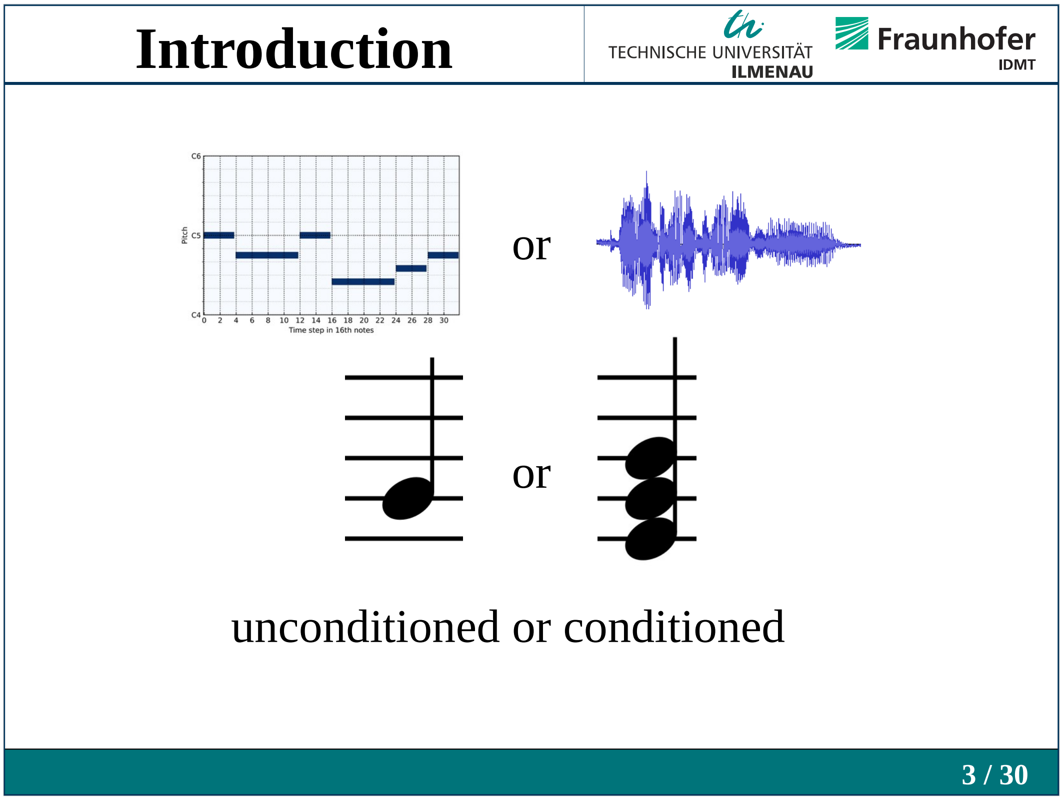
TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

Music generation could be **supportive for composition or live performances**.

Picture retrieved July 16, 2023 from
https://pixabay.com/photos/music-producer-studio-actor-audio-4507819/

- **What is music generation?**
- **why do we care?**
  - create music for creative productions
  - support during composition or live performances



- **symbolic vs. waveform**
- **monophonic vs. polyphonic**
- **conditioned vs. unconditioned**
- **Western music**
- **objectives**
  - summarize state of the art
  - re-implement MusicVAE
    - training data: two-bar monophonic sequences
  - generate excerpts and discuss their quality

Introduction: Objectives

(1) review **state of the art**

(2) **re-implement** MusicVAE

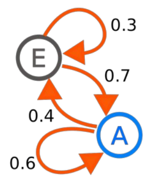
(3) **evaluate quality** of generated excerpts

State of the Art

State of the Art

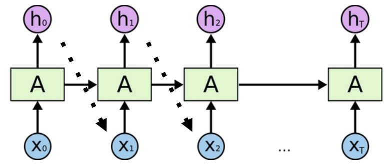
TECHNISCHE UNIVERSITÄT ILMENAU

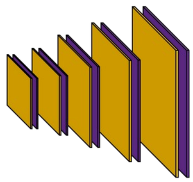
Fraunhofer IDMT



(1)

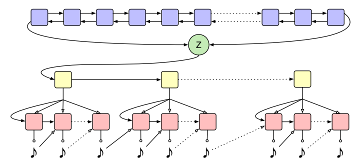
(2)





(3)

(4)



(1) Picture retrieved July 17, 2023 from https://en.wikipedia.org/wiki/Markov_chain#/media/File:Markovkate_01.svg

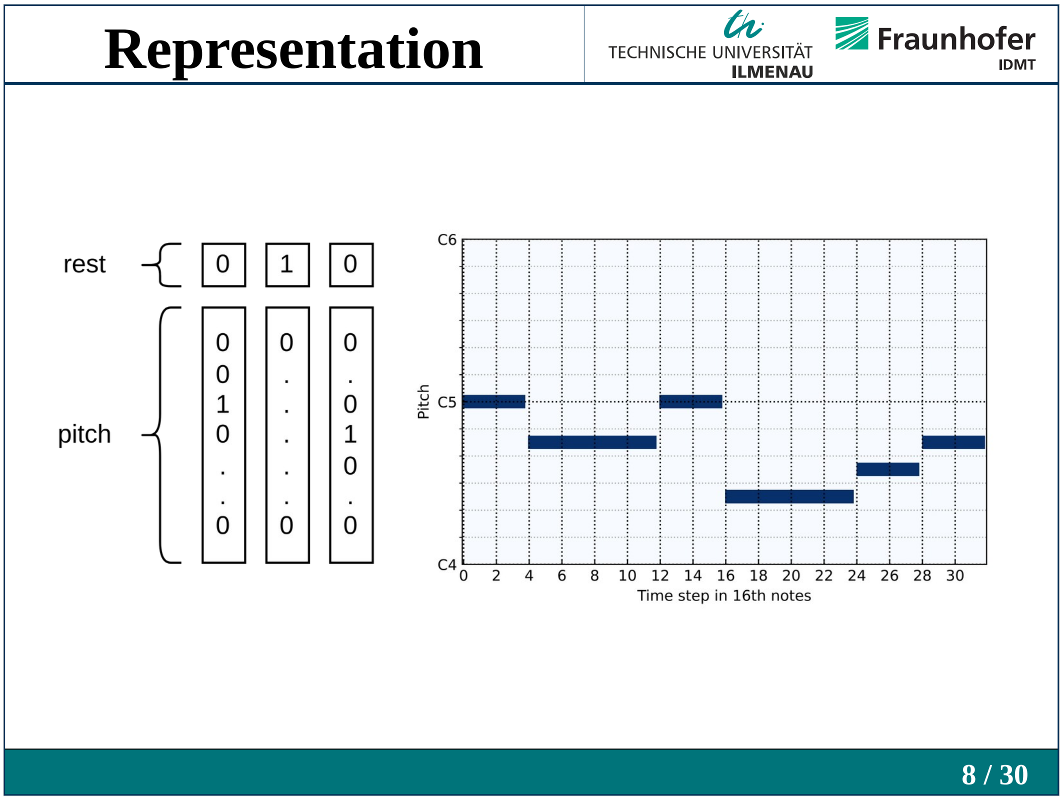
(2) Picture retrieved and changed July 17, 2023 from https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-unrolled.png

(3) Picture retrieved and changed July 17, 2023 from https://clinicadl.readthedocs.io/en/latest/images/transfer_learning.png

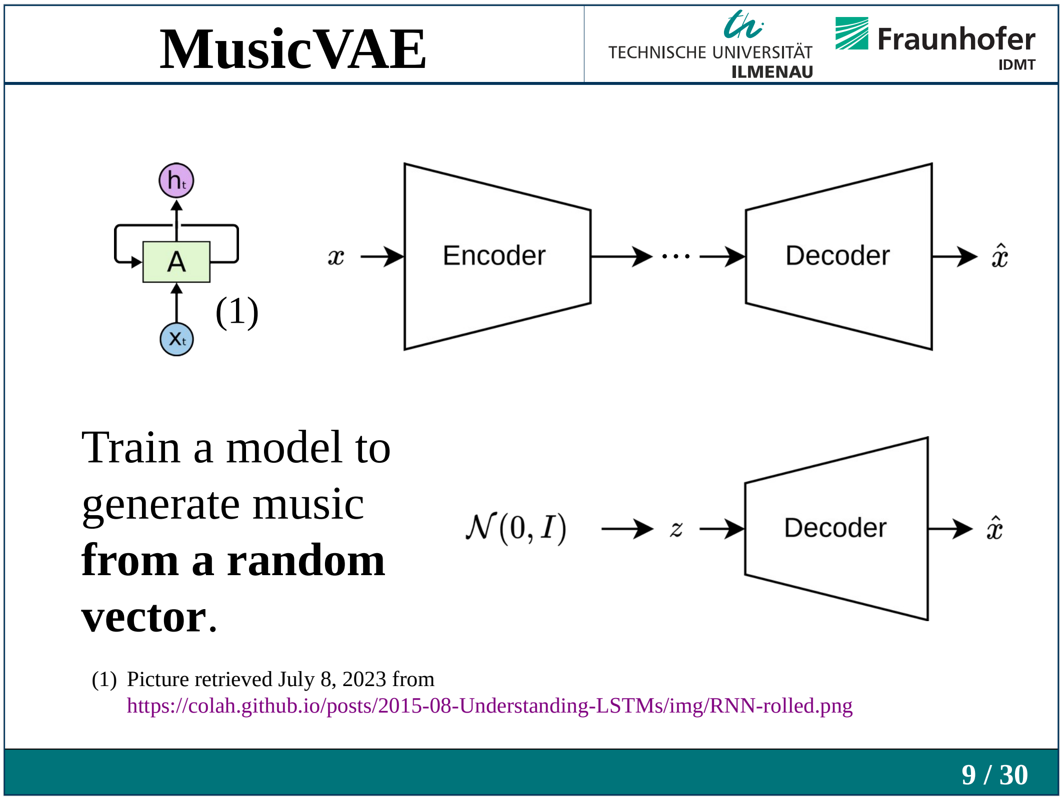
(4) Picture retrieved and changed July 17, 2023 from https://magenta.tensorflow.org/assets/music_vae/architecture.png

- **approaches**
  - historically: Illiac Suite, use markov models
  - just RNNs, LSTMs
    - generate **autoregressively**
    - **difference to Markov models**: MM looks back limited amount of time, RNNs potentially unlimited time
  - MidiNet
    - generate **pianoroll** using **CNN**
  - MusicVAE
    - **hierarchical**
  - TransformerVAE
    - leads to **similar reconstruction acc.**
- **evaluation procedures**
  - quality?
    - listening study
    - objective measures
  - originality?
    - how many notes are equal (regardless of transp.)
  - deficit: only compared to other DL appr.

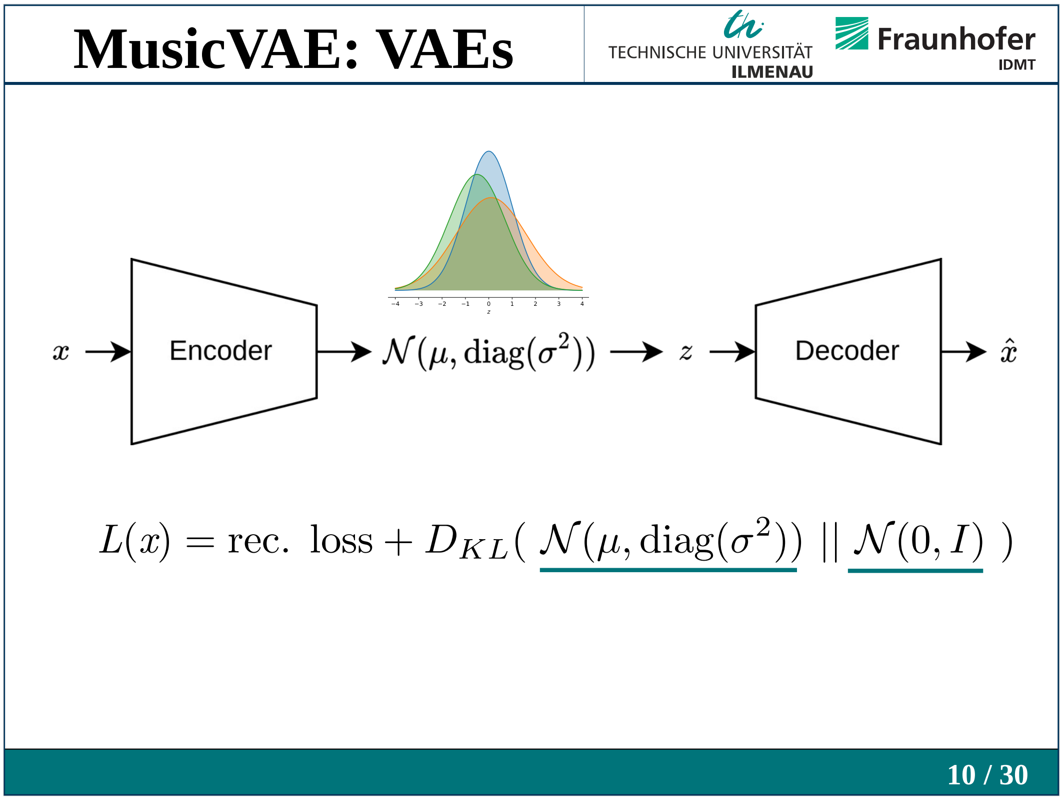
MusicVAE



- **data and representation**
  - there are symbols for each pitch plus a symbol for “no pitch”



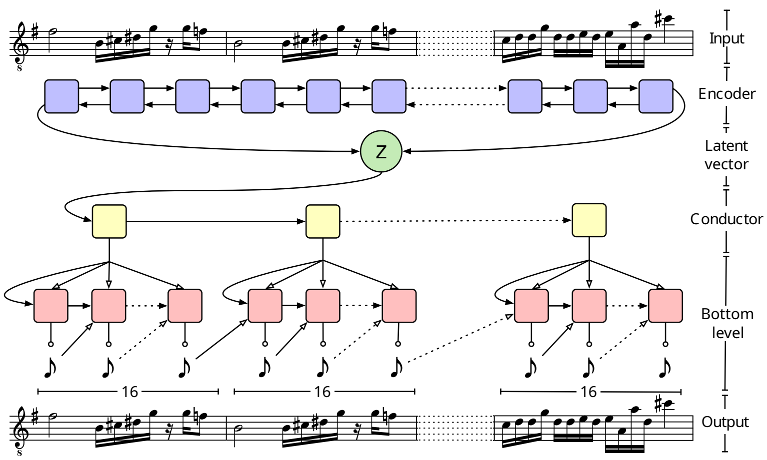
- **general structure** of MusicVAE
  - music → encoder → z with special requirements → decoder → music
- **music is encoded as sequence of vectors.**
  - **each vector represents one 16th note**
- first: RNNs, then VAEs



- x → μ, σ → dist → z → xhat
- **reconstruction loss**
  - cross entropy
- **KL divergence**

MusicVAE

TECHNISCHE UNIVERSITÄT ILMENAU Fraunhofer IDMT



11 / 30

<u>Hierarchical</u>

- **Encoder:**
  - two stacked BLSTMs
- **Decoder:**
  - hierarchical
  - conductor LSTM
    - one vector per bar
  - bottom-level LSTM
    - bar from conductor vector

<u>Flat</u>

- **Encoder:**
  - simple BLSTM
- **Decoder:**
  - multi-layer LSTM

Implementation: Training

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

$$L(x) = \text{rec. loss} + \underline{\beta}\, \max[\, D_{KL},\ \underline{\lambda}\,]$$

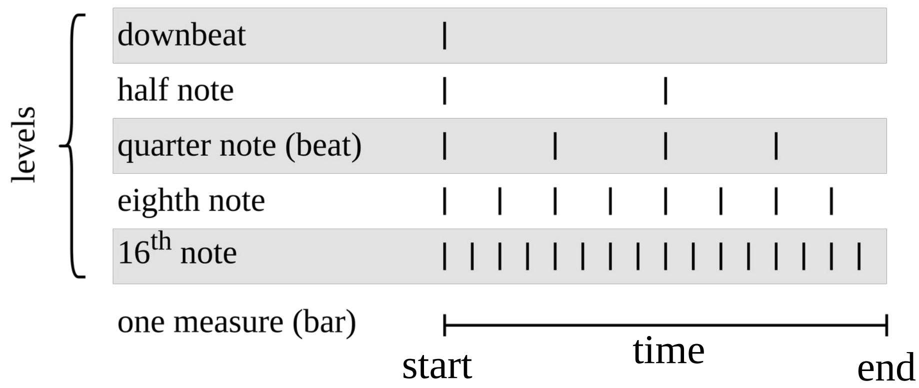
- ***β* and *λ* chosen after grid search**
  - $\beta = 1$; $\lambda = 33.3$
- optimizer
  - Adam
  - learning rate (LR) = 10^{-3}
- batch size = 64
- weight decay
  - L_2 regularization with weight 10^{-6}
- LR scheduling
  - customized variant of `ReduceLROnPlateau`
- early stopping was used

- **loss function**
  - β, λ constant
- **training techniques**
  - Adam, 1e-3
  - batch size 64
  - L2 wd: 1e-6
  - customized ReduceLROnPlateau
  - early stopping

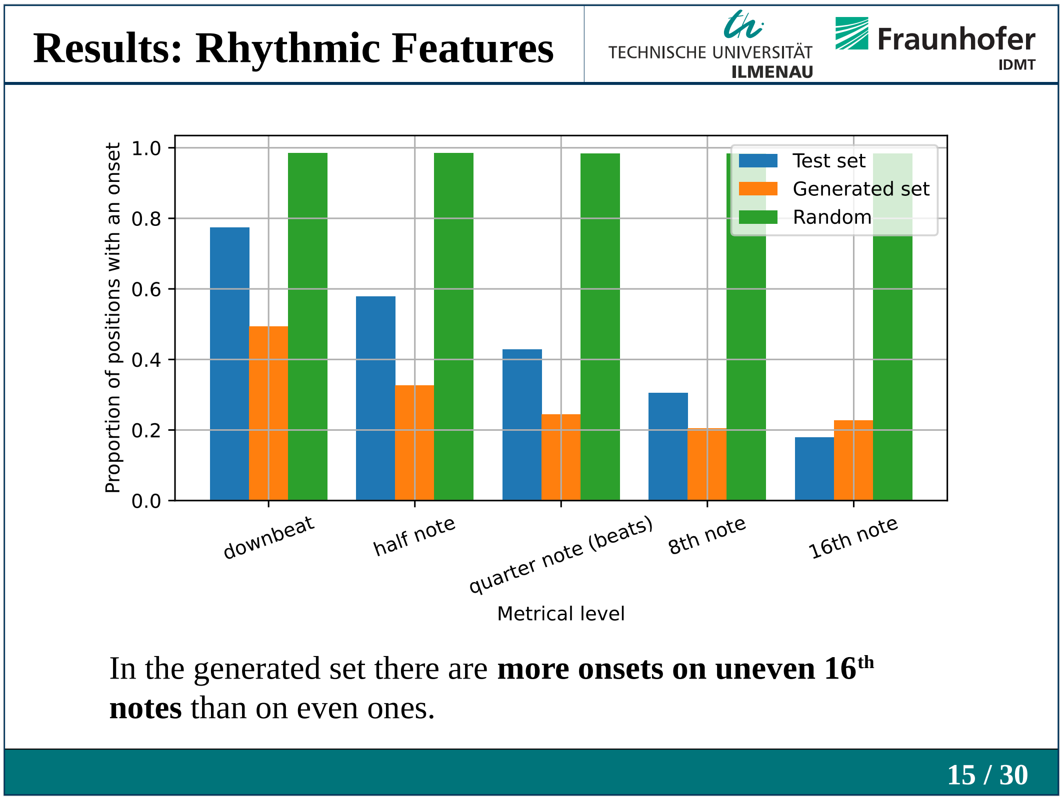
TECHNISCHE UNIVERSITÄT ILMENAU
Fraunhofer IDMT

Results:
Rhythmic Features

Results: Rhythmic Features



Note onsets can be assigned to a **metrical level**.

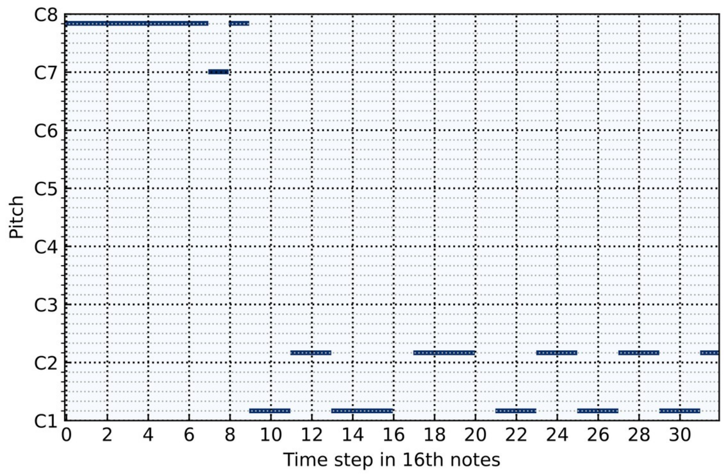


- **onset proportion**
  - …
- **note length**, avg note length
  - …

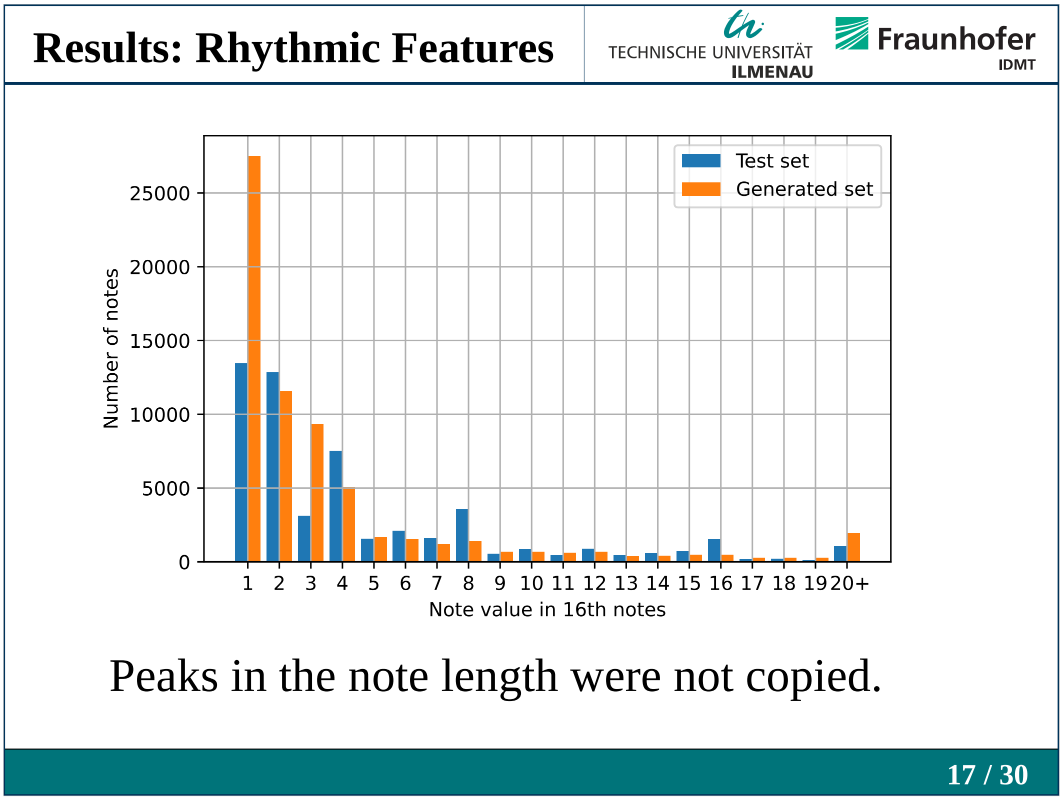
Results: Rhythmic Features

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



Sequence Nr. 30 (Figure 6.4)



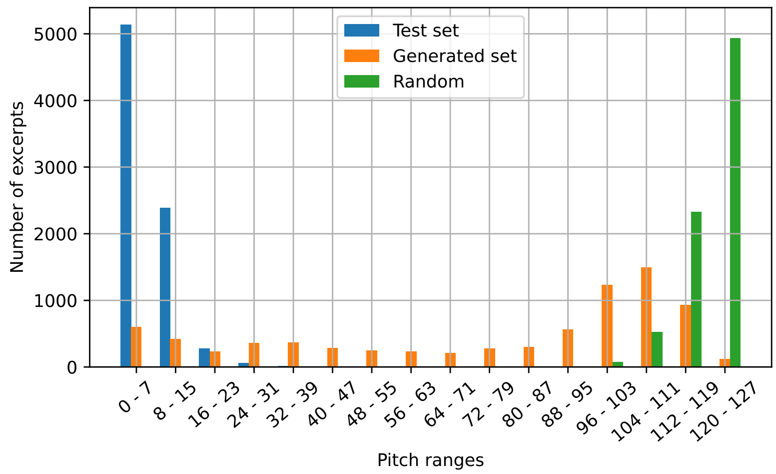
- **onset proportion**
  - ...
- **note length**, avg note length
  - ...

Results:
Melodic Features

Results: Melodic Features

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



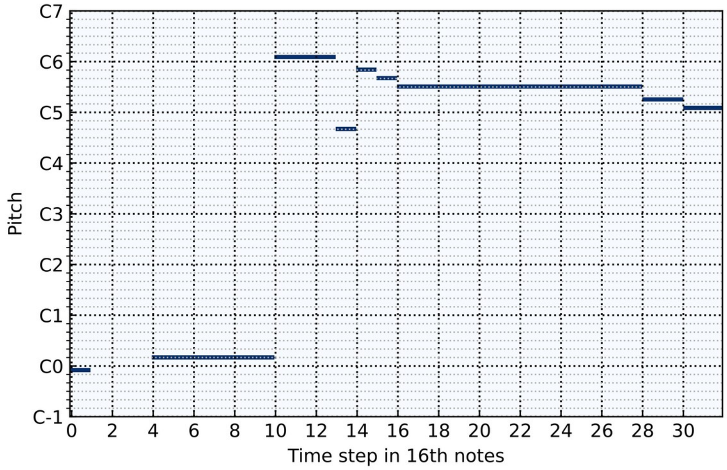
There are **high pitch jumps** in the generated excerpts.

- pitch range
  - ...

Results: Melodic Features

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

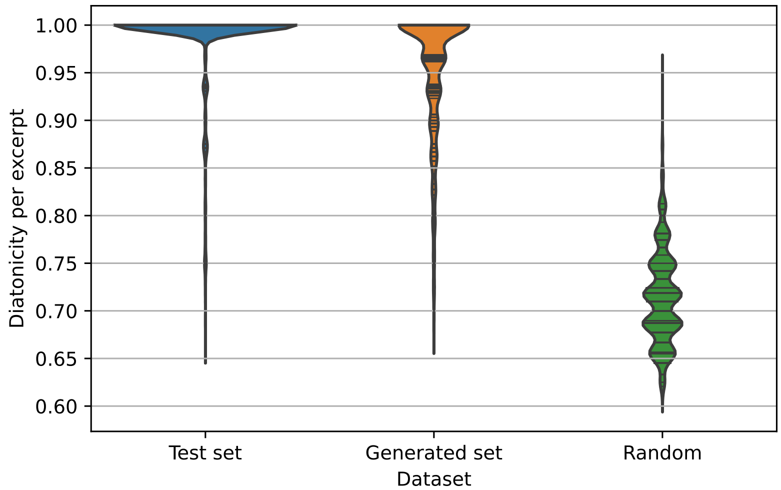


Sequence Nr. 24 (Figure 6.7)

Results: Melodic Features

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

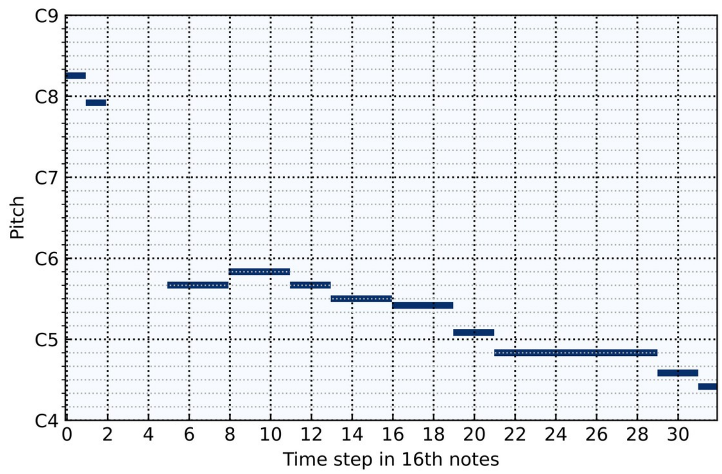


Generated excerpts are **mostly diatonic**, but there are odd notes.

- diatonicity
  - how much it stays in one key

Results: Melodic Features

TECHNISCHE UNIVERSITÄT ILMENAU
Fraunhofer IDMT



Sequence Nr. 13 (Figure 6.11)

TECHNISCHE UNIVERSITÄT ILMENAU
Fraunhofer IDMT

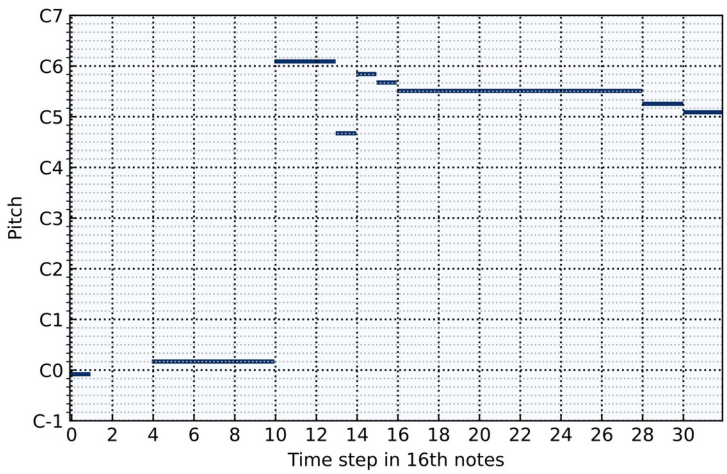
Results: Qualitative Evaluation

- high pitch jumps
- rhythmic differences where not noticed
- ignoring few odd notes, still pleasant and coherent

Results: Melodic Features

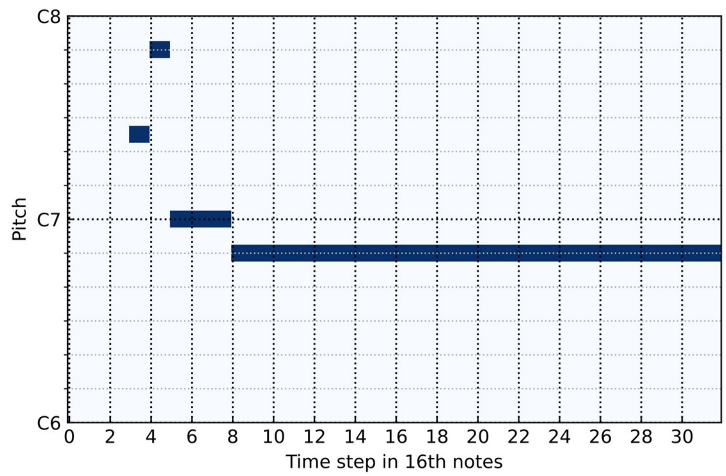
TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



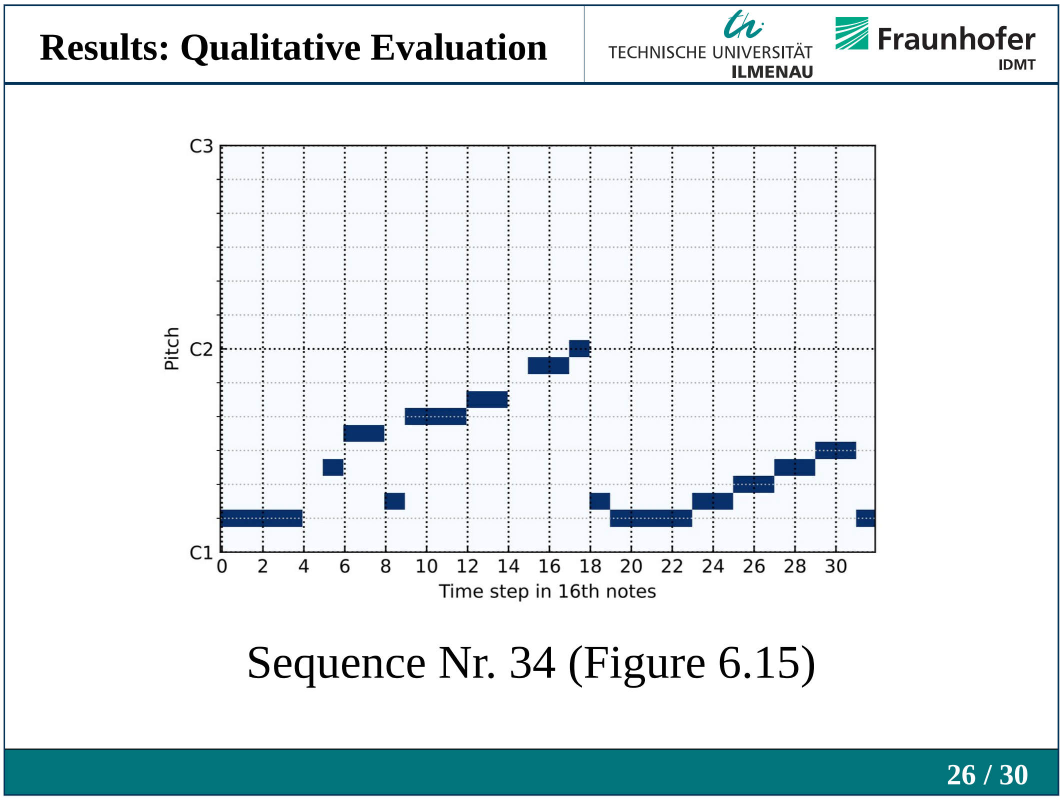
Sequence Nr. 24 (Figure 6.7)

Results: Qualitative Evaluation



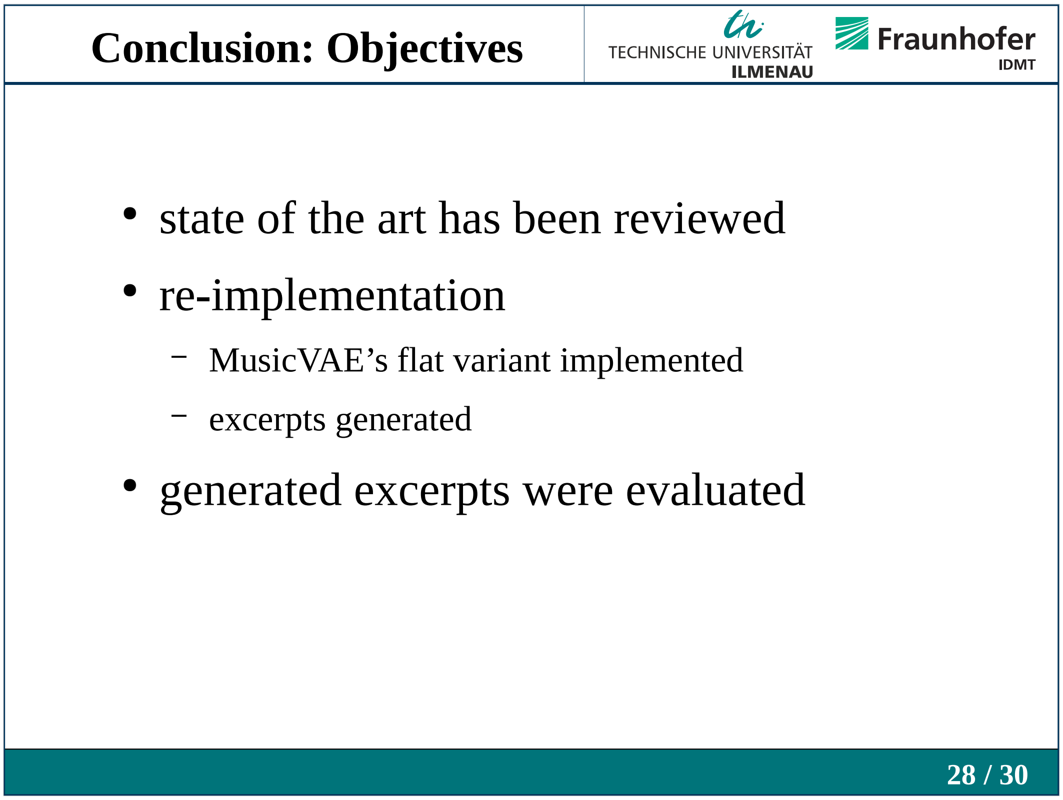
Sequence Nr. 50 (Figure 6.12)

- chimes



- rhythmically interesting

Conclusion



- What has been done
  - MusicVAE re-implemented, excerpts generated & evaluated
  - generated excerpts similar to training data, but still differences
  - apart from these, generated melodies sounded pleasant
- Future work
  - larger dataset, analysis of it
  - regularization techniques to reduce overfitting
  - sudden jumps in the loss → investigate reason and improve training
  - improve hierarchical model
  - condition the latent space →
  - evaluate latent space in a different way

Conclusion: Findings

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

- more **onsets on uneven 16th notes**
- single **non-diatonic notes** & **pitch jumps**
- mostly musically coherent & **pleasant-sounding**

- What has been done
  - MusicVAE re-implemented, excerpts generated & evaluated
  - generated excerpts similar to training data, but still differences
  - apart from these, generated melodies sounded pleasant
- Future work
  - larger dataset, analysis of it
  - regularization techniques to reduce overfitting
  - sudden jumps in the loss → investigate reason and improve training
  - improve hierarchical model
  - condition the latent space →
  - evaluate latent space in a different way

Conclusion: Future Work

TECHNISCHE UNIVERSITÄT ILMENAU
Fraunhofer IDMT

- recreate training **dataset**
- adjust training **procedure**
- **extend** model
  - hierarcical, polyphonic, conditioned

- recreate training dataset
  - larger dataset
  - analyse training dataset more thoroughly before training
  - consider alternative representations
- adjust training procedure
  - tune weight decay
  - scheduled sampling & annealing of β
  - more thorough hyperparameter search
  - find out **reason for jumps in the loss curve** & adjust training accordingly
- extend model
  - implement hierarchical model

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

Thank you!

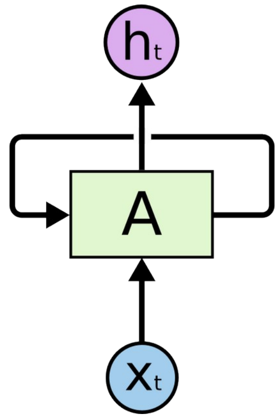
Recurrent Neural Networks

MusicVAE: RNNs

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

RNNs are neural networks with feedback.



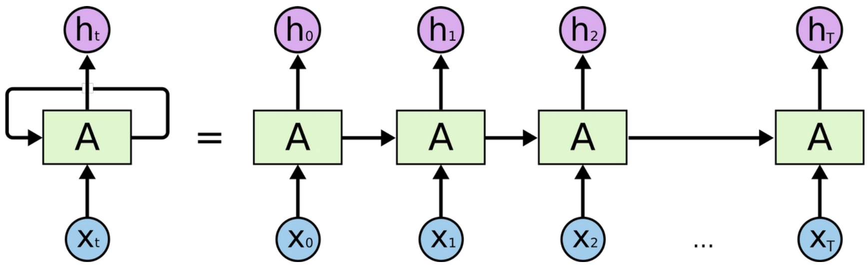
Picture retrieved July 8, 2023 from
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-rolled.png

- RNNs can be used to process time-series data

MusicVAE: RNNs

TECHNISCHE UNIVERSITÄT ILMENAU
Fraunhofer IDMT

RNNs can be represented unrolled.

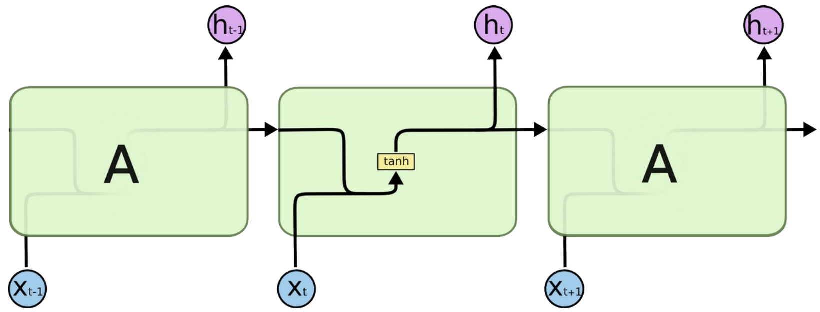


Picture retrieved and changed July 8, 2023 from
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/RNN-unrolled.png

MusicVAE: RNNs

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



$$h_t = \tanh(W_i x_t + b_i + W_h h_{t-1} + b_h)$$

Standart RNNs have some drawbacks.

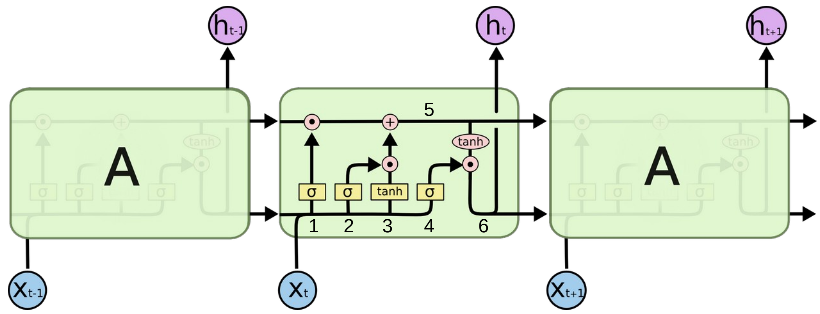
Picture retrieved July 8, 2023 from
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-SimpleRNN.png

- **problems:**
  - …

MusicVAE: RNNs

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



$$1: \; i_t = \sigma(\underline{W_{ii}x_t + b_{ii}} + \underline{W_{hi}h_{t-1} + b_{hi}}) \qquad \sigma(x) = \text{sigmoid}(x)$$

$$2: \; f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf})$$

$$3: \; g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \qquad 5: \; c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$4: \; o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \qquad 6: \; h_t = o_t \odot \tanh(c_t)$$

Picture retrieved and changed July 8, 2023 from
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-chain.png

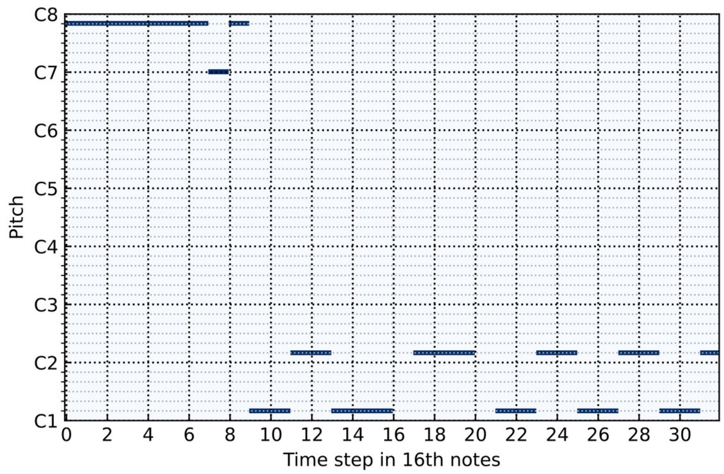
- LSTMs
  - forget gate
  - input gate
  - output gate

Generated Excerpts

Generated Excerpts

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT

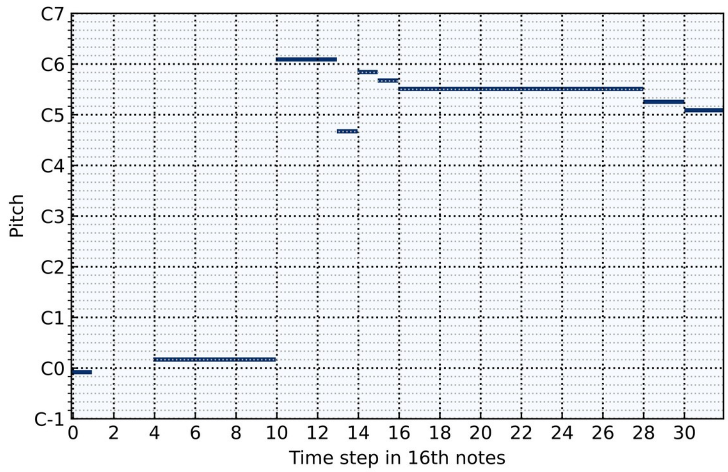


Sequence Nr. 30 (Figure 6.4)

Generated Excerpts

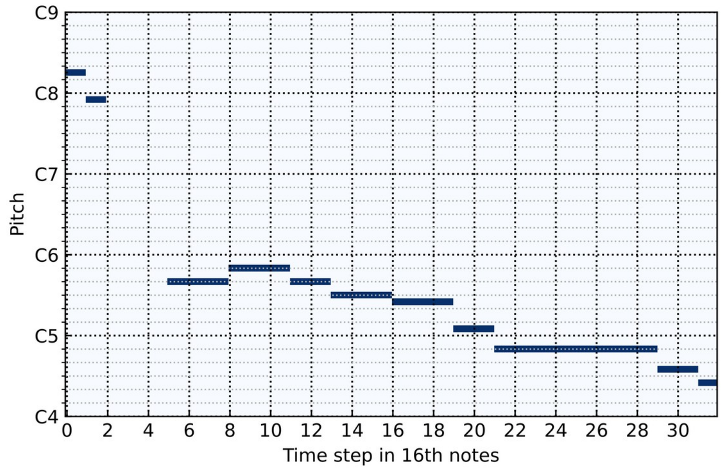
TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



Sequence Nr. 24 (Figure 6.7)

Results: Melodic Features

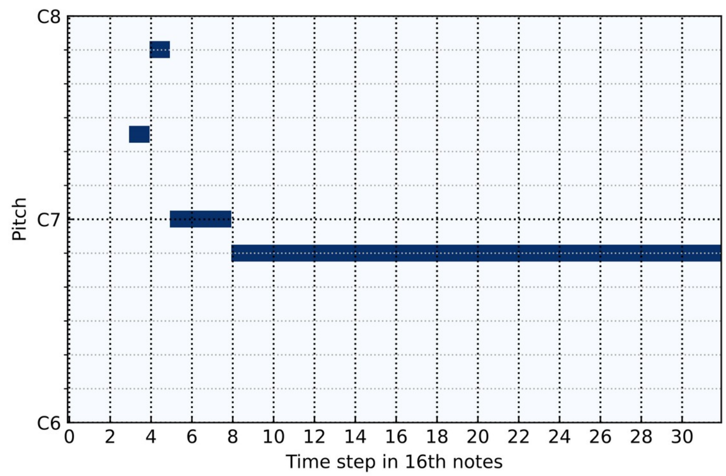


Sequence Nr. 13 (Figure 6.11)

Generated Excerpts

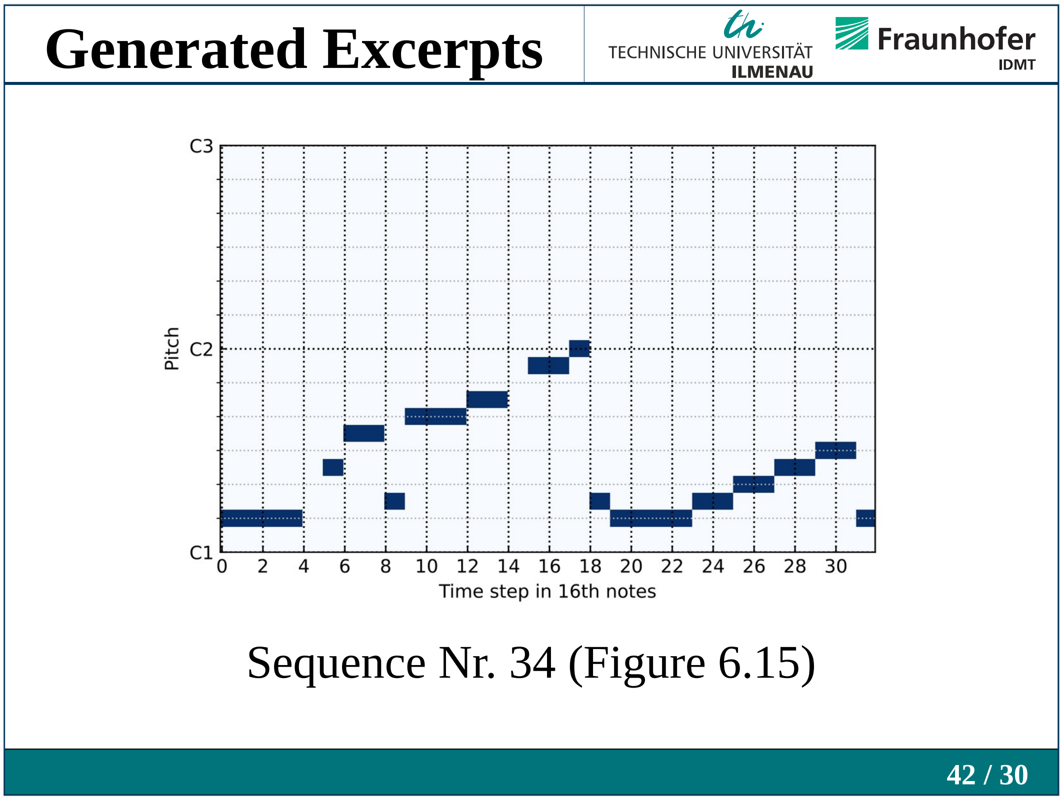
TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



Sequence Nr. 50 (Figure 6.12)

- chimes



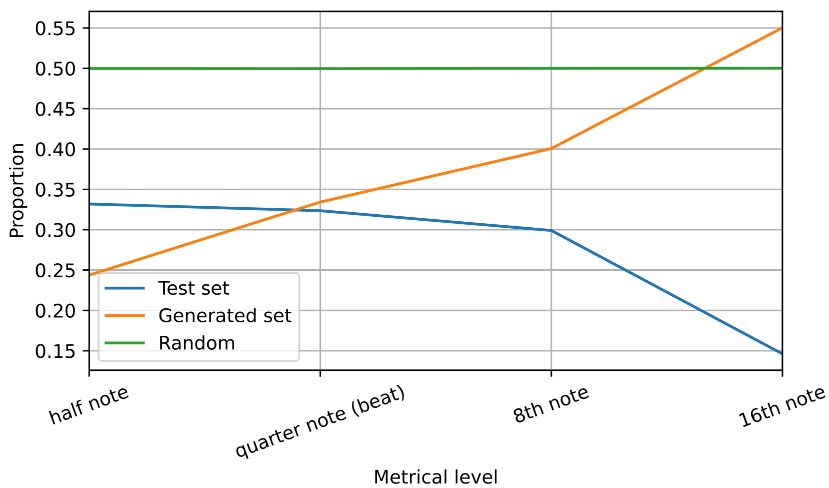
- rhythmically interesting

Evaluation

Results: Rhythmic Features

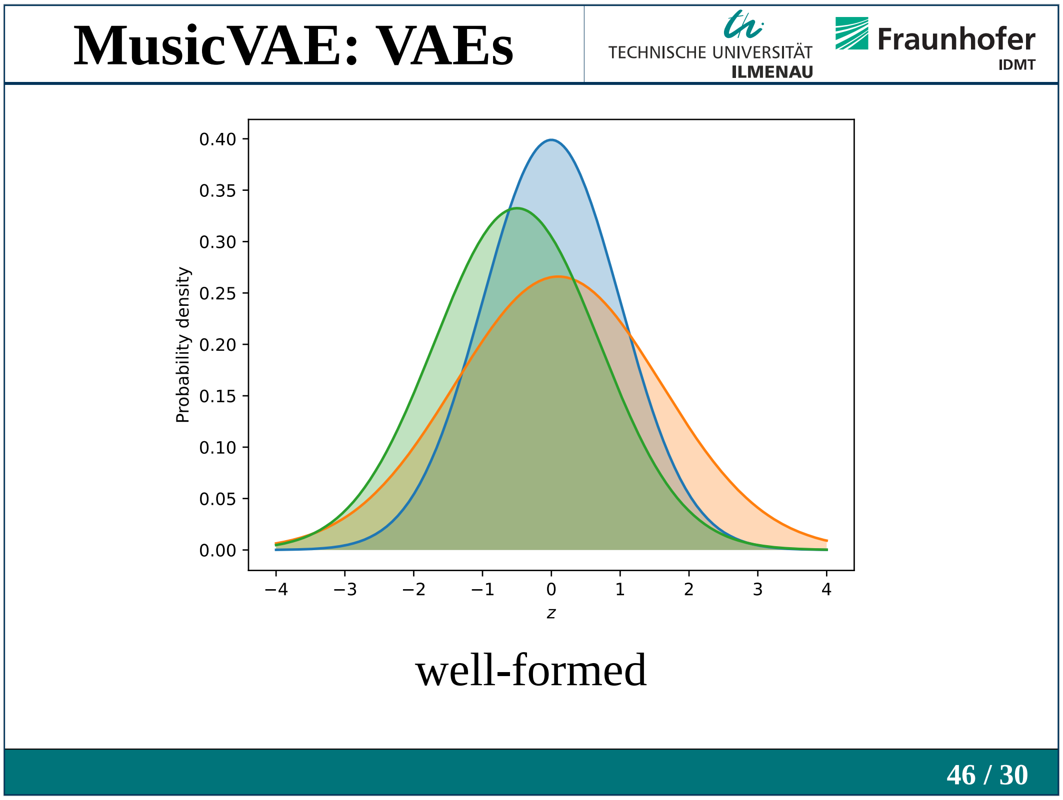
TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



In the generated set there are **more onsets on uneven 16th notes** than on even ones.

Latent Space

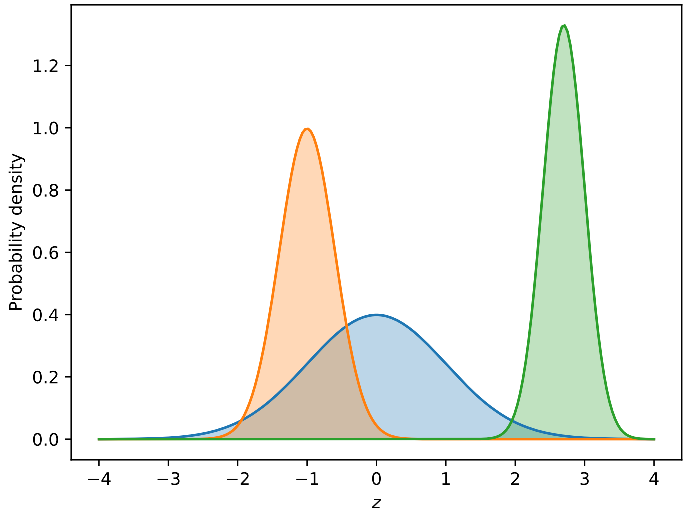


- **Figure: desirable**

MusicVAE: VAEs

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT



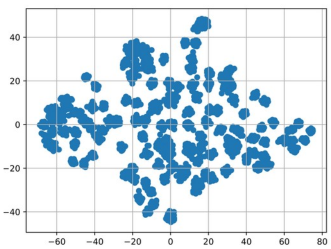
not well-formed

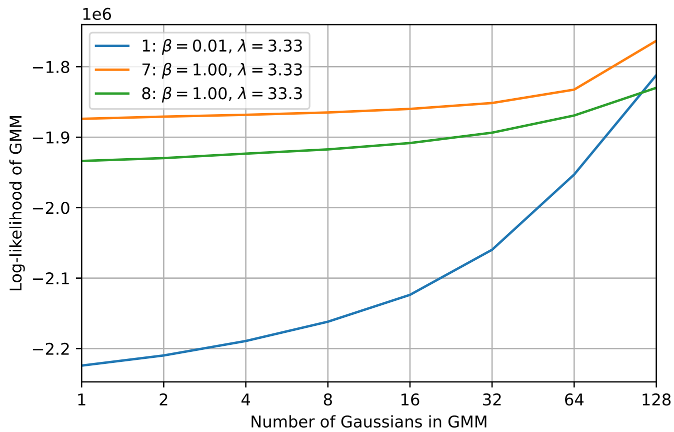
- **Figure: undesirable**

Results: Training

TECHNISCHE UNIVERSITÄT ILMENAU

Fraunhofer IDMT





The structure of the latent spaces of Models 1, 7, and 8 was examined.

- **grid search**
  - …
- structure of **latent space**