DRUM TRANSCRIPTION VIA JOINT BEAT AND DRUM MODELING USING CONVOLUTIONAL RNNs

Richard Vogl

richard.vogl@tuwien.ac.at

ifs.tuwien.ac.at/~vogl

21st Vienna Deep Learning Meetup

15th of October 2018

TECHNISCHE UNIVERSITÄT WIEN

mir group

JOHANNES KEPLER UNIVERSITÄT LINZ

Institute of Computational Perception

DRUM TRANSCRIPTION VIA JOINT BEAT AND DRUM MODELING USING CONVOLUTIONAL RNNs

Richard Vogl[1,2], Matthias Dorfer[2], Gerhard Widmer[2], Peter Knees[1]

richard.vogl@tuwien.ac.at, matthias.dorfer@jku.at, gerhard.widmer@jku.at, peter.knees@tuwien.ac.at

ISMIR

2017, SUZHOU, CHINA

1 TECHNISCHE UNIVERSITÄT WIEN

ifs

mir group

2 JOHANNES KEPLER UNIVERSITÄT LINZ

Institute of Computational Perception

PART 1
AUTOMATIC DRUM TRANSCRIPTION

Task Definition, Problem Modeling, Architectures

PART 2
MULTI-TASK LEARNING

Metadata for Transcripts

PART 1
AUTOMATIC DRUM TRANSCRIPTION

Task Definition, Problem Modeling, Architectures

PART 2
MULTI-TASK LEARNING

Metadata for Transcripts

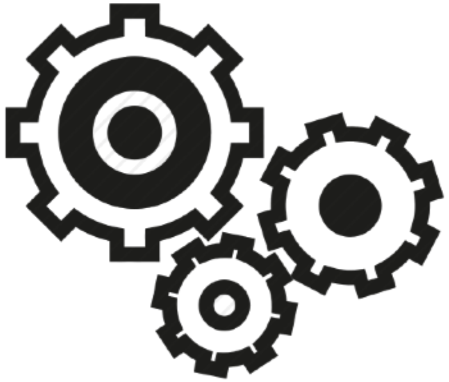
WHAT IS DRUM TRANSCRIPTION?

audio

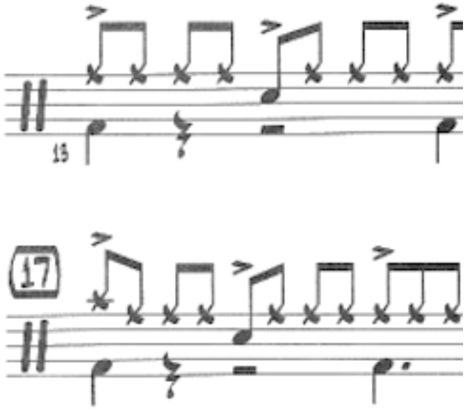
WHAT IS DRUM TRANSCRIPTION?

audio

ADT system



symbolic representation
of drum notes



sheet music

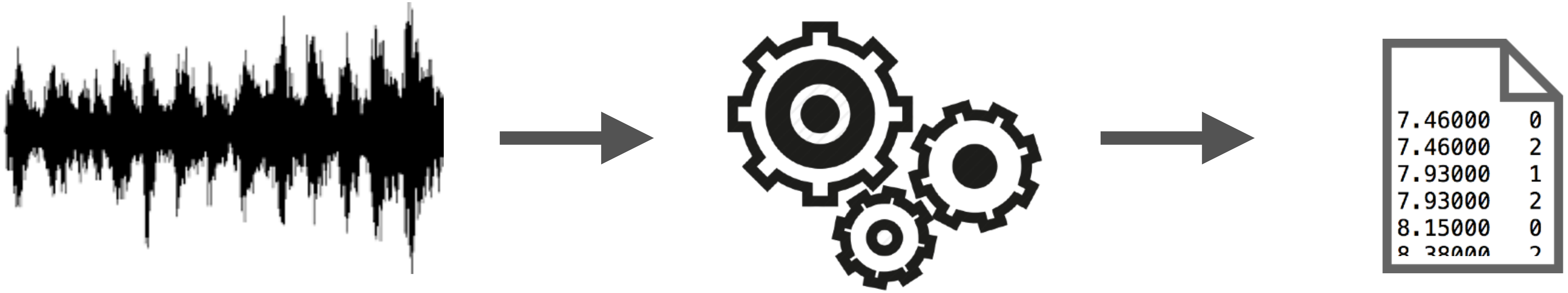
- **Input:** western popular music containing drums
- **Output:** symbolic representation of notes played by drum instruments

WHAT IS DRUM TRANSCRIPTION?

audio

ADT system

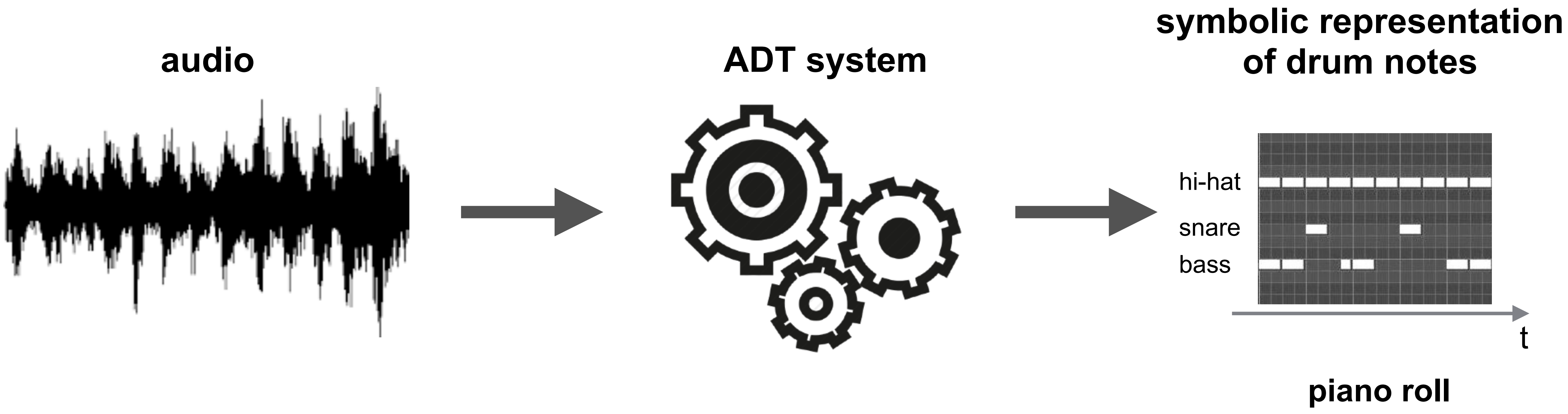
symbolic representation
of drum notes



list of events

- **Input:** western popular music containing drums
- **Output:** symbolic representation of notes played by drum instruments

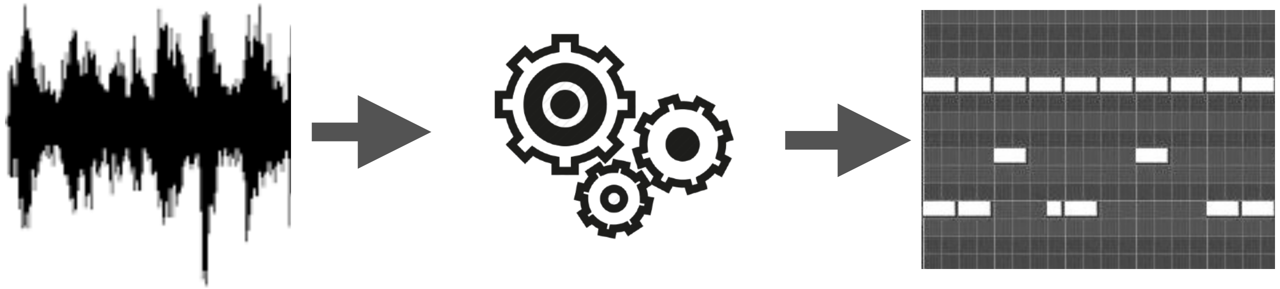
WHAT IS DRUM TRANSCRIPTION?



- **Input:** western popular music containing drums
- **Output:** symbolic representation of notes played by drum instruments

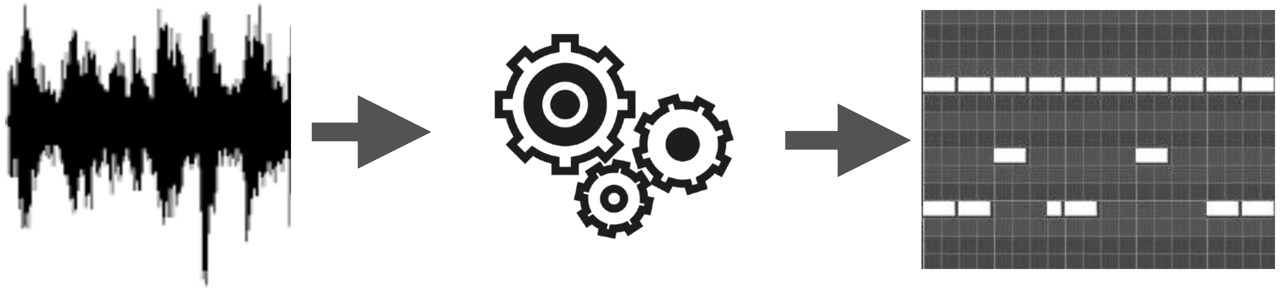
WHAT IS DRUM TRANSCRIPTION?

audio **ADT system** **notes**



WHAT IS DRUM TRANSCRIPTION?

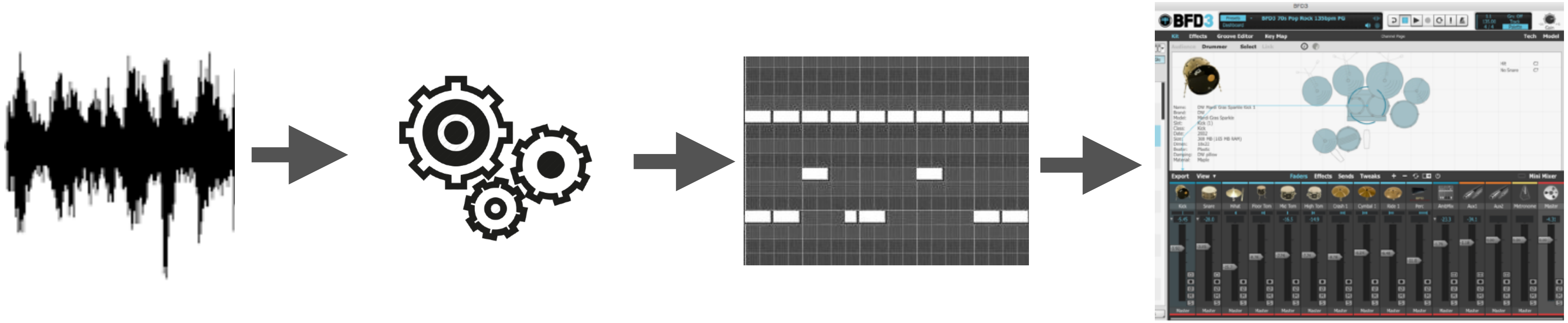
audio ADT system notes



♫

WHAT IS DRUM TRANSCRIPTION?

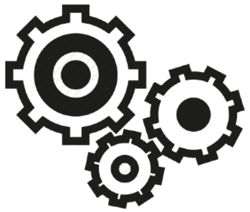
audio ADT system notes drum sampler



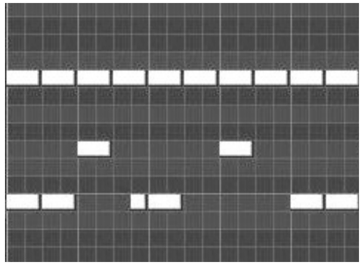
WHAT IS DRUM TRANSCRIPTION?

audio

ADT system



notes

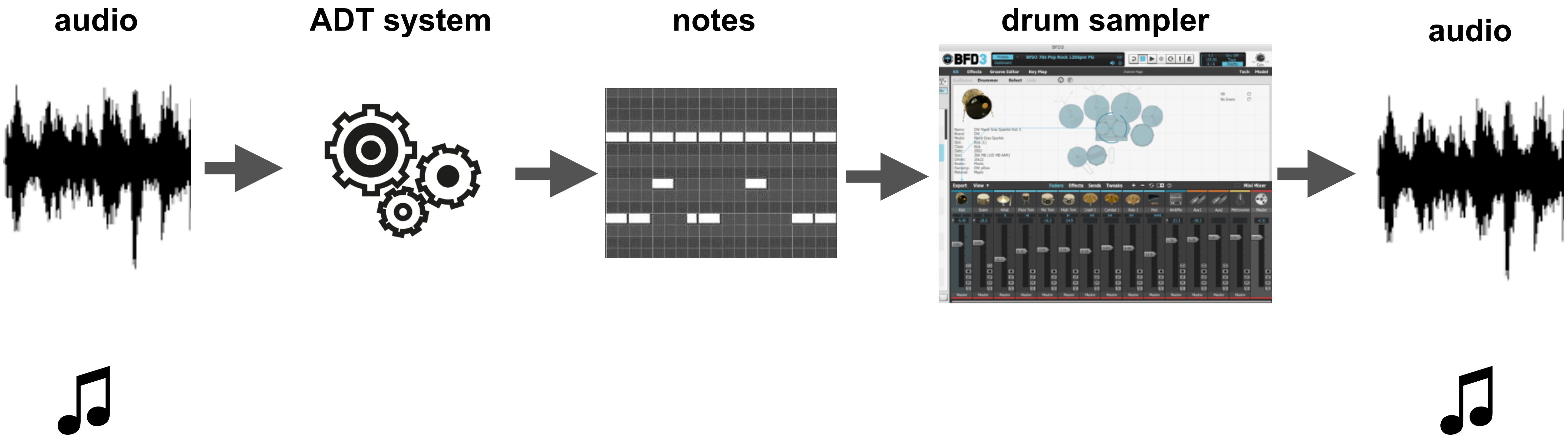


drum sampler

audio

♫

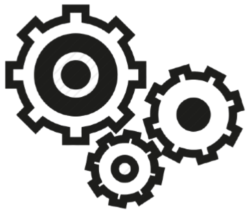
WHAT IS DRUM TRANSCRIPTION?



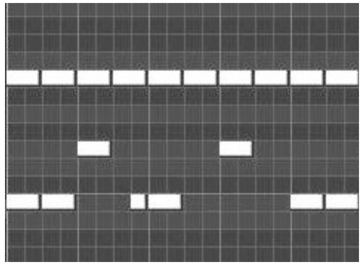
WHAT IS DRUM TRANSCRIPTION?

audio

ADT system



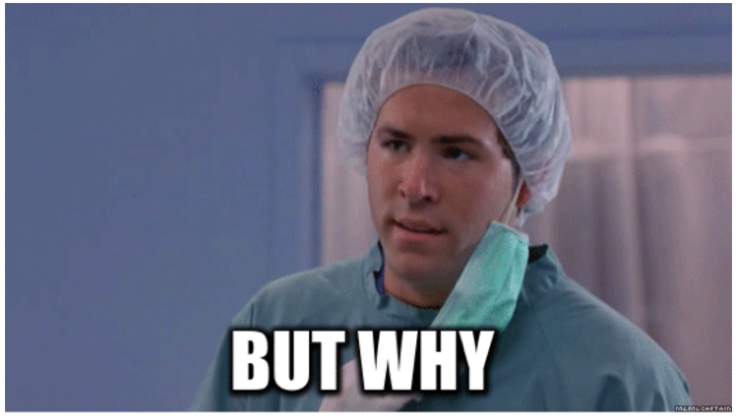
notes



drum sampler

audio

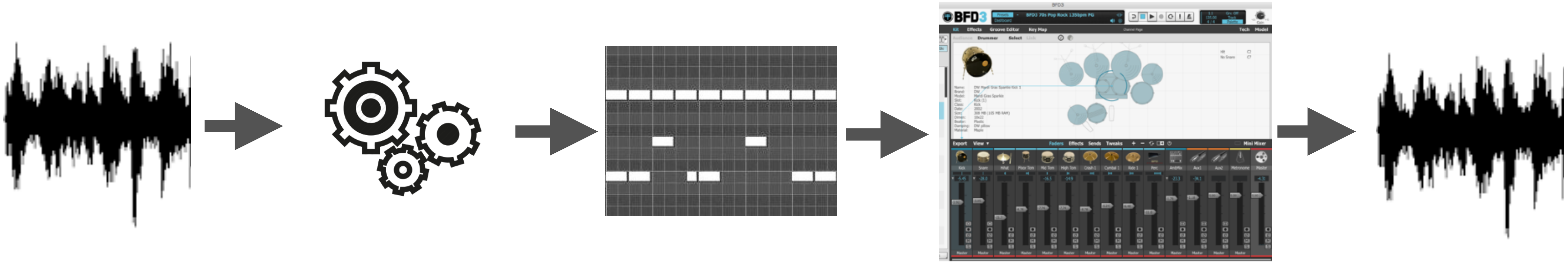
♫

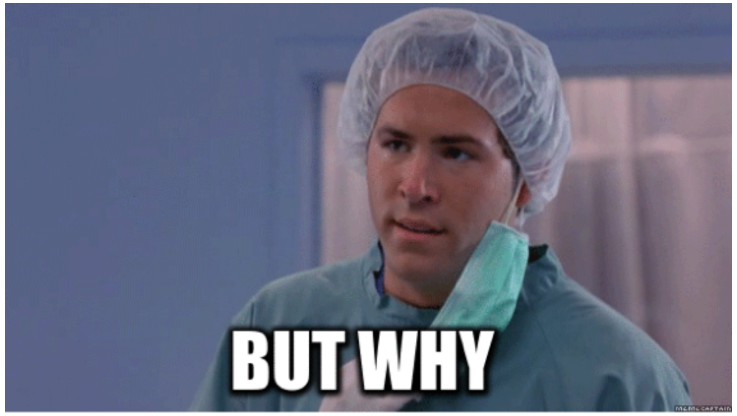


♫

WHAT IS DRUM TRANSCRIPTION?

audio | ADT system | notes | drum sampler | audio

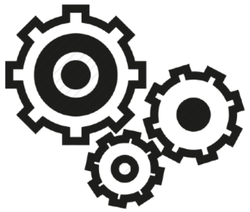




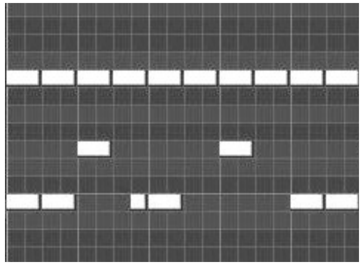
WHAT IS DRUM TRANSCRIPTION?

audio

ADT system



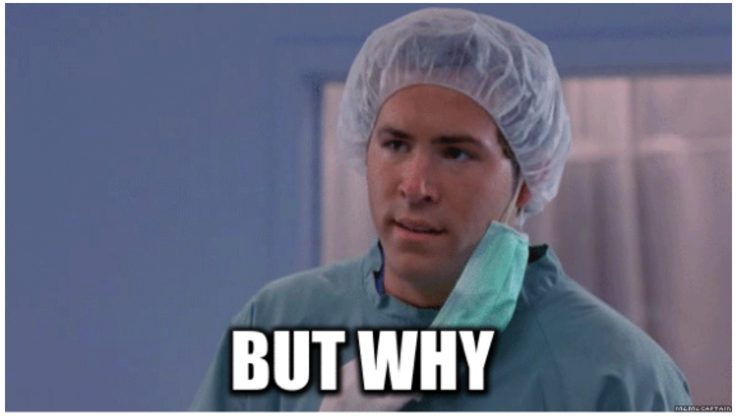
notes



drum sampler

audio

♫ ♫

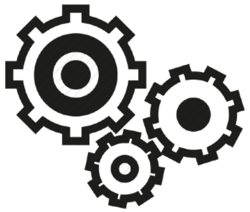


♫ ♫

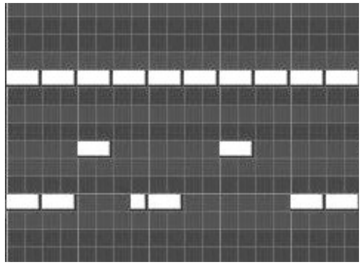
WHAT IS DRUM TRANSCRIPTION?

audio

ADT system

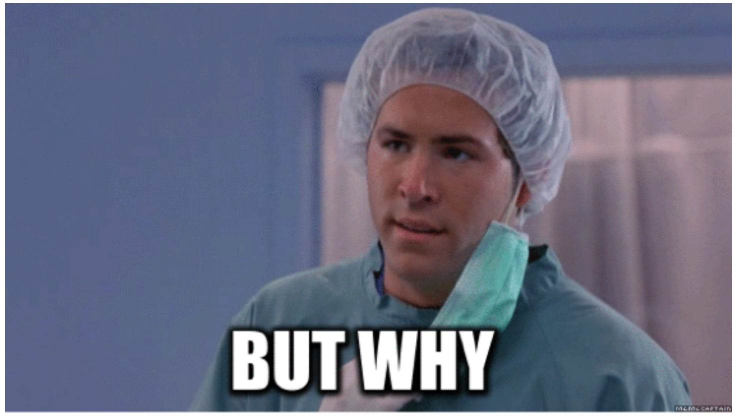


notes



drum sampler

audio



WHY DRUM TRANSCRIPTION?

WHY DRUM TRANSCRIPTION?

- Wide range of application

WHY DRUM TRANSCRIPTION?

- Wide range of application
  - Generate **sheet music**

WHY DRUM TRANSCRIPTION?

- Wide range of application
  - Generate **sheet music**
  - **Music production**
    sampling / remixing / resynthesis

WHY DRUM TRANSCRIPTION?

- Wide range of application
  - Generate **sheet music**
  - **Music production**
    sampling / remixing / resynthesis
  - Higher level **MIR tasks**
    use drum patterns for other tasks
    genre classification
    song segmentation

FOCUSED INSTRUMENTS

FOCUSED INSTRUMENTS

- ADT methods focus bass drum (BD) snare (SD) and hi-hat (HH)

FOCUSED INSTRUMENTS

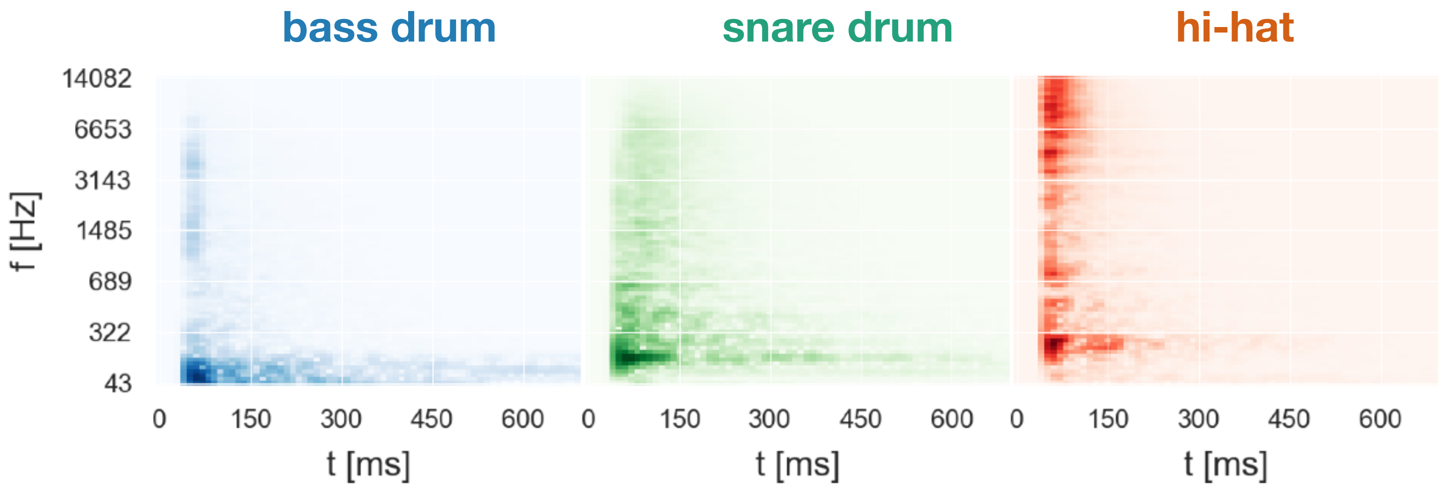
- ADT methods focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets

FOCUSED INSTRUMENTS

- ADT methods focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**

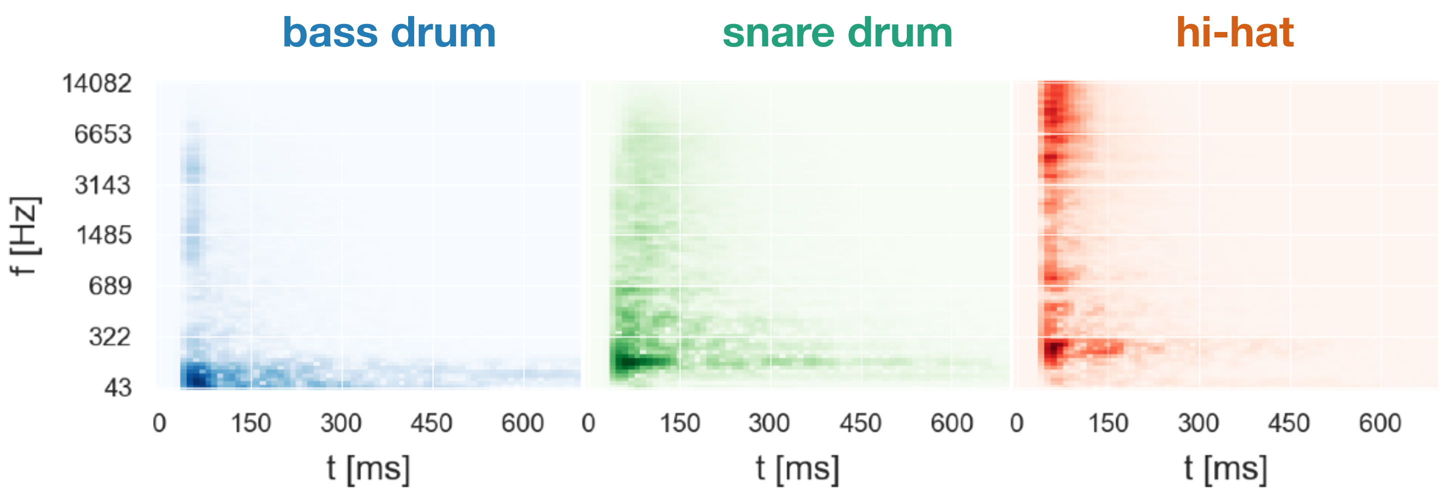
FOCUSED INSTRUMENTS

- ADT methods focus bass drum (**BD**) snare (**SD**) and hi-hat (**HH**)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy distribution**



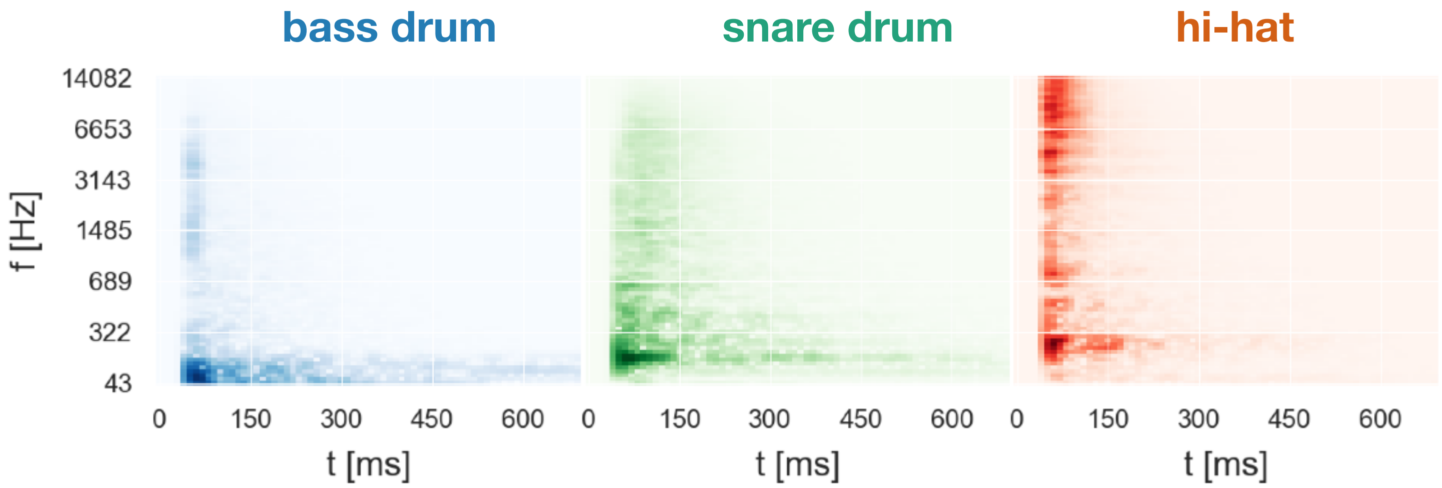
FOCUSED INSTRUMENTS

- ADT methods focus bass drum (**BD**) snare (**SD**) and hi-hat (**HH**)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy distribution**



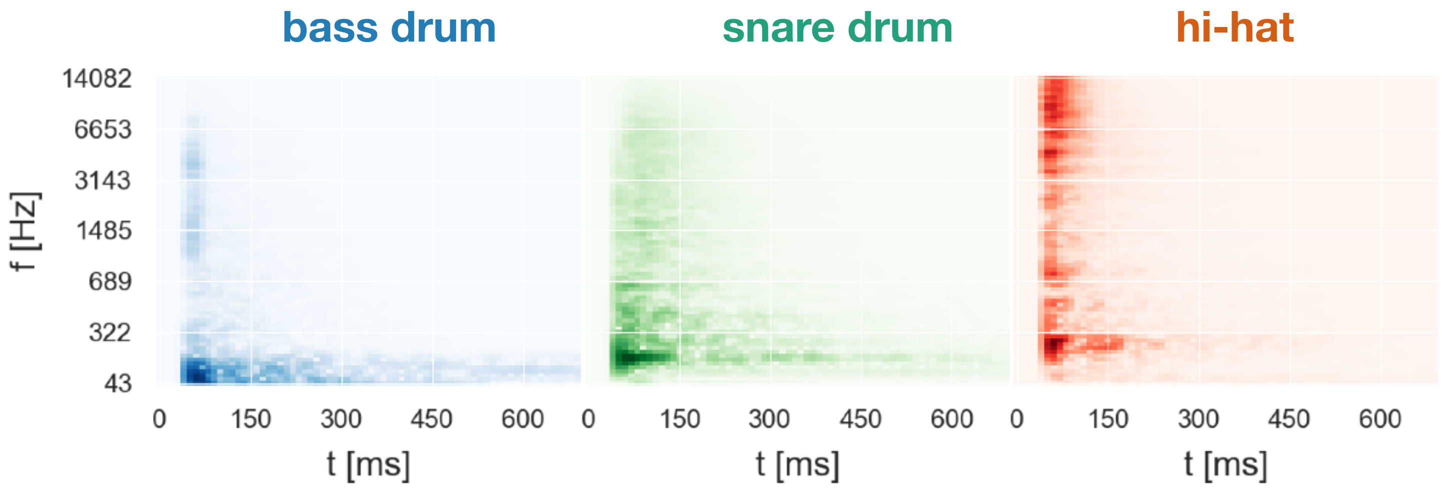
FOCUSED INSTRUMENTS

- ADT methods focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy distribution**

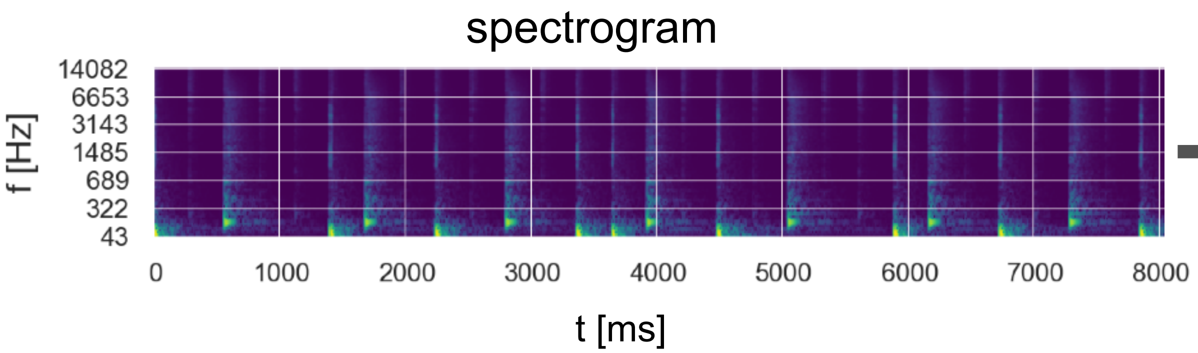


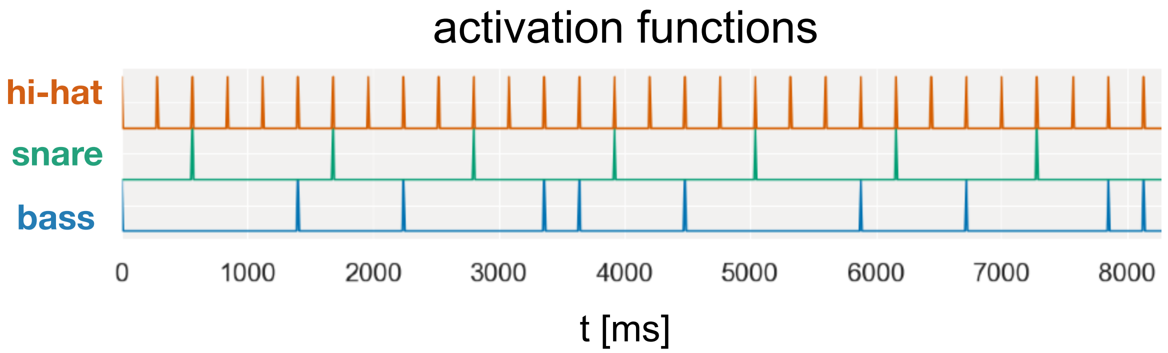
FOCUSED INSTRUMENTS

- ADT methods focus bass drum (BD) snare (SD) and hi-hat (HH)
  - Make up **majority of notes** in datasets
  - Beat defining / **most important**
  - Well **separated spectral energy distribution**



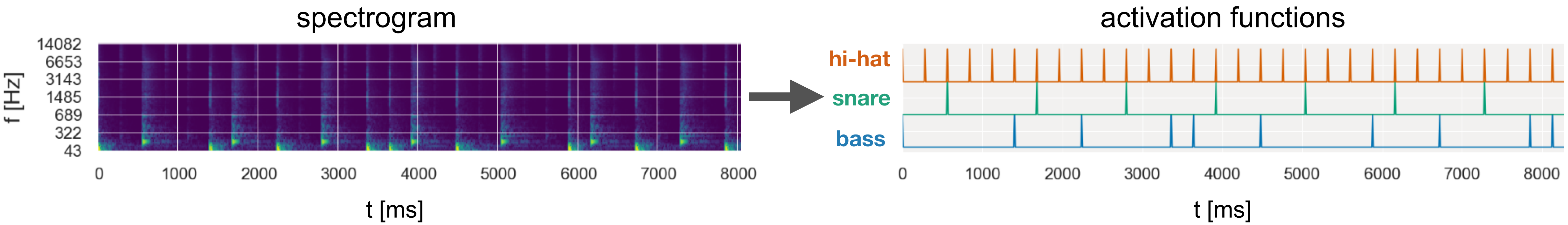
STATE OF THE ART





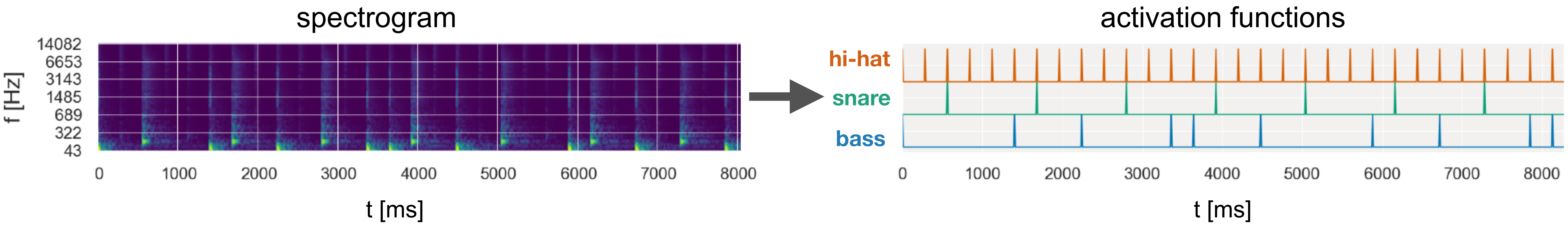
STATE OF THE ART

- End-to-end / **activation-function-based**
- **Neural Networks** and **NMF-based** approaches



STATE OF THE ART

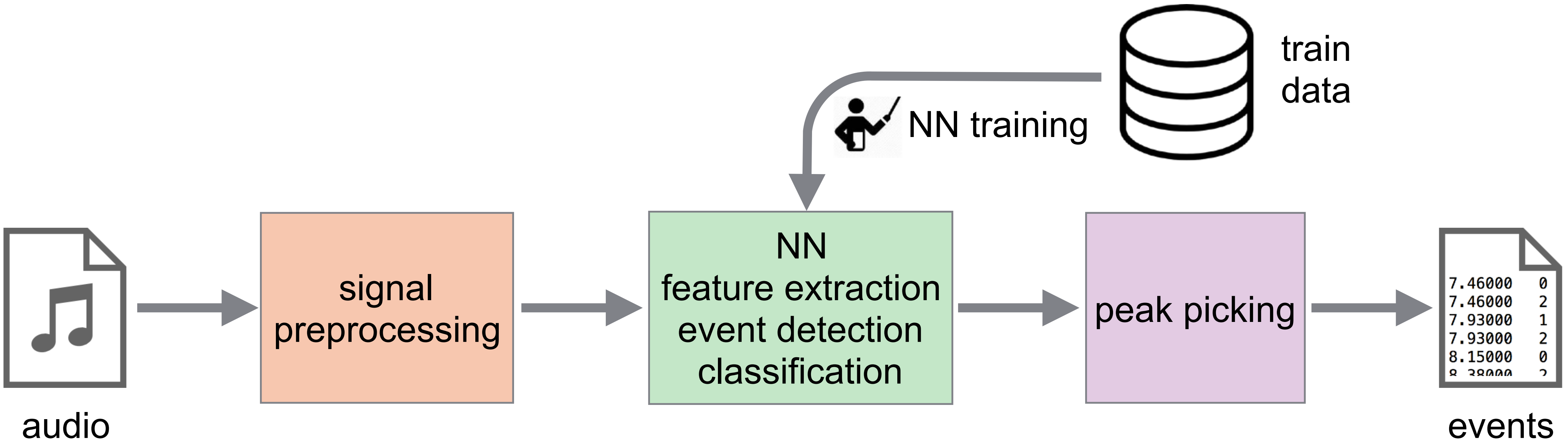
- End-to-end / **activation-function-based**
- **Neural Networks** and **NMF-based** approaches



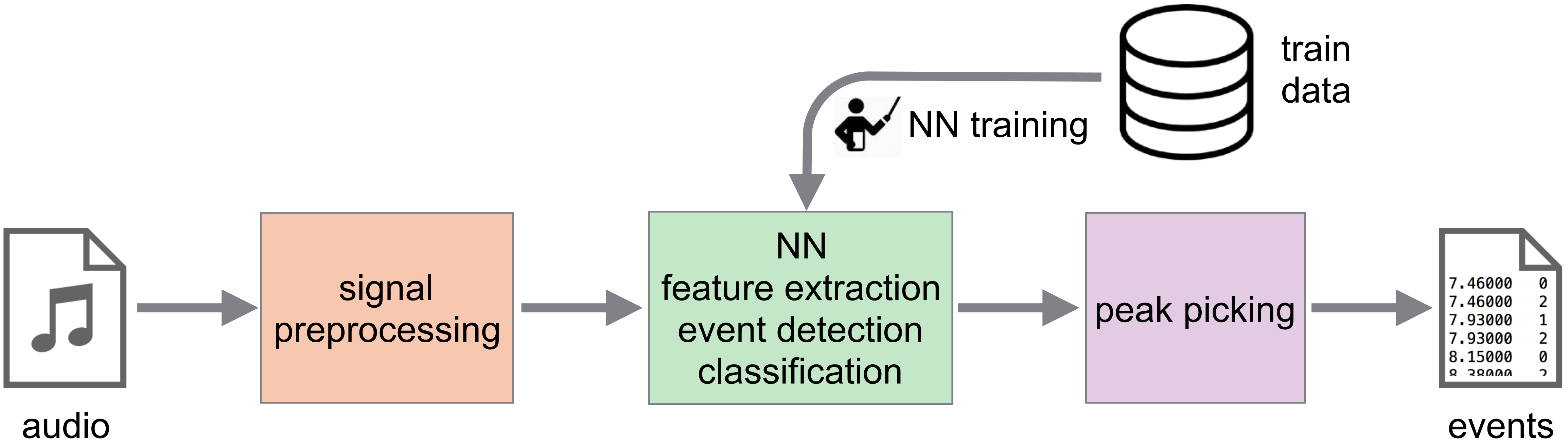
- Overview Article

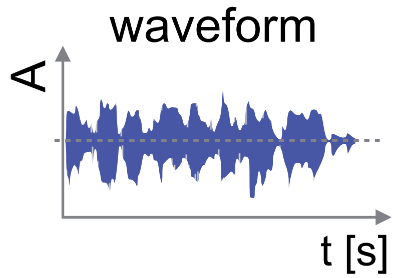
  Wu, C.-W., Dittmar, C., Southall, C.,Vogl, R., Widmer, G., Hockman, J., Müller, M., Lerch, A.: "**An Overview of Automatic Drum Transcription**," IEEE TASLP, vol. 26, no. 9, Sept. 2018.

SYSTEM OVERVIEW

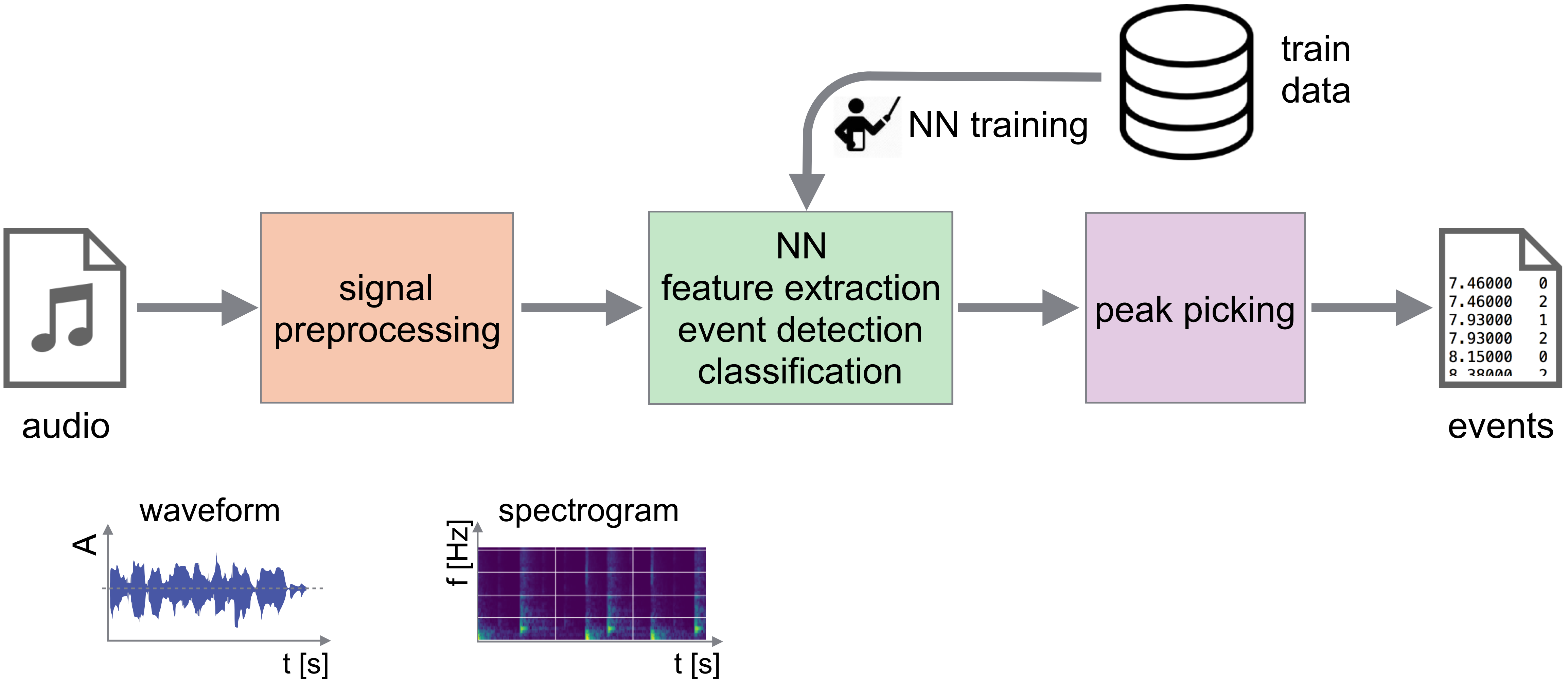


SYSTEM OVERVIEW

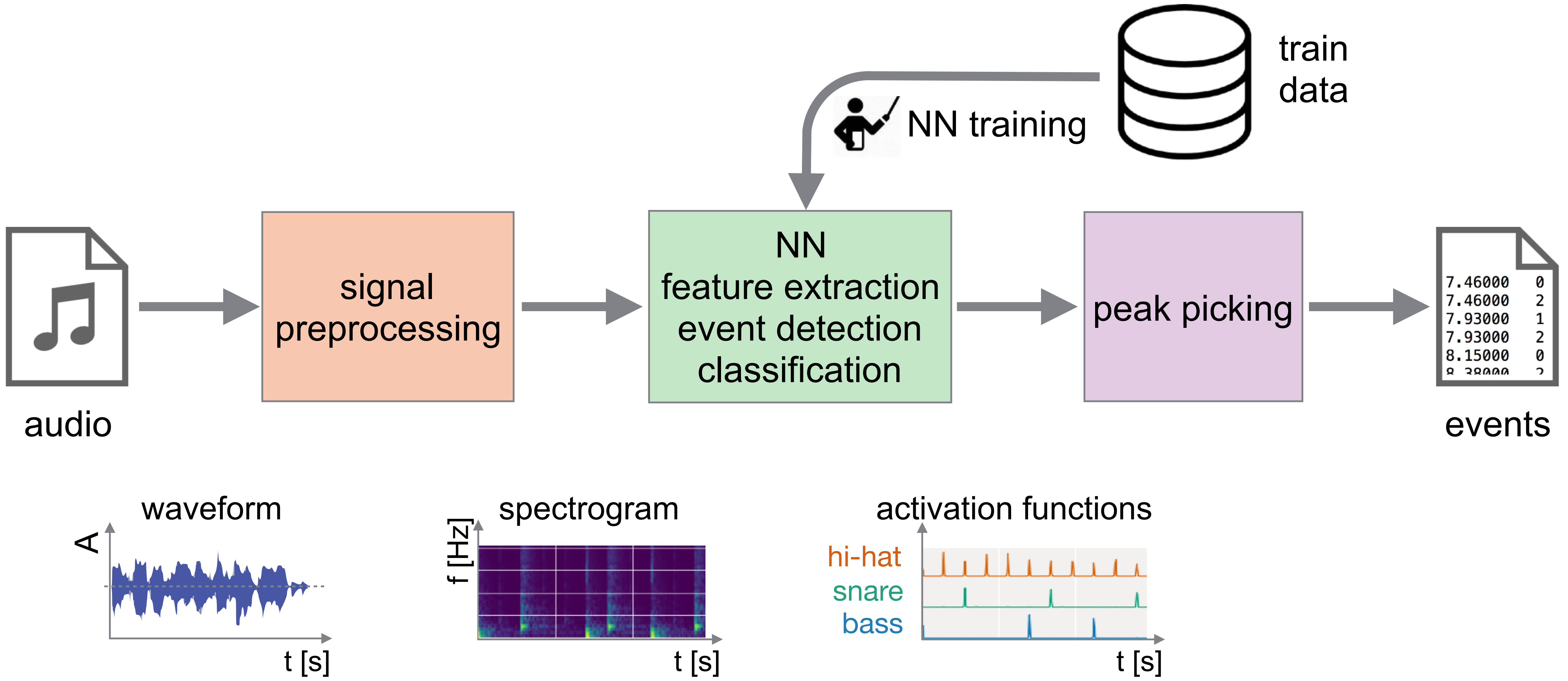




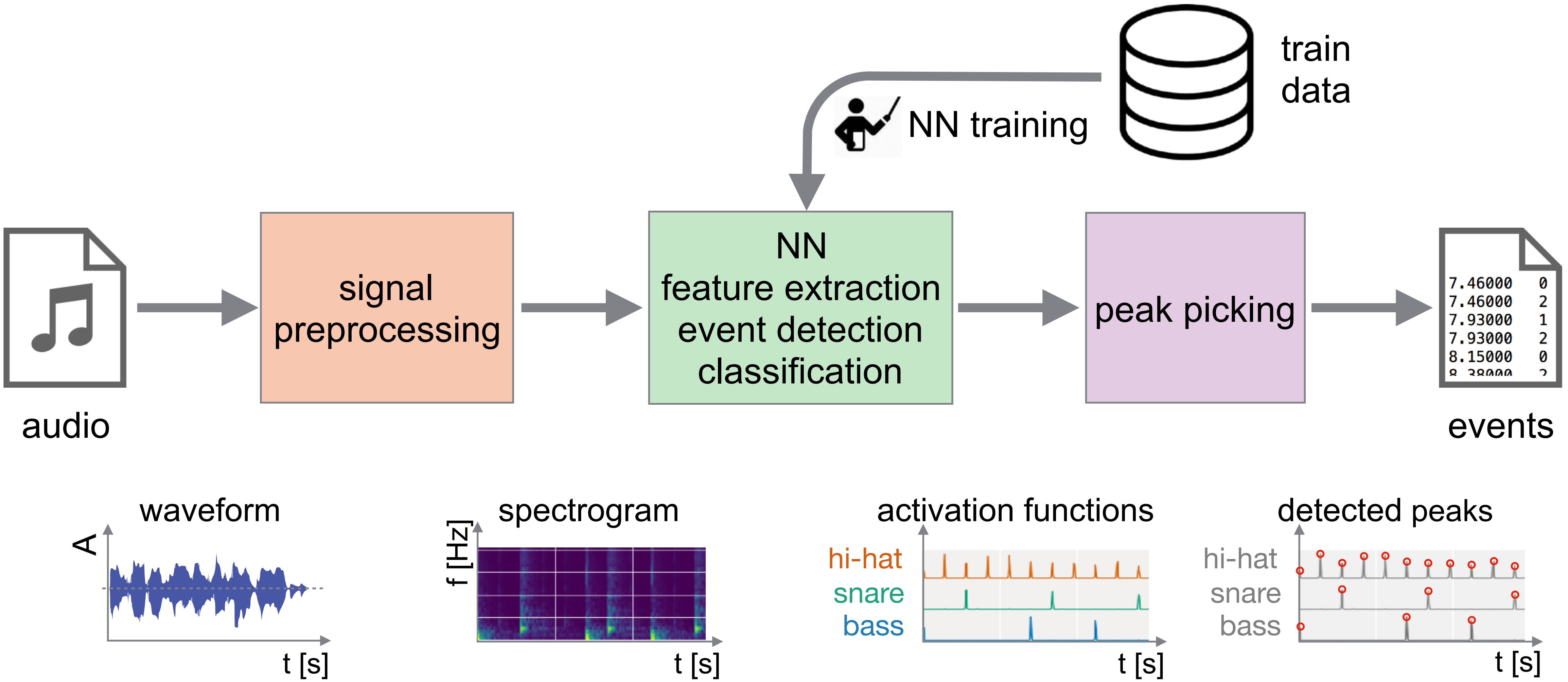
SYSTEM OVERVIEW



SYSTEM OVERVIEW



SYSTEM OVERVIEW



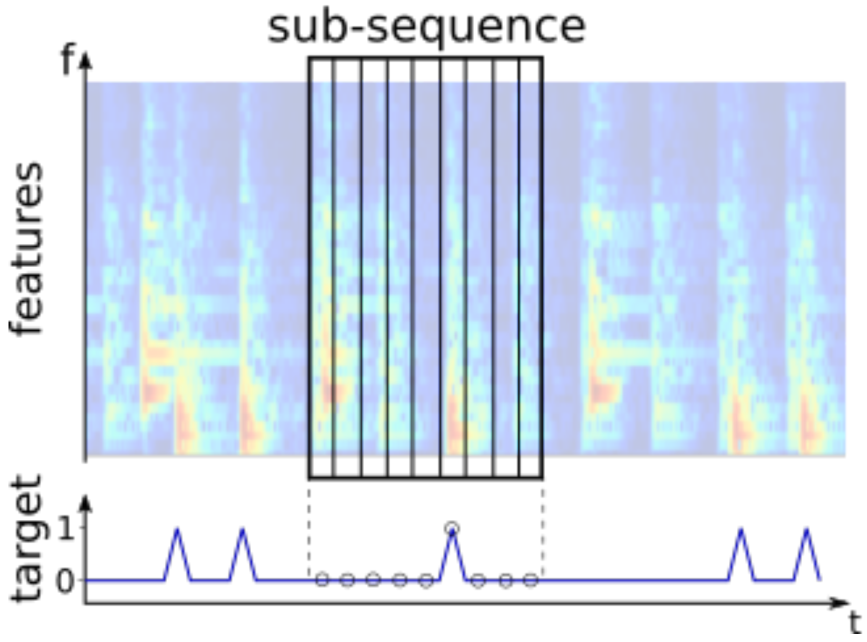
NETWORK MODELS — RNN

NETWORK MODELS — RNN

- Processing of spectrogram frames as **sequential data**

NETWORK MODELS — RNN

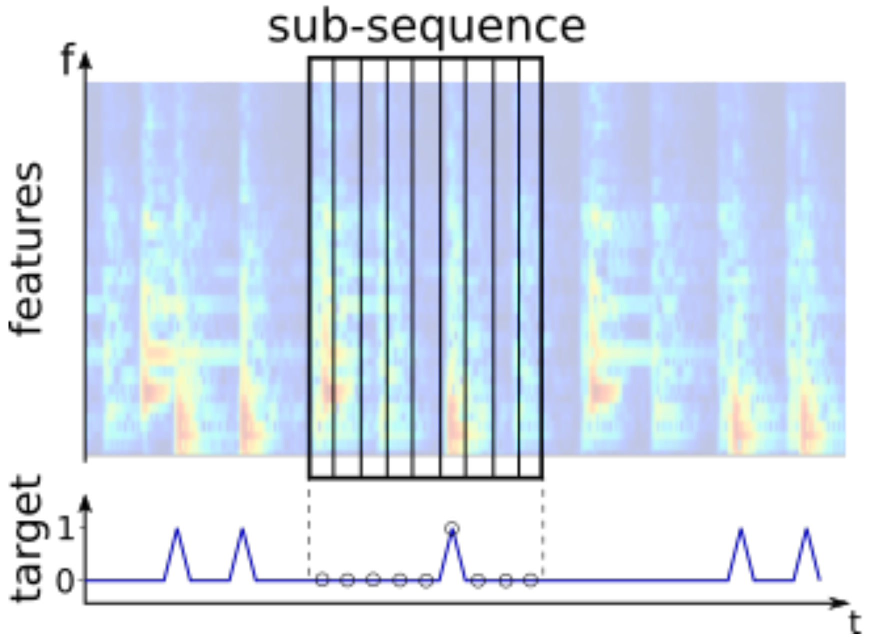
- Processing of spectrogram frames as **sequential data**
- **Frame-wise detection** of instrument onsets



RNN train data sample

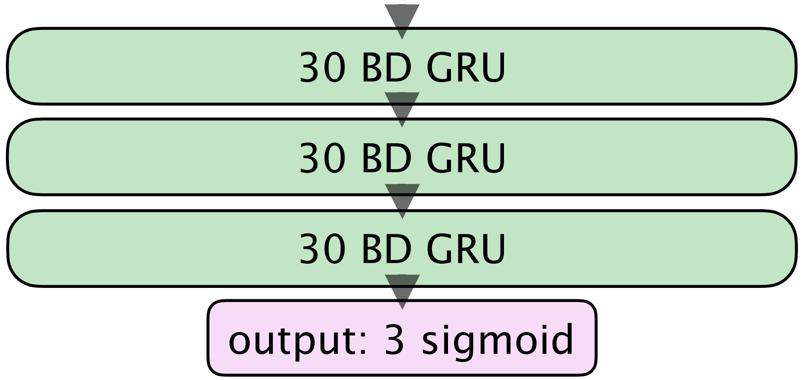
NETWORK MODELS — RNN

- Processing of spectrogram frames as **sequential data**
- **Frame-wise detection** of instrument onsets



RNN train data sample

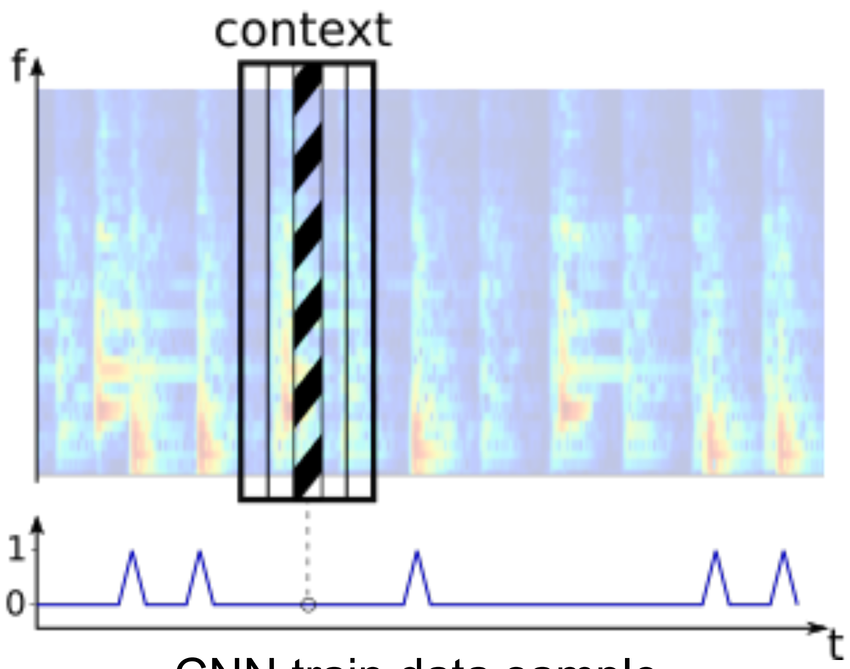
bidirectional RNN architecture with GRUs:



NETWORK MODELS — CNN

NETWORK MODELS — CNN

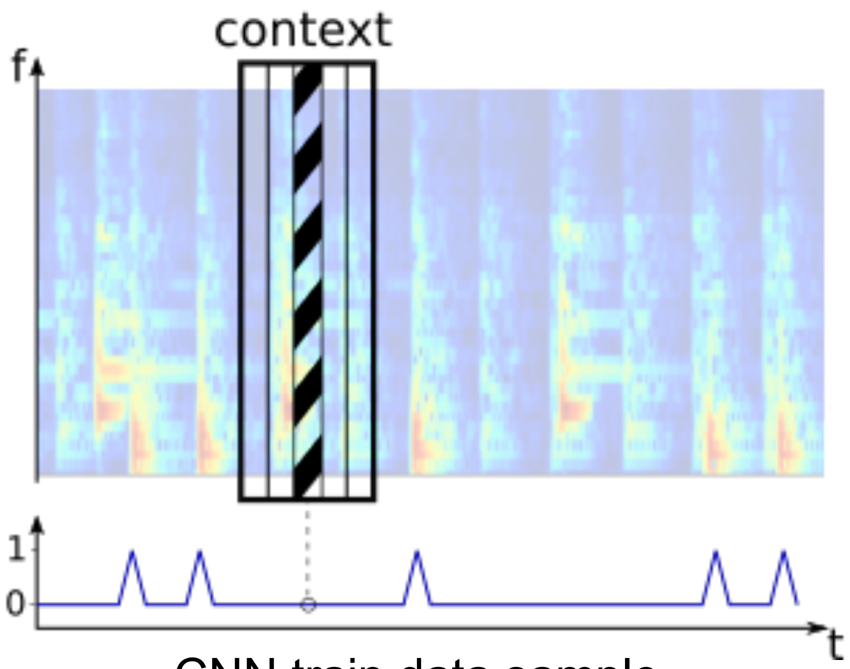
- Operate on **small windows** of spectrogram (current frame + **spectral context**)



CNN train data sample

NETWORK MODELS — CNN

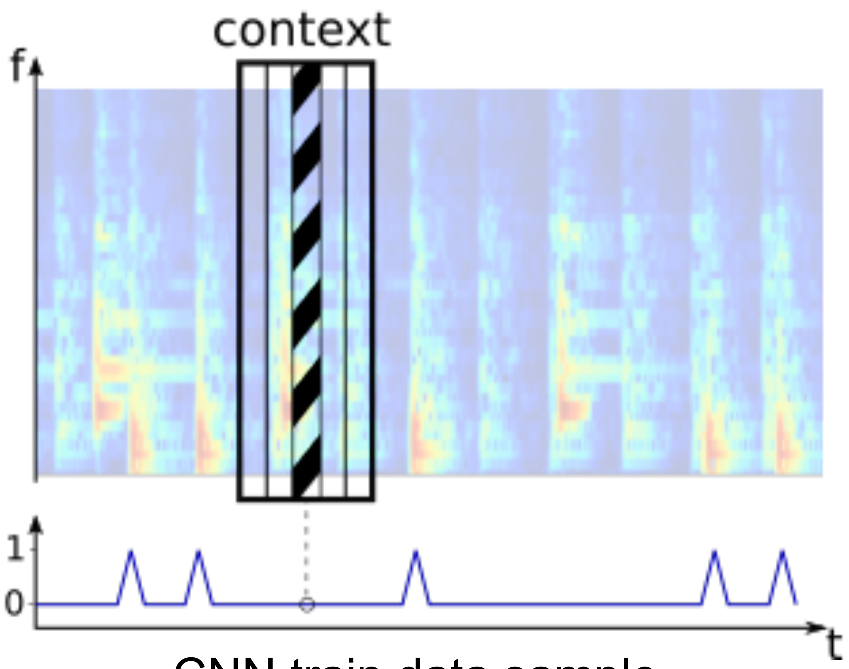
- Operate on **small windows** of spectrogram (current frame + **spectral context**)
- **Acoustic modeling** of drum sounds



CNN train data sample

NETWORK MODELS — CNN

- Operate on **small windows** of spectrogram (current frame + **spectral context**)
- **Acoustic modeling** of drum sounds



CNN train data sample

VGG - style architecture:

- 2 x conv: 32 x 3x3 (batch norm)
- max pool: 3x3
- 2 x conv: 64 x 3x3 (batch norm)
- max pool: 3x3
- 2 x dense: 256 ReLU
- output: 3 sigmoid

NETWORK MODELS — CRNN

NETWORK MODELS — CRNN

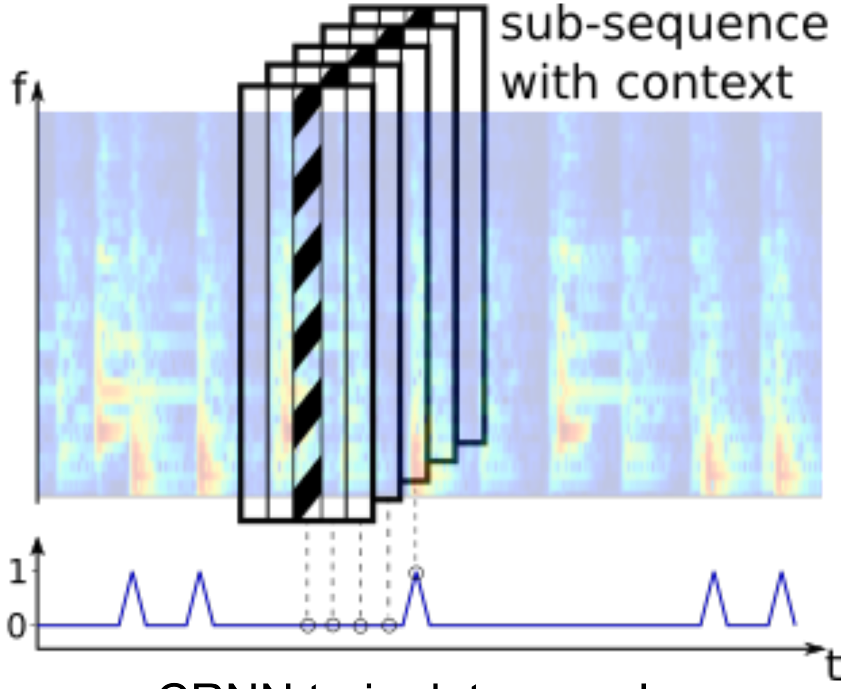
- Low-level CNN for **acoustic modeling**

NETWORK MODELS — CRNN

- Low-level CNN for **acoustic modeling**
- High-level RNN for ***music language model***

NETWORK MODELS — CRNN

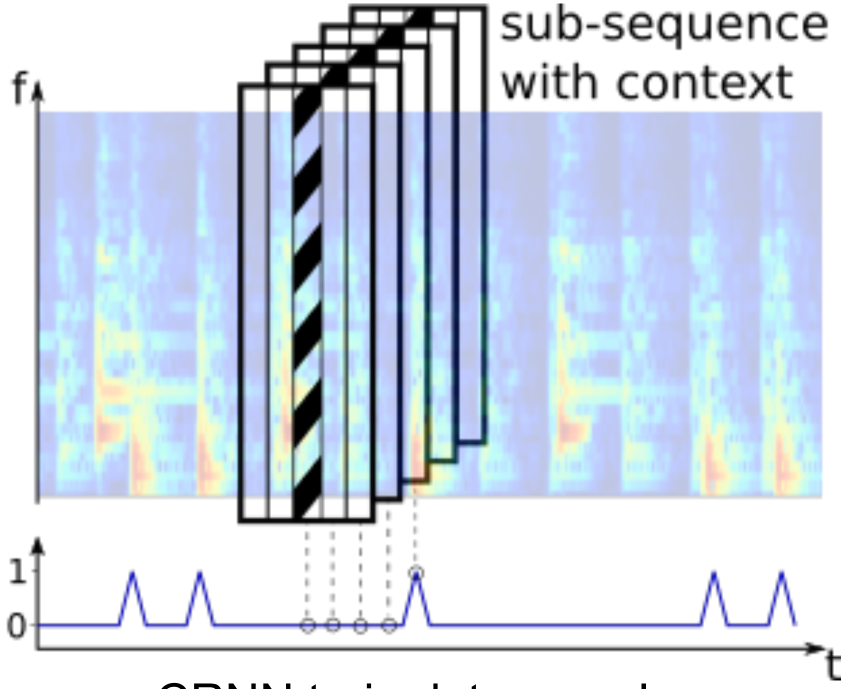
- Low-level CNN for **acoustic modeling**
- High-level RNN for ***music language model***



CRNN train data sample

NETWORK MODELS — CRNN

- Low-level CNN for **acoustic modeling**
- High-level RNN for ***music language model***



CRNN train data sample

stacked CNN + RNN architecture:

2 x conv: 32 x 3x3 (batch norm)

max pool: 3x3

2 x conv: 64 x 3x3 (batch norm)

max pool: 3x3

3 x RNN: 50 BD GRU

output: 3 sigmoid

WHY IS CONTEXT RELEVANT?

WHY IS CONTEXT RELEVANT?

- Instruments from the same class often **sound quite different**
  Similar sound for different instruments

snare drums: ♫

crash v.s. splash: ♫

WHY IS CONTEXT RELEVANT?

- Instruments from the same class often **sound quite different**
  Similar sound for different instruments

snare drums: ♫ crash v.s. splash: ♫

WHY IS CONTEXT RELEVANT?

- Instruments from the same class often **sound quite different**
  Similar sound for different instruments

snare drums: ♫ crash v.s. splash: ♫

WHY IS CONTEXT RELEVANT?

- Instruments from the same class often **sound quite different**
  Similar sound for different instruments

snare drums: crash v.s. splash:

- When **humans** transcribe drums
  - **Function** in a track equally important (snare drum v.s. backbeat)

WHY IS CONTEXT RELEVANT?

- Instruments from the same class often **sound quite different**
  Similar sound for different instruments

snare drums: ♫ crash v.s. splash: ♫

- When **humans** transcribe drums
  - **Function** in a track equally important (snare drum v.s. backbeat)
  - **Inaudible** onsets will be filled in if **expected**

WHY IS CONTEXT RELEVANT?

- Instruments from the same class often **sound quite different**
  Similar sound for different instruments

snare drums: crash v.s. splash:

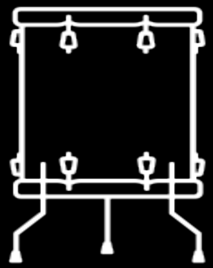
- When **humans** transcribe drums
  - **Function** in a track equally important (snare drum v.s. backbeat)
  - **Inaudible** onsets will be filled in if **expected**

- ***Music Language Model***

BASS DRUM OR LOW TOM?

♫ ♫

1: bass drum



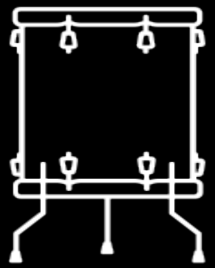
2: floor tom

3: ? ? ?

BASS DRUM OR LOW TOM?

♫ ♫

1: bass drum



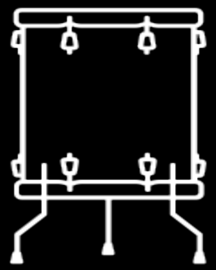
2: floor tom

3: ? ? ?

BASS DRUM OR LOW TOM?

♫ ♫

1: bass drum



2: floor tom

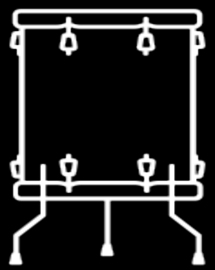
3: ? ? ?

BASS DRUM OR LOW TOM?

♫ ♫ ♫

context

1: bass drum



2: floor tom

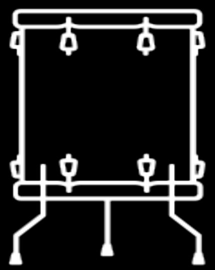
3: ? ? ?

BASS DRUM OR LOW TOM?

♫ ♫ ♫

context

1: bass drum



2: floor tom

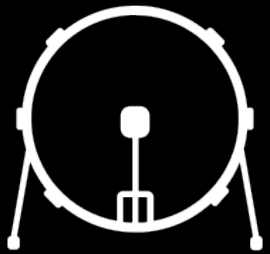
?

3: ? ? ?

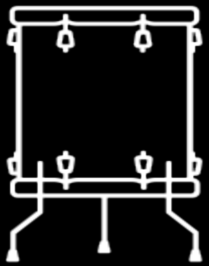
BASS DRUM OR LOW TOM?

♫ ♫ ♫

context



1: bass drum



2: floor tom

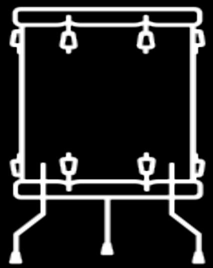
3: bass drum

BASS DRUM OR LOW TOM?

♫ ♫ ♫

context

1: bass drum



2: floor tom

DATASETS

DATASETS

- **IDMT-SMT-Drums** [Dittmar and Gärtner 2014]
  - **Solo** drum tracks, recorded, synthesized, and sampled
  - 95 tracks, total: **24m**, onsets: 8004

DATASETS

SMT (simple!)

- **IDMT-SMT-Drums** [Dittmar and Gärtner 2014]
  - **Solo** drum tracks, recorded, synthesized, and sampled
  - 95 tracks, total: **24m**, onsets: 8004

♫

DATASETS

SMT (simple!)

- **IDMT-SMT-Drums** [Dittmar and Gärtner 2014]
  - **Solo** drum tracks, recorded, synthesized, and sampled
  - 95 tracks, total: **24m**, onsets: 8004
- **ENST-Drums** [Gillet and Richard 2006]
  - Recordings, three drummers on different drum kits, **optional accompaniment**
  - 64 tracks, total: **1h**, onsets: 22391

DATASETS

SMT (simple!)

- **IDMT-SMT-Drums** [Dittmar and Gärtner 2014]
  - **Solo** drum tracks, recorded, synthesized, and sampled
  - 95 tracks, total: **24m**, onsets: 8004

♫

- **ENST-Drums** [Gillet and Richard 2006]
  - Recordings, three drummers on different drum kits, **optional accompaniment**
  - 64 tracks, total: **1h**, onsets: 22391

ENST solo ♫ ♫

(harder!)

DATASETS

SMT (simple!)

- **IDMT-SMT-Drums** [Dittmar and Gärtner 2014]
  - **Solo** drum tracks, recorded, synthesized, and sampled
  - 95 tracks, total: **24m**, onsets: 8004
- **ENST-Drums** [Gillet and Richard 2006]
  - Recordings, three drummers on different drum kits, **optional accompaniment**
  - 64 tracks, total: **1h**, onsets: 22391

ENST solo (harder!)

ENST acc. (difficult!)

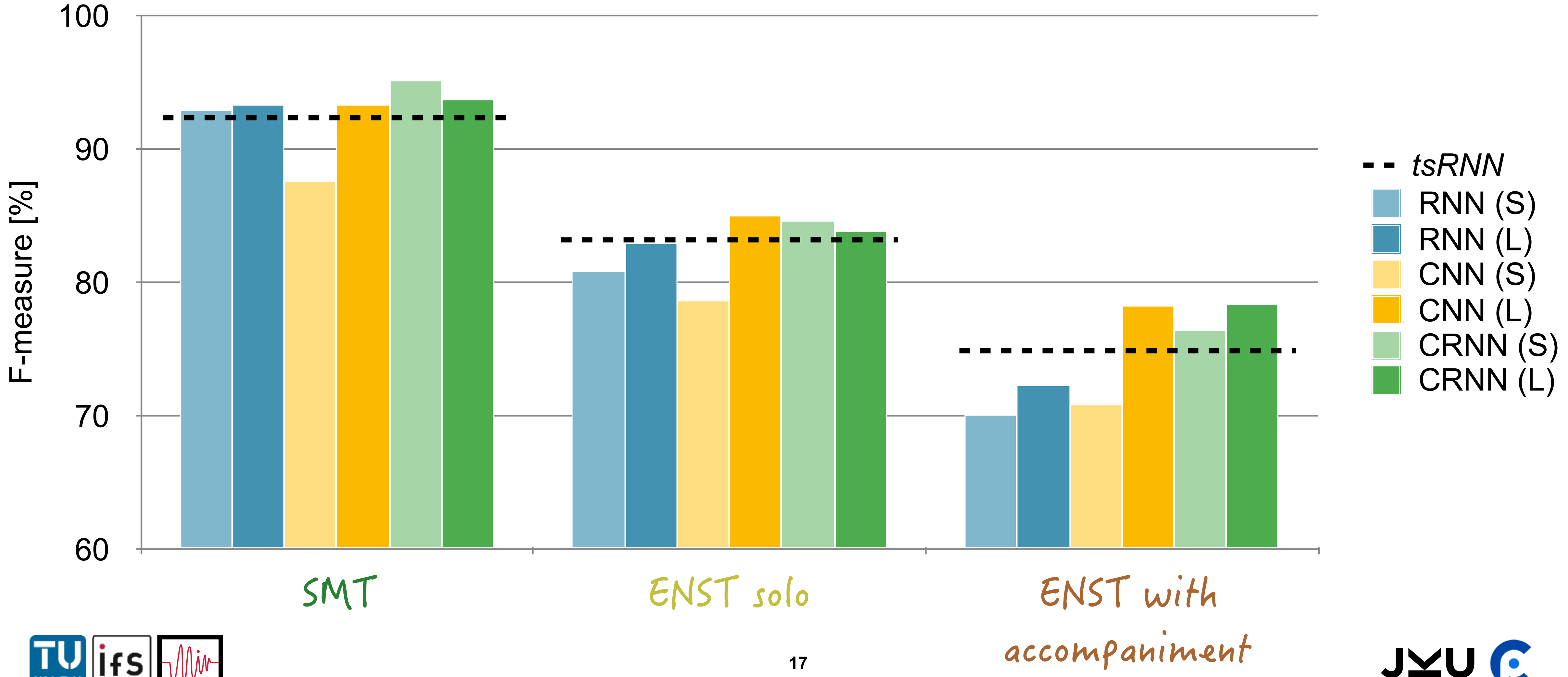
NETWORK MODELS

Architecture

	Frames	Context	Conv. Layers	Rec. Layers	Dense Layers
RNN (S)	100	—	—	2x50 GRU	—
RNN (L)	400	—	—	3x30 GRU	—
CNN (S)	—	9	2 x 32 3x3 filt. 3x3 max pooling 2 x 64 3x3 filt. 3x3 max pooling all w/ batch norm.	—	2x256
CNN (L)	—	25		—	2x256
CRNN (S)	100	9		2x50 GRU	—
CRNN (L)	400	13		3x60 GRU	—
tsRNN	*baseline* [Vogl et al. ICASSP'17]				

- Early stopping
- Batch normalization
- L2 norm
- Dropout
- ADAM optimizer

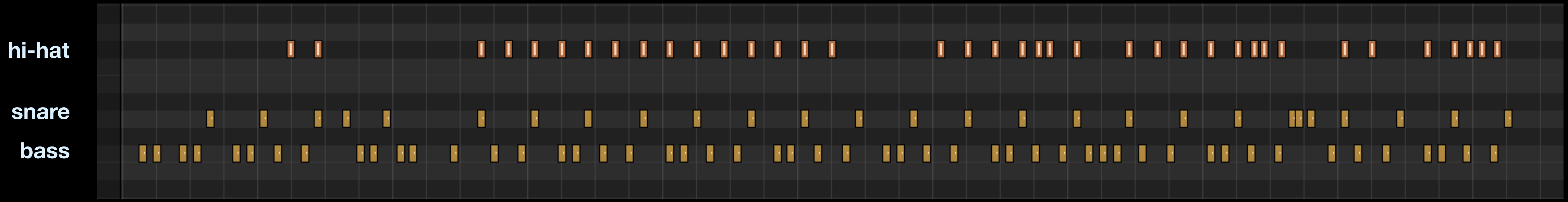
RESULTS



100
90
80
70
60
F-measure [%]
tsRNN
RNN (S)
RNN (L)
CNN (S)
CNN (L)
CRNN (S)
CRNN (L)
SMT
ENST solo
ENST with accompaniment

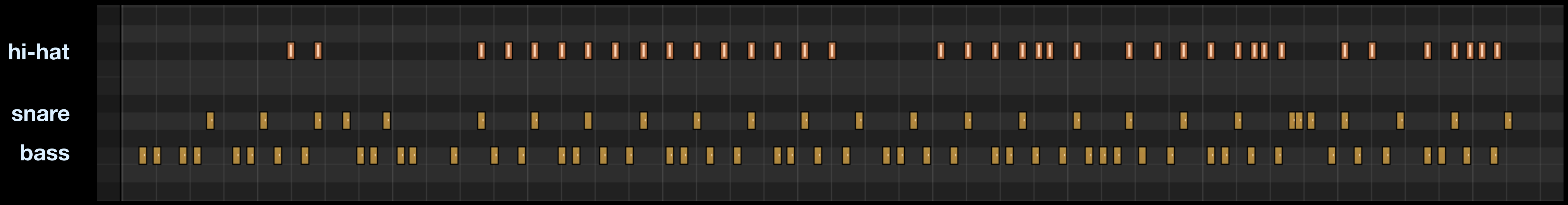
HOW DOES IT SOUND?

“Punk” MEDLEY DB



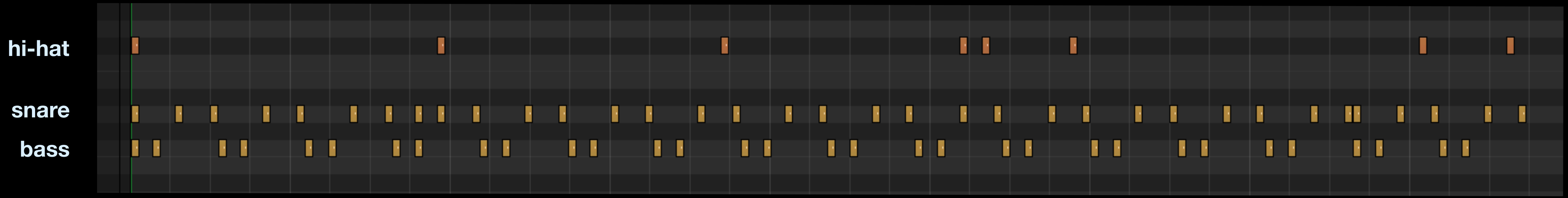
HOW DOES IT SOUND?

"Punk" MEDLEY DB



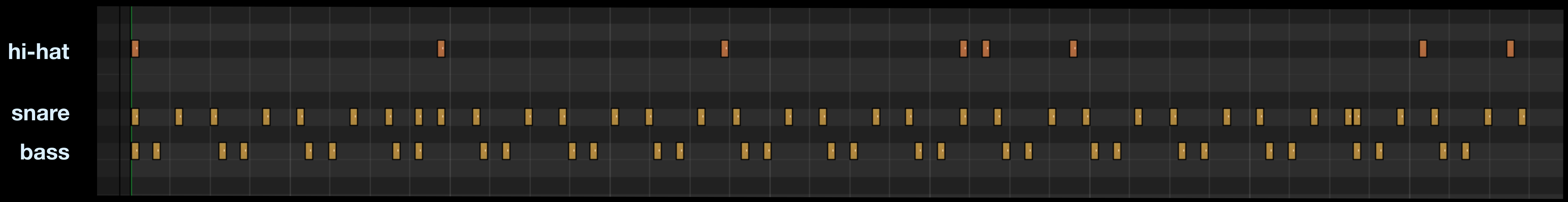
HOW DOES IT SOUND?

“Hendrix” MEDLEY DB



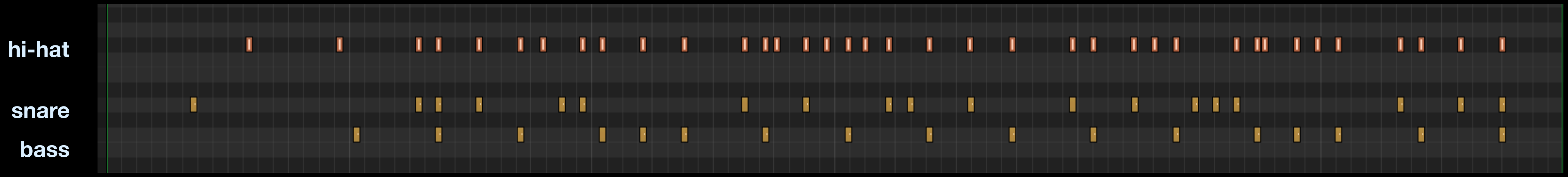
HOW DOES IT SOUND?

“Hendrix” MEDLEY DB



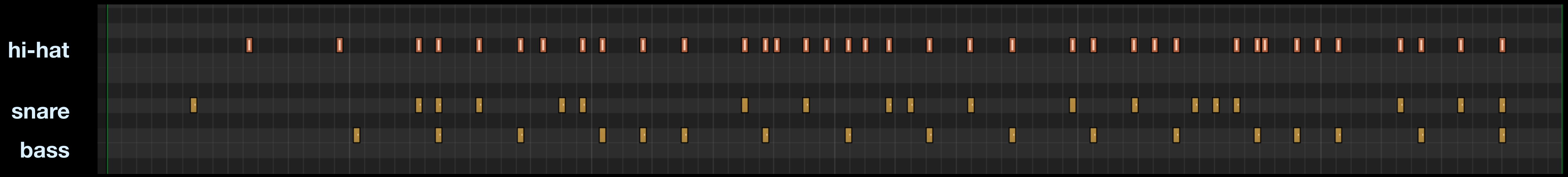
HOW DOES IT SOUND?

Alexa, play some music…



HOW DOES IT SOUND?

Alexa, play some music…



PART 1
AUTOMATIC DRUM TRANSCRIPTION

Task Definition, Problem Modeling, Architectures

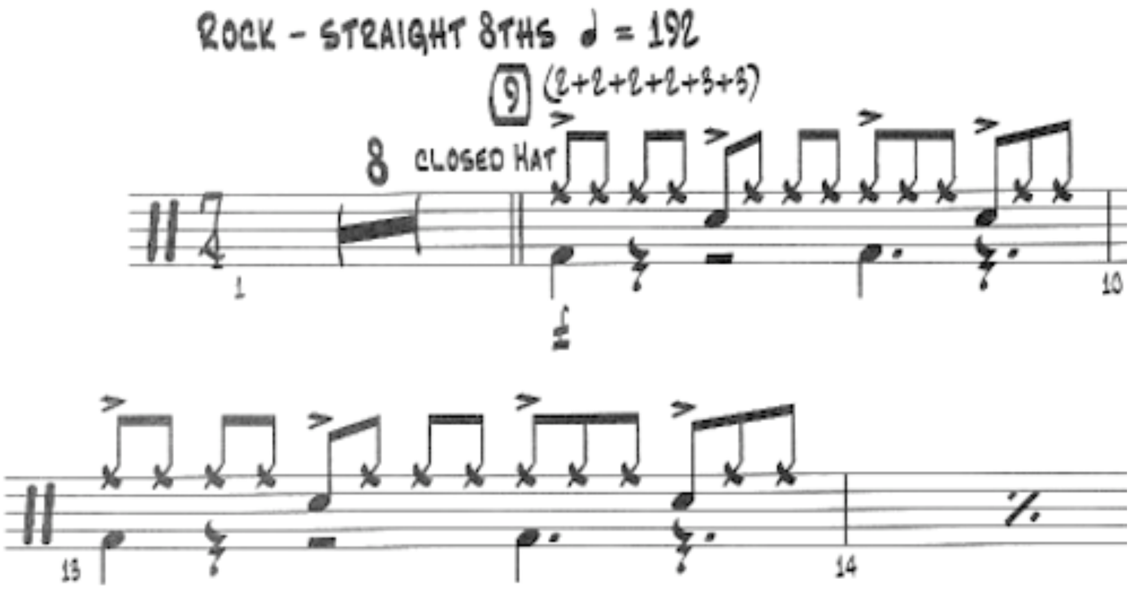
PART 2
MULTI-TASK LEARNING

Metadata for Transcripts

LIMITATIONS OF CURRENT SYSTEMS

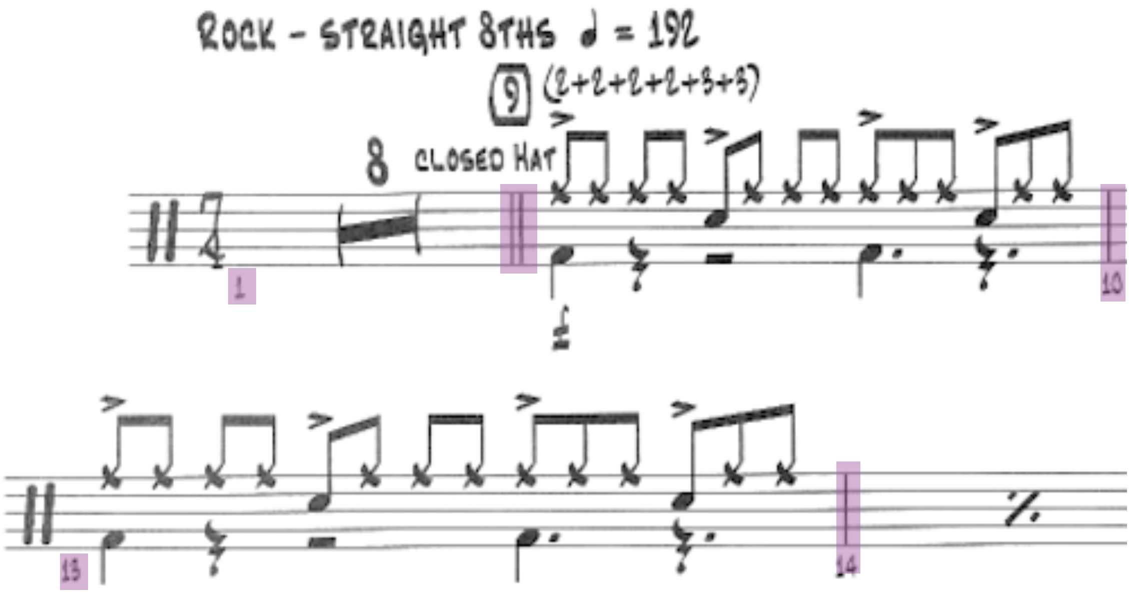
LIMITATIONS OF CURRENT SYSTEMS

- Do not produce additional information for transcripts
  drum onset detection vs ***drum transcription***



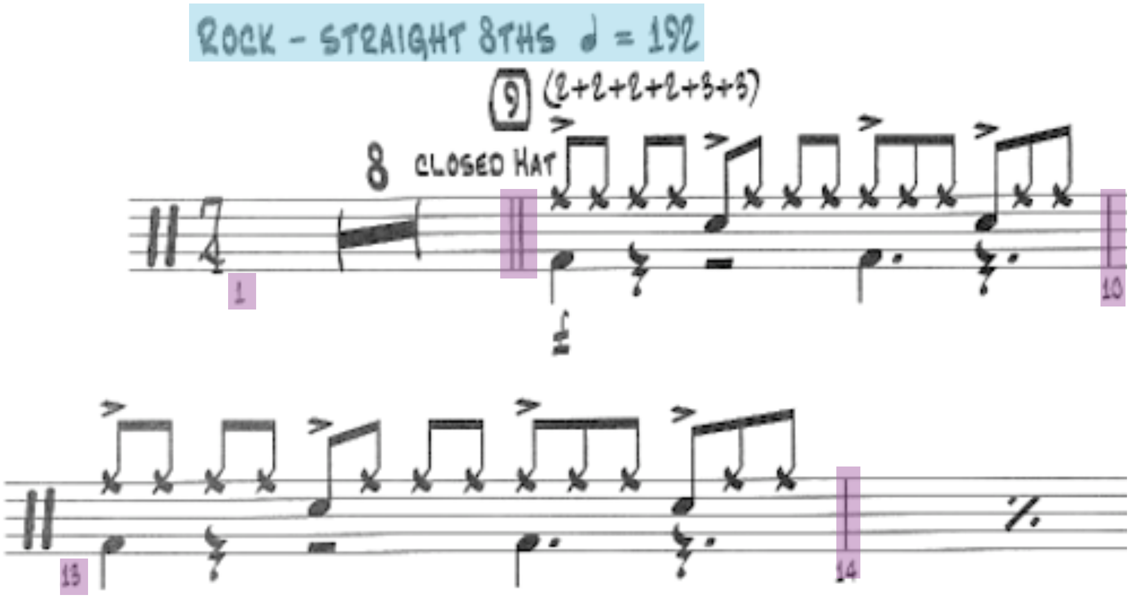
LIMITATIONS OF CURRENT SYSTEMS

- Do not produce additional information for transcripts
  drum onset detection vs ***drum transcription***
  - bars lines



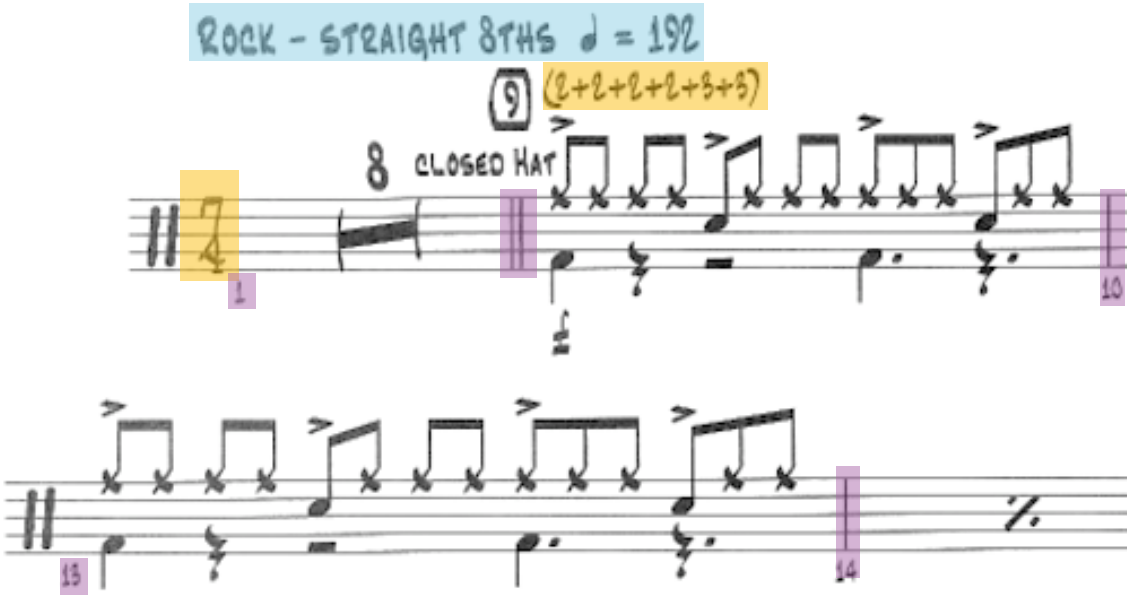
LIMITATIONS OF CURRENT SYSTEMS

- Do not produce additional information for transcripts
  drum onset detection vs ***drum transcription***
  - bars lines
  - tempo



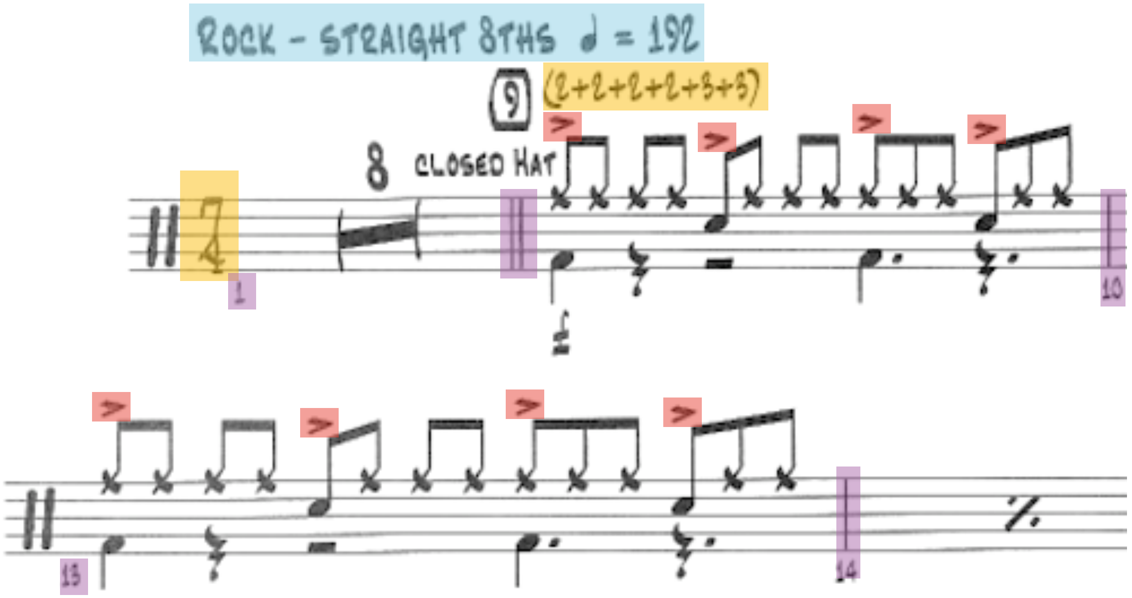
LIMITATIONS OF CURRENT SYSTEMS

- Do not produce additional information for transcripts
  drum onset detection vs ***drum transcription***
  - bars lines
  - tempo
  - meter



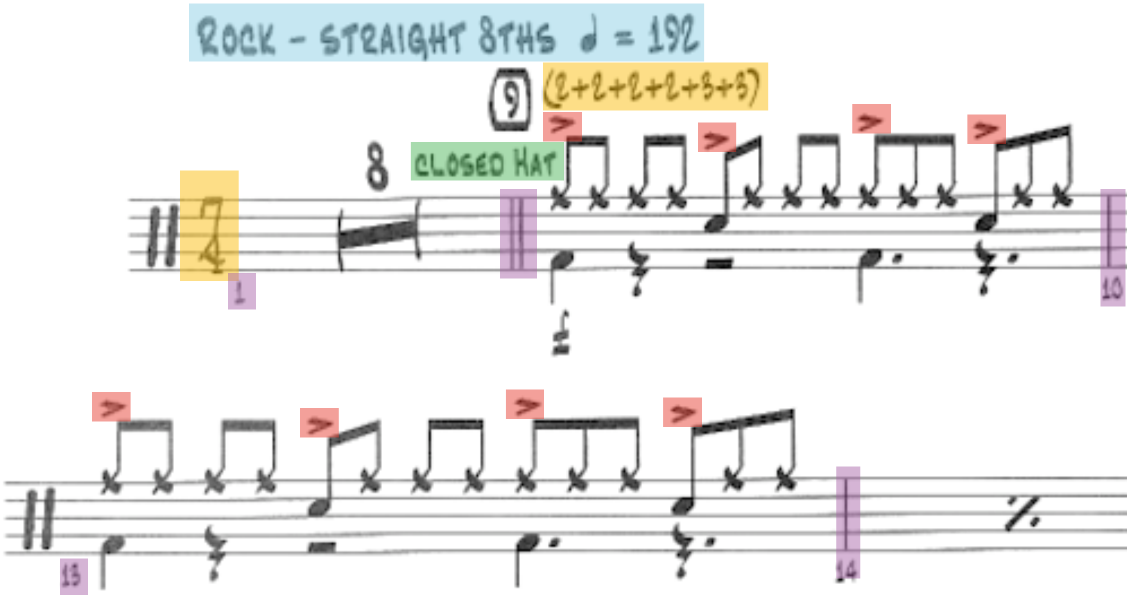
LIMITATIONS OF CURRENT SYSTEMS

- Do not produce additional information for transcripts
  drum onset detection vs ***drum transcription***
  - bars lines
  - tempo
  - meter
  - dynamics / accents



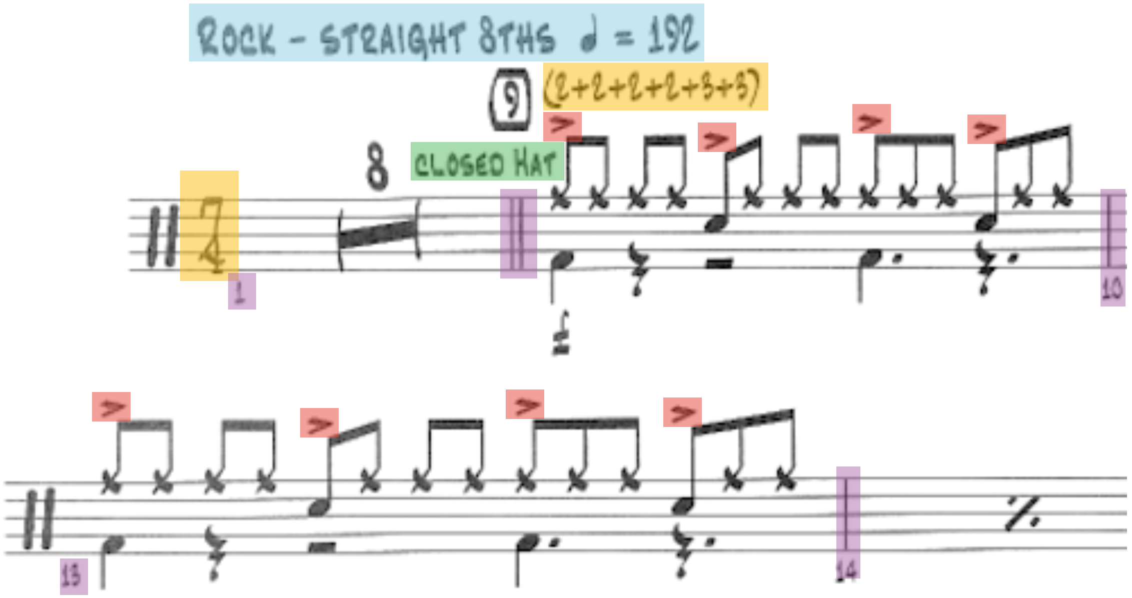
LIMITATIONS OF CURRENT SYSTEMS

- Do not produce additional information for transcripts
  drum onset detection vs ***drum transcription***
  - bars lines
  - tempo
  - meter
  - dynamics / accents
  - stroke / playing technique



LIMITATIONS OF CURRENT SYSTEMS

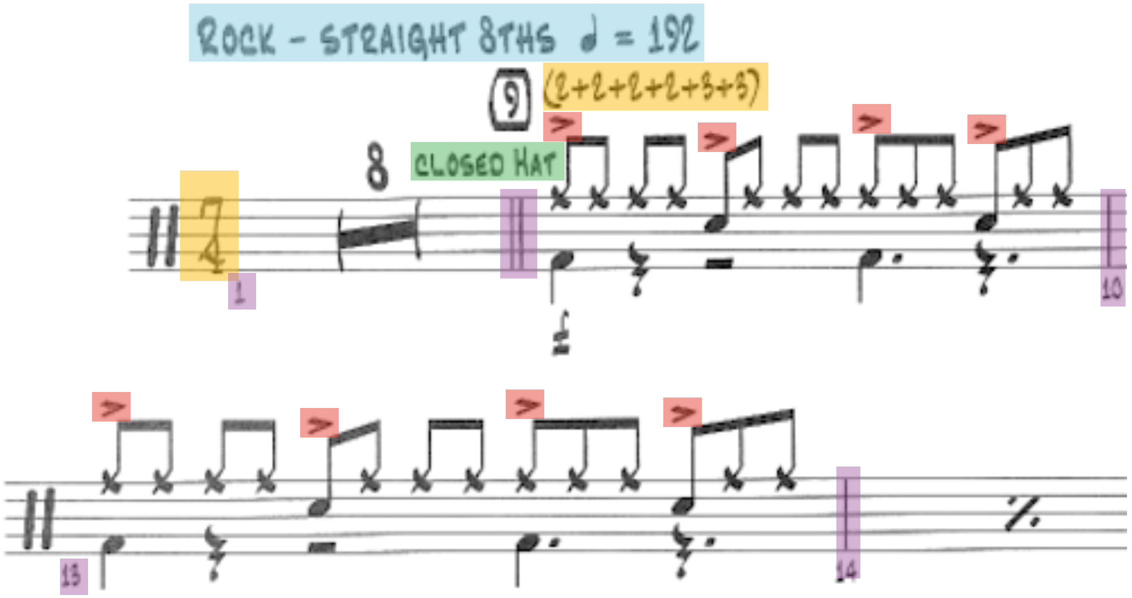
- Do not produce additional information for transcripts
  drum onset detection vs ***drum transcription***
  - bars lines
  - tempo
  - meter
  - dynamics / accents
  - stroke / playing technique
- Only three instrument classes



Richard Vogl, Gerhard Widmer, and Peter Knees, "**Towards multi-instrument drum transcription**," in *Proc. 21th Intl. Conf. on Digital Audio Effects (DAFx18), Aveiro, Portugal, Sep. 2018.*

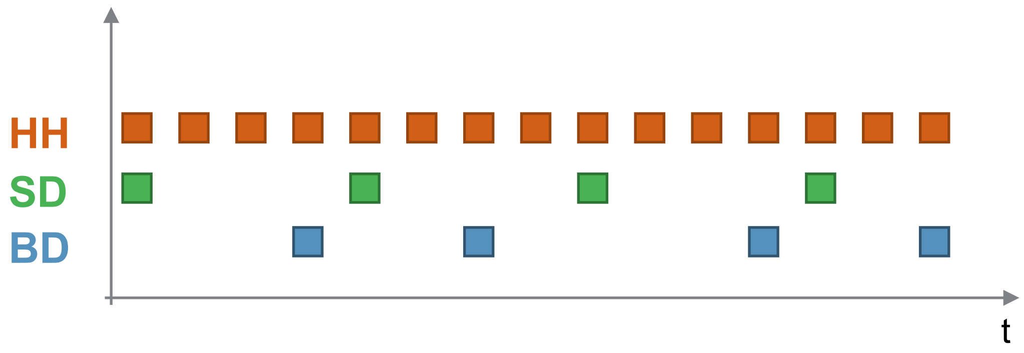
LIMITATIONS OF CURRENT SYSTEMS

- Do not produce additional information for transcripts
  drum onset detection vs ***drum transcription***
  - bars lines
  - tempo
  - meter
  - dynamics / accents
  - stroke / playing technique
- Only three instrument classes

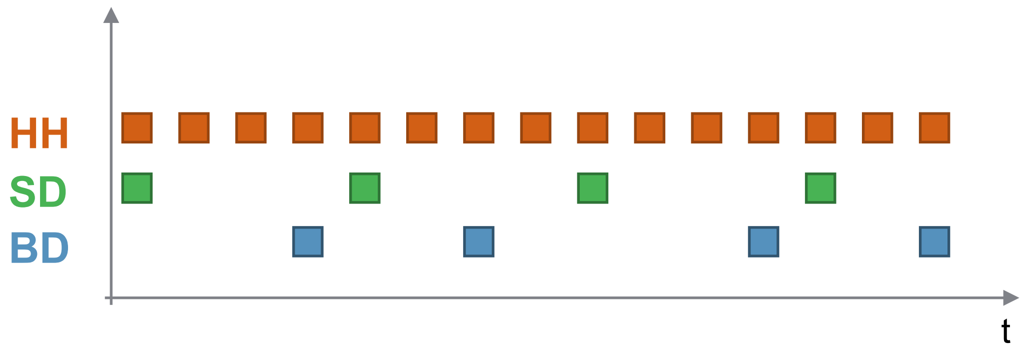


Richard Vogl, Gerhard Widmer, and Peter Knees, "**Towards multi-instrument drum transcription**," in *Proc. 21th Intl. Conf. on Digital Audio Effects (DAFx18), Aveiro, Portugal, Sep. 2018.*

ADDITIONAL INFORMATION FOR TRANSCRIPTS

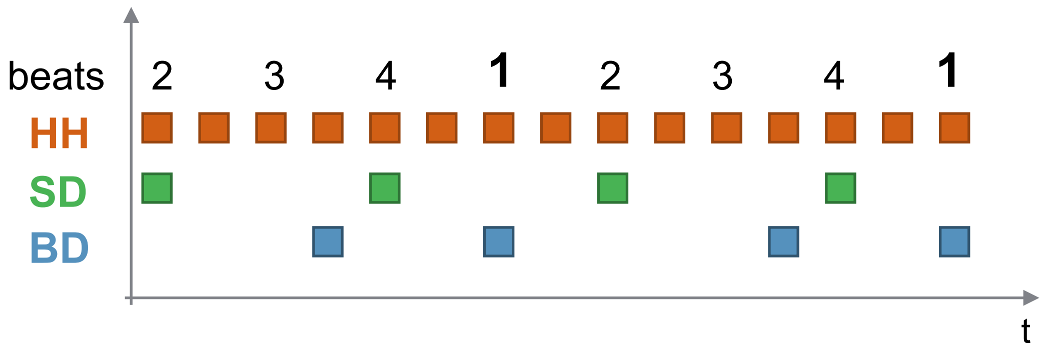


ADDITIONAL INFORMATION FOR TRANSCRIPTS



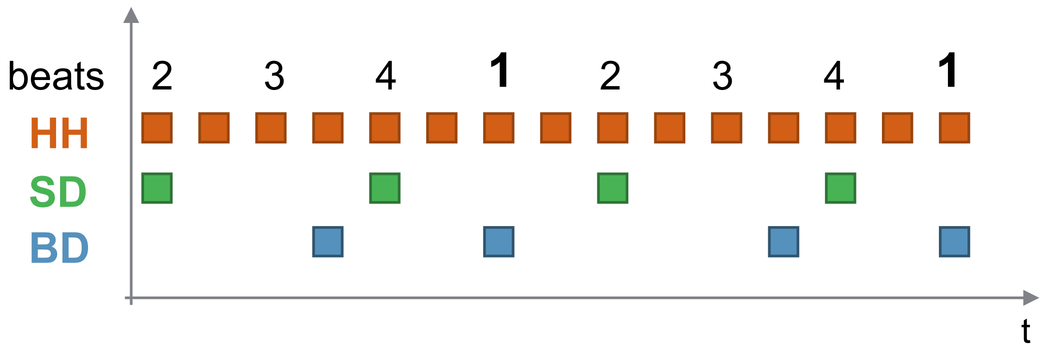
ADDITIONAL INFORMATION FOR TRANSCRIPTS

- Use **beat and downbeat tracking** to get:



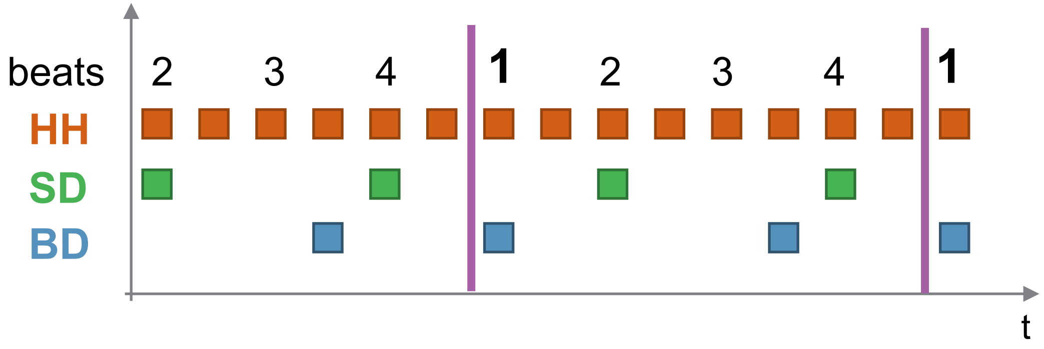
ADDITIONAL INFORMATION FOR TRANSCRIPTS

- Use **beat and downbeat tracking** to get:



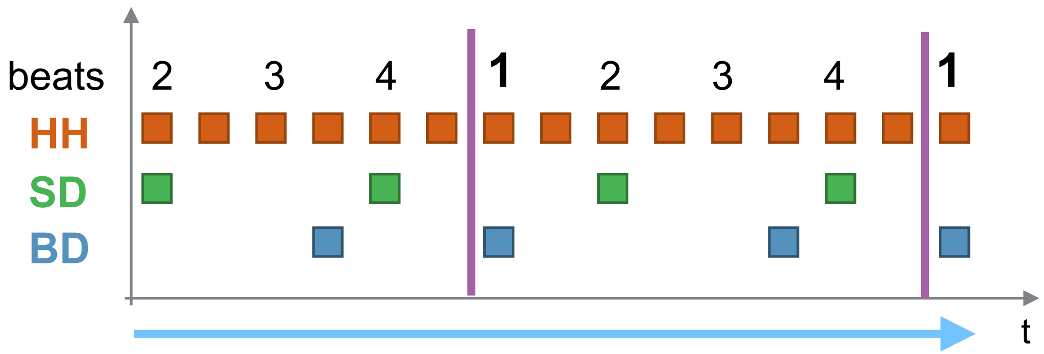
ADDITIONAL INFORMATION FOR TRANSCRIPTS

- Use **beat and downbeat tracking** to get:
  - bars lines



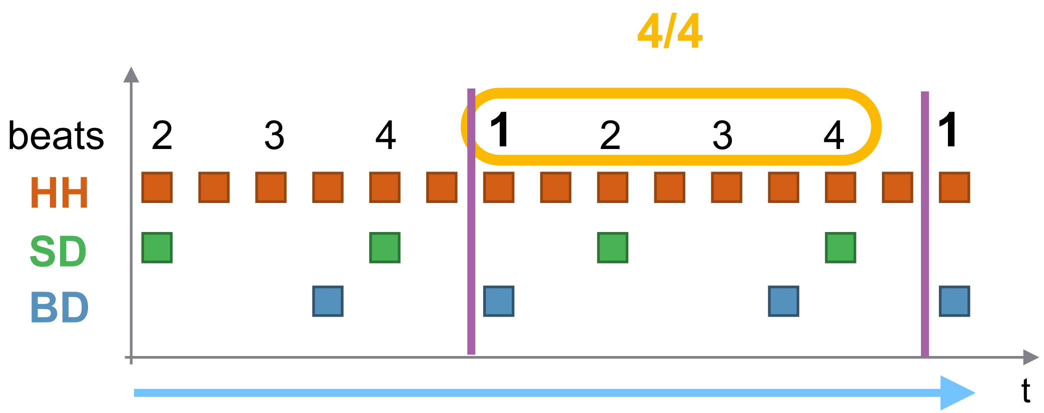
ADDITIONAL INFORMATION FOR TRANSCRIPTS

- Use **beat and downbeat tracking** to get:
  - bars lines
  - tempo



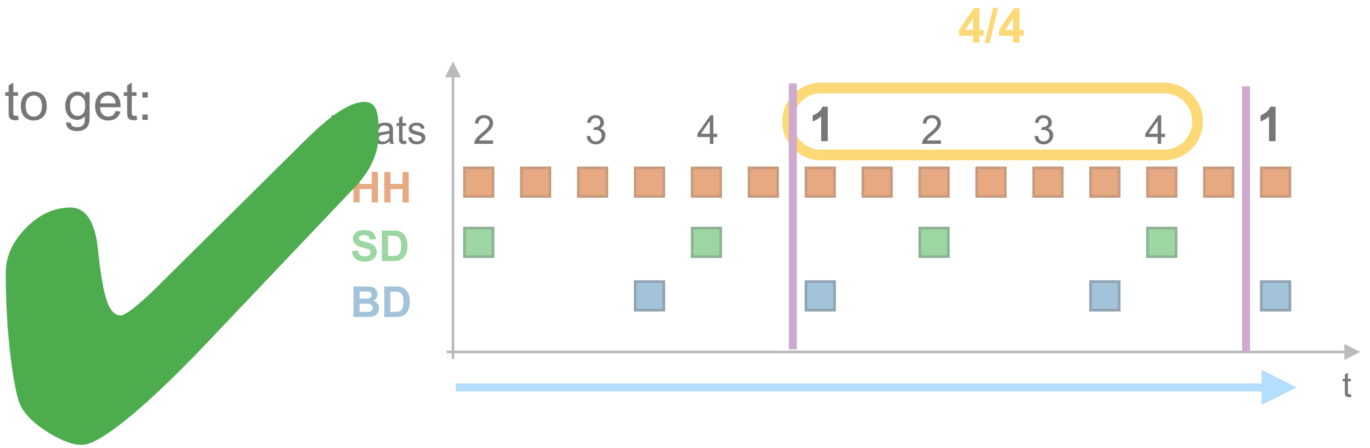
ADDITIONAL INFORMATION FOR TRANSCRIPTS

- Use **beat and downbeat tracking** to get:
  - bars lines
  - tempo
  - meter

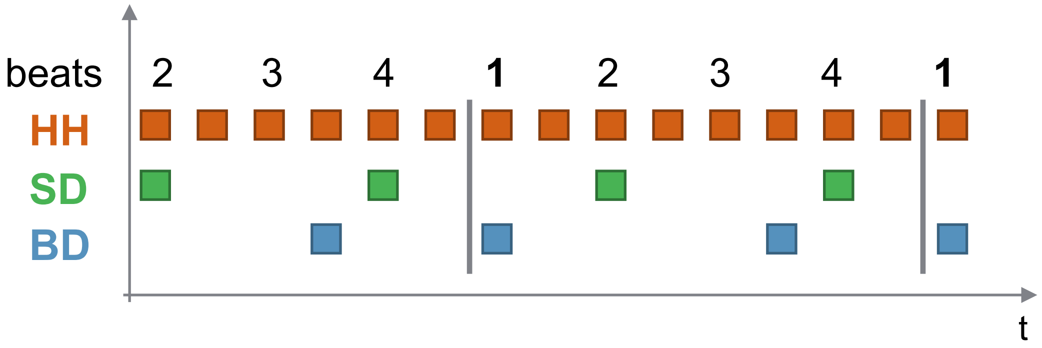


ADDITIONAL INFORMATION FOR TRANSCRIPTS

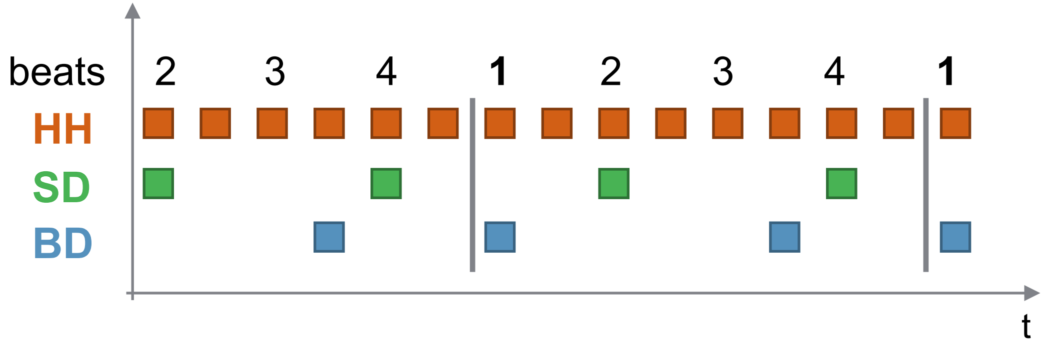
- Use **beat and downbeat tracking** to get:
  - bars lines
  - tempo
  - meter



LEVERAGE BEAT INFORMATION

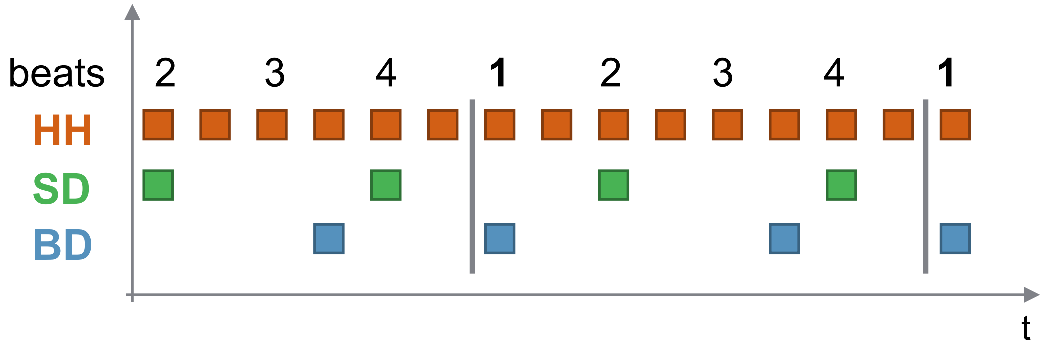


LEVERAGE BEAT INFORMATION



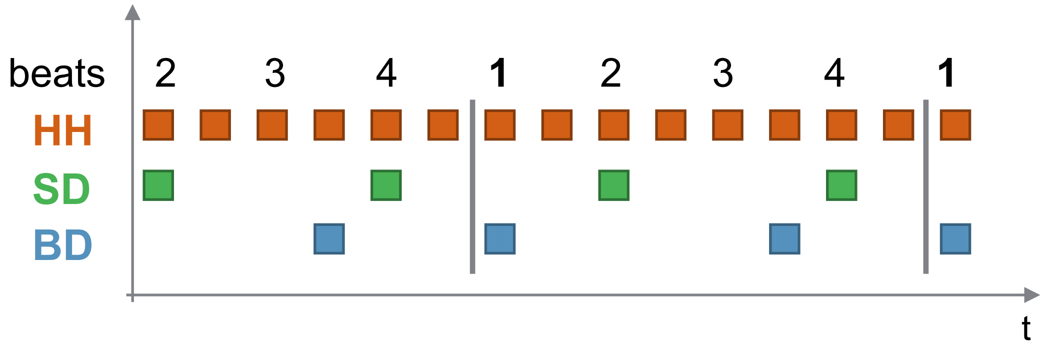
- Beats are **highly correlated** with drum patterns (drum onset locations / repetitive patterns)

LEVERAGE BEAT INFORMATION



- Beats are **highly correlated** with drum patterns
(drum onset locations / repetitive patterns)
- Assume that **prior knowledge** of beats is helpful for drum transcription

LEVERAGE BEAT INFORMATION



- Beats are **highly correlated** with drum patterns (drum onset locations / repetitive patterns)
- Assume that **prior knowledge** of beats is helpful for drum transcription
- Use **multi-task learning** for beats and drums

MULTI-TASK LEARNING

MULTI-TASK LEARNING

- Training one model to solve **multiple related tasks**

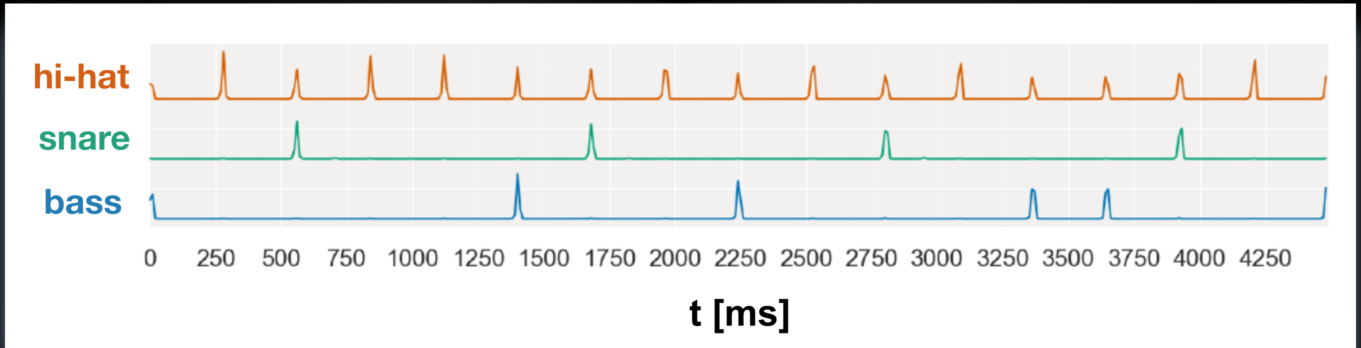
MULTI-TASK LEARNING

- Training one model to solve **multiple related tasks**
  - **Improve performance** for each subtask ➡ context!

MULTI-TASK LEARNING

- Training one model to solve **multiple related tasks**
  - **Improve performance** for each subtask ➡ context!

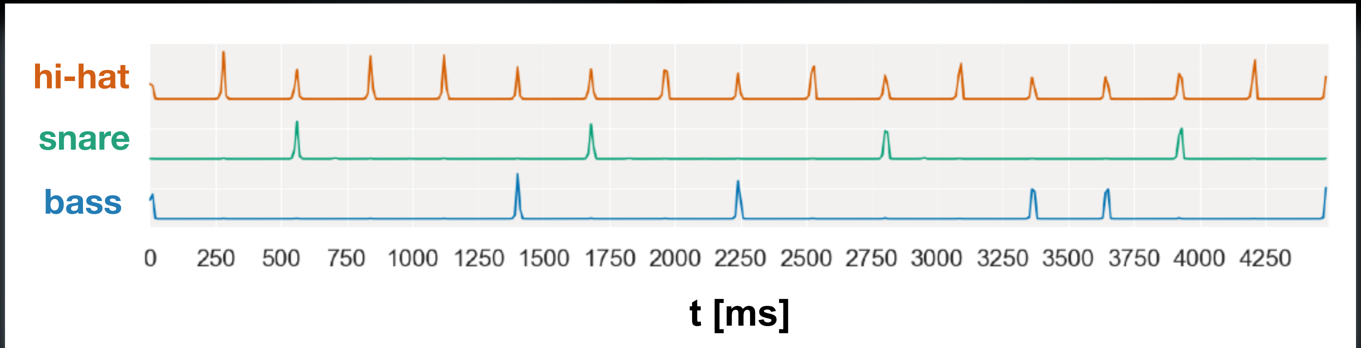
- Individual activation functions are already **learned using multi-task learning**



MULTI-TASK LEARNING

- Training one model to solve **multiple related tasks**
  - **Improve performance** for each subtask ➡ context!

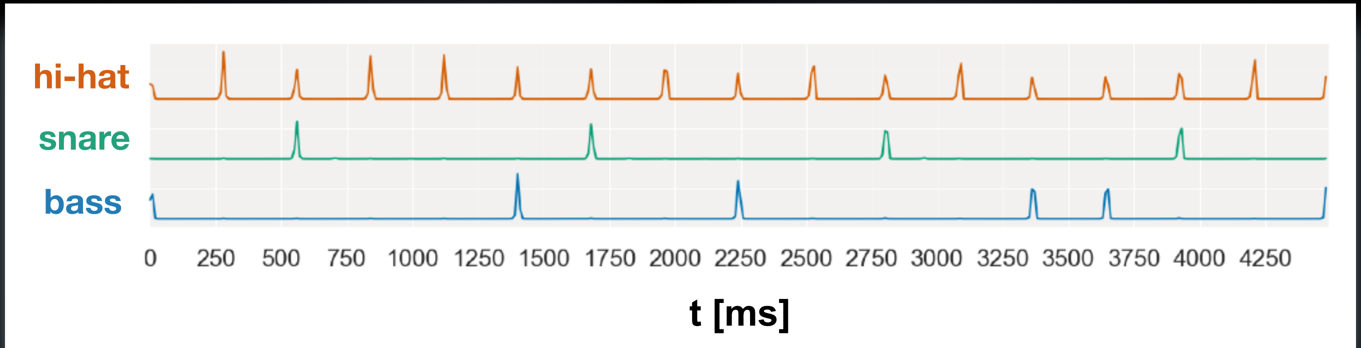
- Individual activation functions are already **learned using multi-task learning**
  - One network for all **instruments**



MULTI-TASK LEARNING

- Training one model to solve **multiple related tasks**
  - **Improve performance** for each subtask ➡ context!

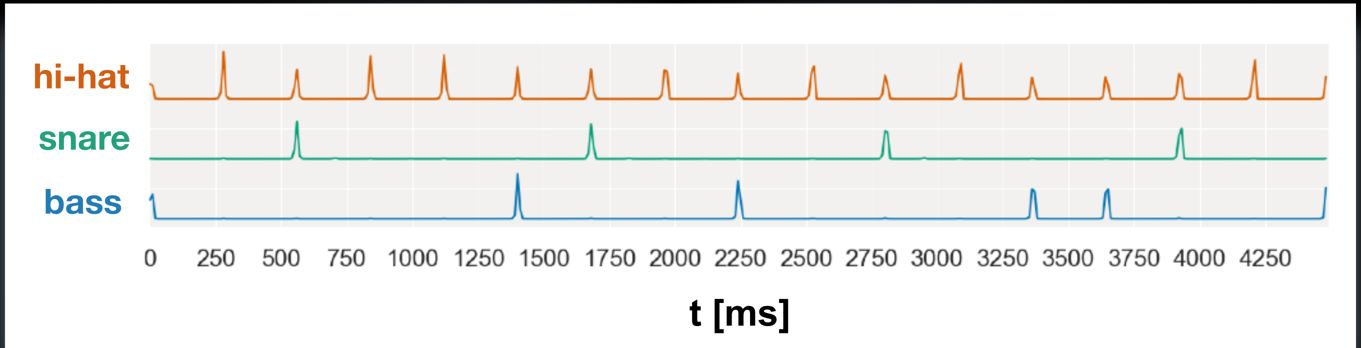
- Individual activation functions are already **learned using multi-task learning**
  - One network for all **instruments**
  - Instrument onsets are **not independent**



MULTI-TASK LEARNING

- Training one model to solve **multiple related tasks**
  - **Improve performance** for each subtask ➡ context!

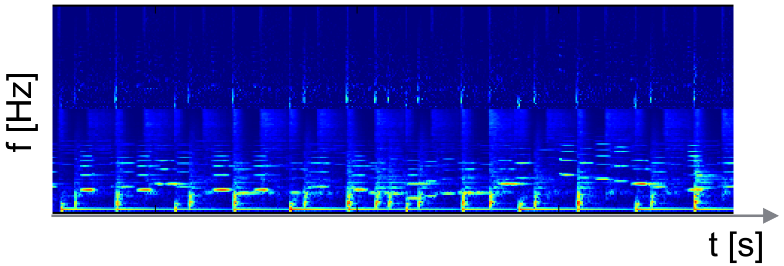
- Individual activation functions are already **learned using multi-task learning**
  - One network for all **instruments**
  - Instrument onsets are **not independent**
  - MIREX results show that **it works better**



MULTI-TASK EXPERIMENTS

input

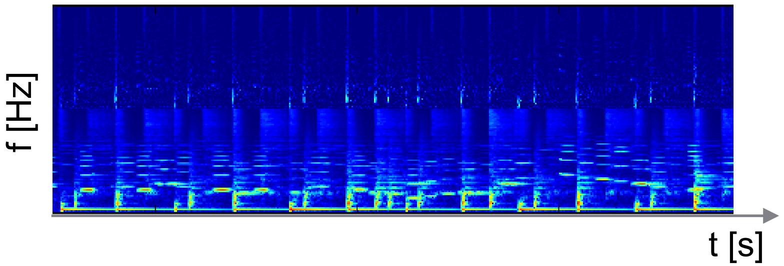
output



MULTI-TASK EXPERIMENTS

input

output

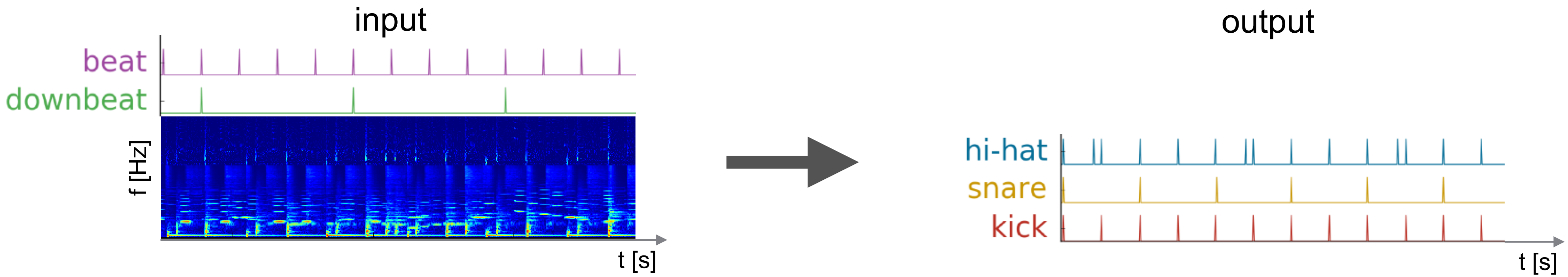


- Three experiments:

MULTI-TASK EXPERIMENTS

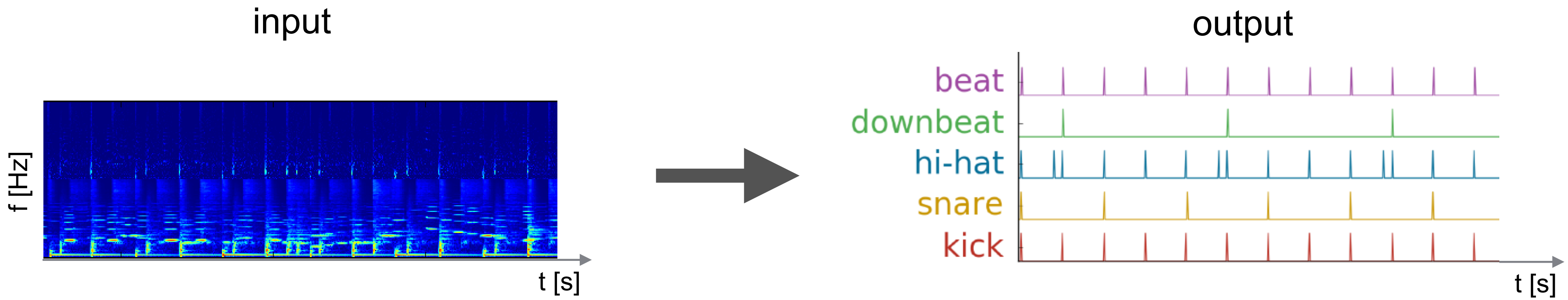
- Three experiments:
  - Training on drum targets (***DT***)

MULTI-TASK EXPERIMENTS



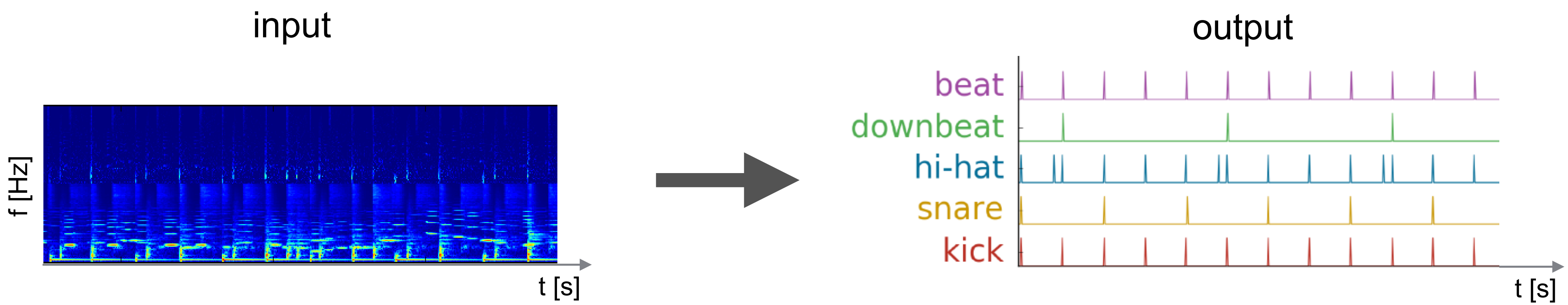
- Three experiments:
  - Training on drum targets (*DT*)
  - Training on drum targets with annotated beats as **additional input** features (*BF*)

MULTI-TASK EXPERIMENTS



- Three experiments:
  - Training on drum targets (*DT*)
  - Training on drum targets with annotated beats as **additional input** features (*BF*)
  - Training on drum and beat targets as **multi-task** problem (*MT*)

MULTI-TASK EXPERIMENTS



- Three experiments:
  - Training on drum targets (*DT*)
  - Training on drum targets with annotated beats as **additional input** features (*BF*)
  - Training on drum and beat targets as **multi-task** problem (*MT*)

- Expected increase in performance for *BF* compared to *DT*

MULTI-TASK EXPERIMENTS

- Three experiments:
  - Training on drum targets (***DT***)
  - Training on drum targets with annotated beats as **additional input** features (***BF***)
  - Training on drum and beat targets as **multi-task** problem (***MT***)

- Expected increase in performance for ***BF*** compared to ***DT***
- Desirable increase in performance for ***MT*** compared to ***DT***

NEW DATASETS

NEW!

RBMA13-Drums [http://ifs.tuwien.ac.at/~vogl/datasets/]

- Free music from the 2013 Red Bull Music Academy, different styles
- 27 tracks, total: **1h 43m**, onsets: 24365
- **drum, beat, and downbeat** annotations

♫ ♫

NEW DATASETS

NEW!

RBMA13-Drums [http://ifs.tuwien.ac.at/~vogl/datasets/]

- Free music from the 2013 Red Bull Music Academy, different styles
- 27 tracks, total: **1h 43m**, onsets: 24365
- **drum, beat, and downbeat** annotations

♫ ♫

RBMA
(super difficult!)

NEW DATASETS

NEW!

RBMA13-Drums [http://ifs.tuwien.ac.at/~vogl/datasets/]

- Free music from the 2013 Red Bull Music Academy, different styles
- 27 tracks, total: **1h 43m**, onsets: 24365
- **drum, beat, and downbeat** annotations

♫ ♫

RBMA
(super difficult!)

RESULTS

Model	DT	BF	MT
RNN (S)	59.8	63.6	64.6
RNN (L)	61.8	64.5	64.3
CNN (S)	66.2	66.7	63.3
CNN (L)	66.8	65.2	64.8
CRNN (S)	65.2	66.1	66.9
CRNN (L)	**67.3**	**68.4**	**67.2**

(Columns: Experiment)

% F-measure for drum onsets, tolerance: ±20ms, 3-fold cross-validation

DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS

RBMA
(super difficult!)

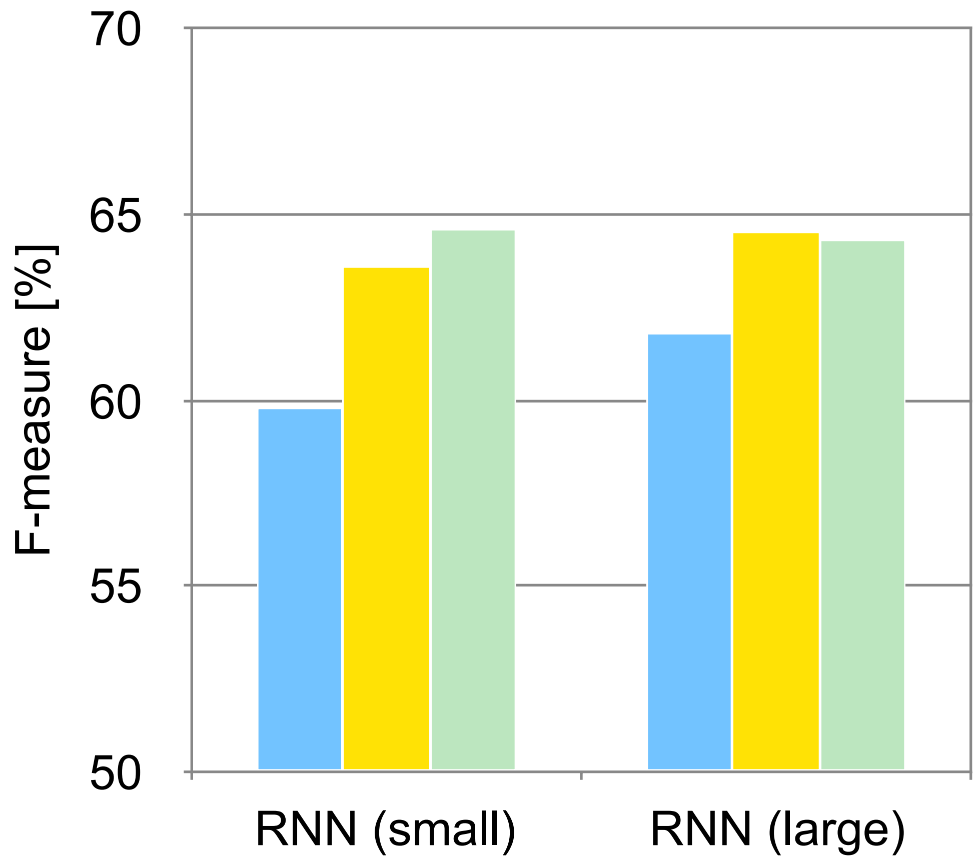
Experiment

Model	DT	BF	MT
RNN (S)	59.8	63.6	64.6
RNN (L)	61.8	64.5	64.3
CNN (S)	66.2	66.7	63.3
CNN (L)	66.8	65.2	64.8
CRNN (S)	65.2	66.1	66.9
CRNN (L)	**67.3**	**68.4**	**67.2**

% F-measure for drum onsets, tolerance: ±20ms, 3-fold cross-validation

DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

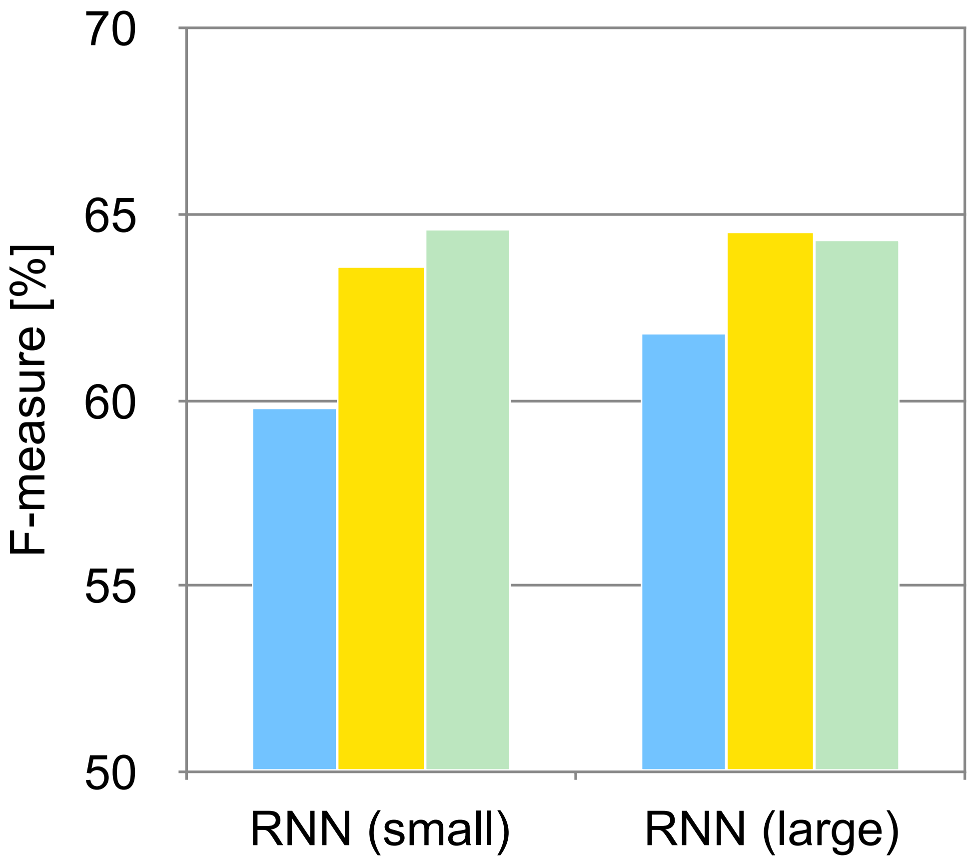
RESULTS: RNNs



DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS: RNNs

Impact of **beats for RNN**s:

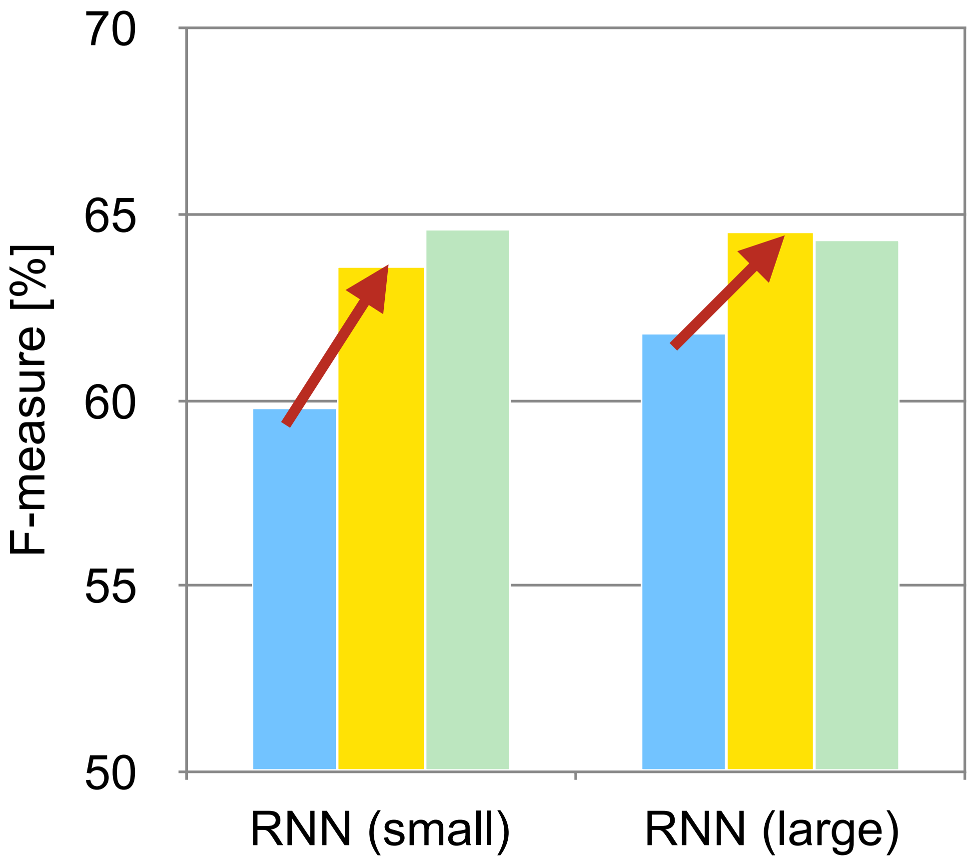


DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS: RNNs

Impact of **beats for RNN**s:

- BF improves for both models ✔

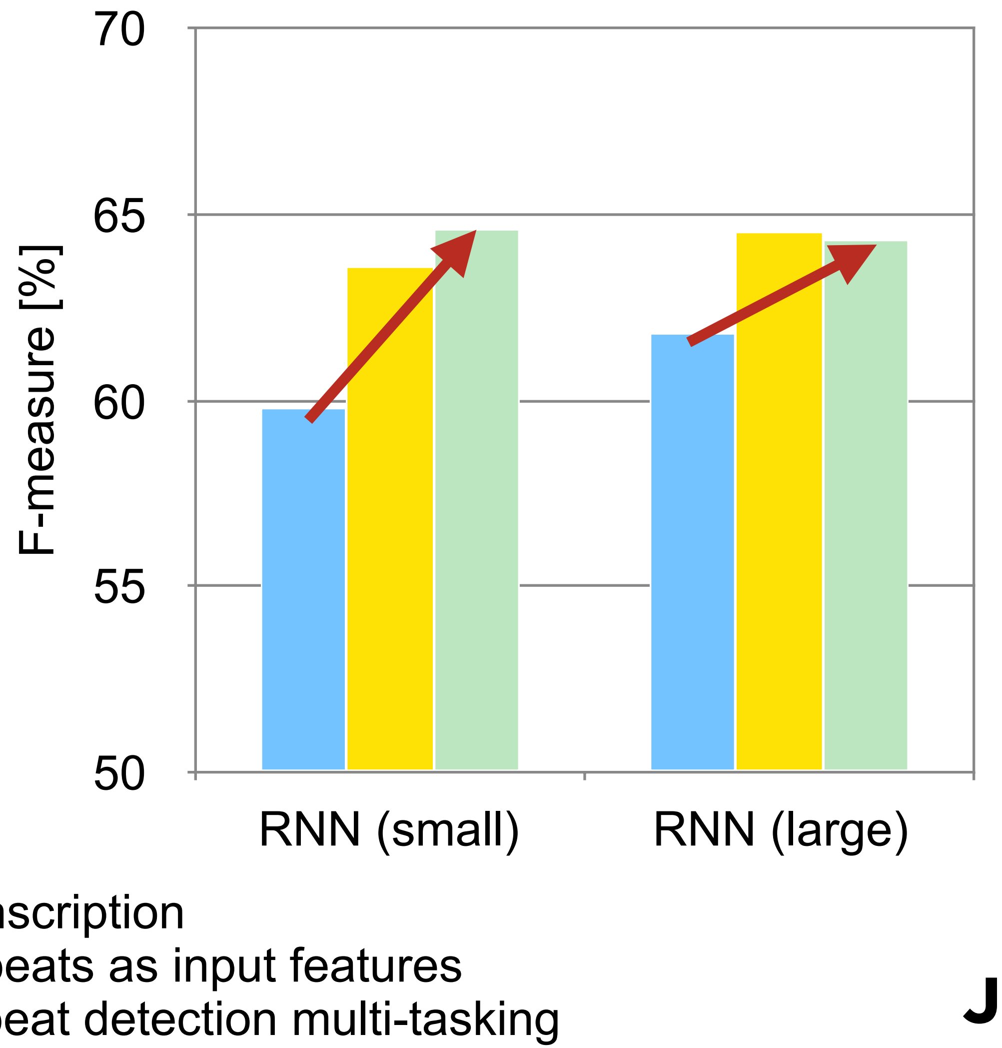


DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

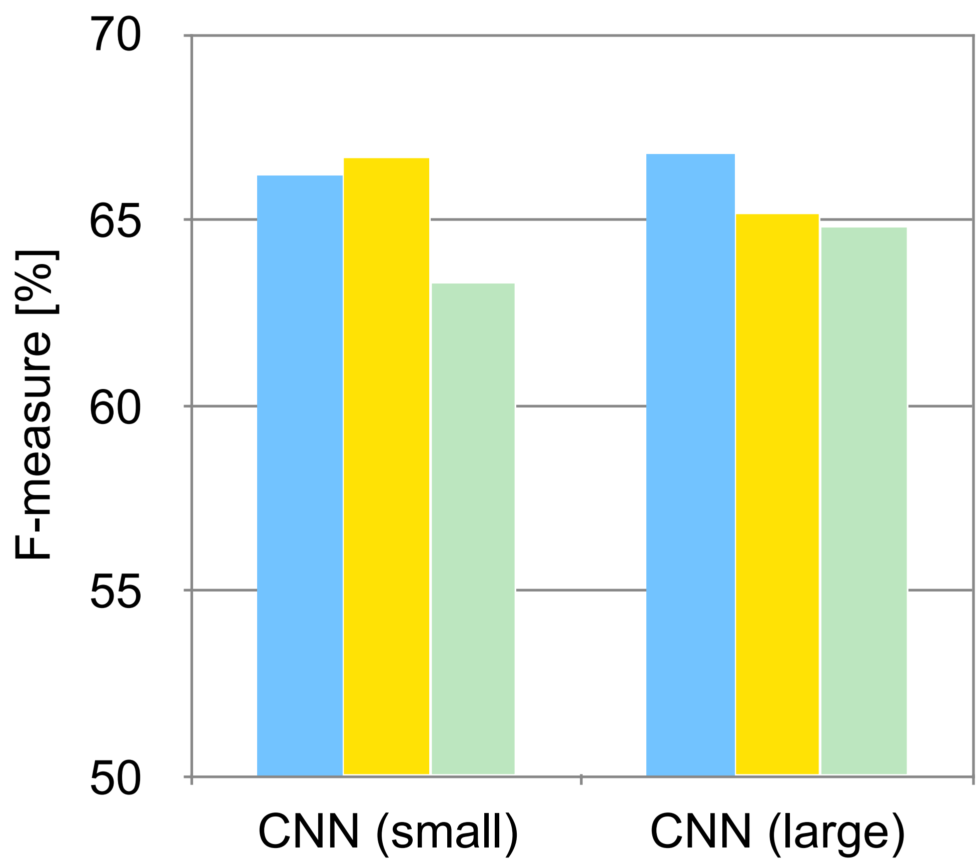
RESULTS: RNNs

Impact of **beats for RNN**s:

- **BF** improves for both models ✔
- **MT** improves for both models ✔



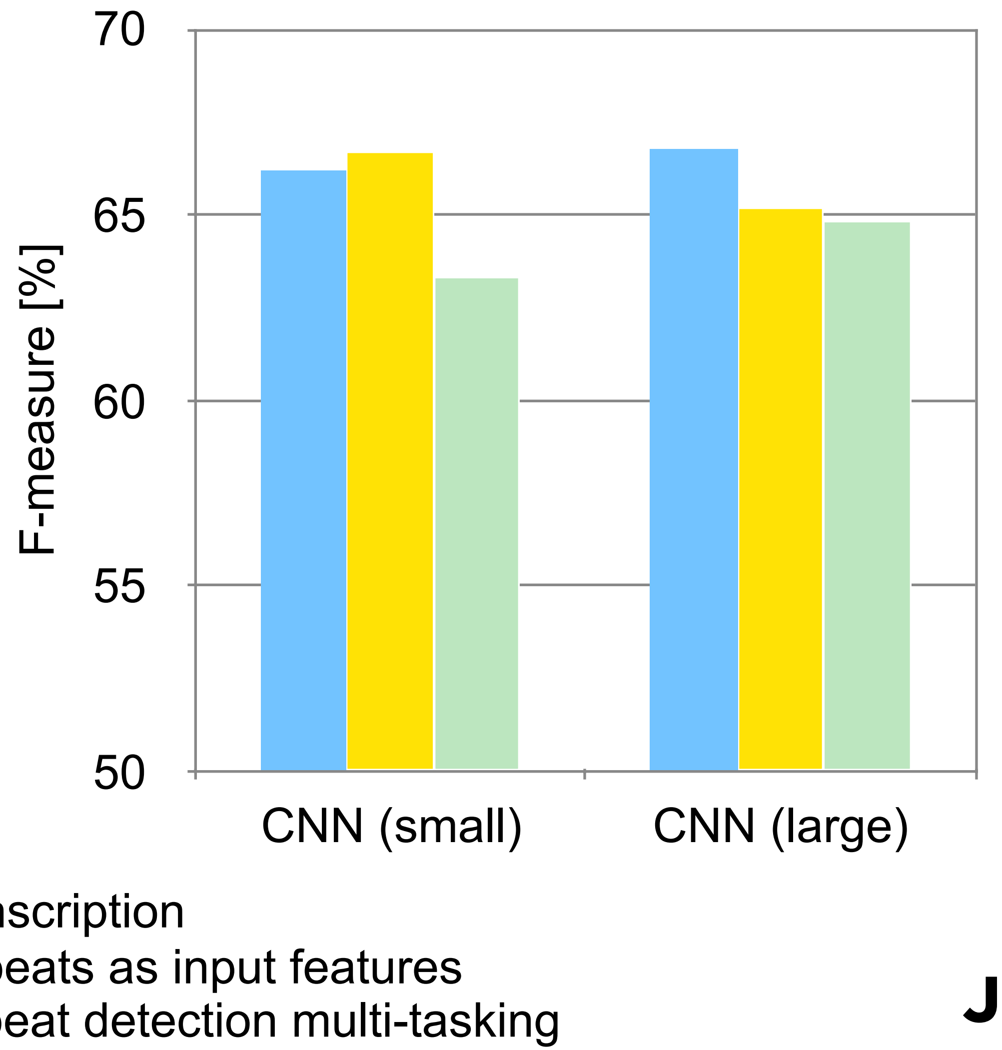
RESULTS: CNNs



DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS: CNNs

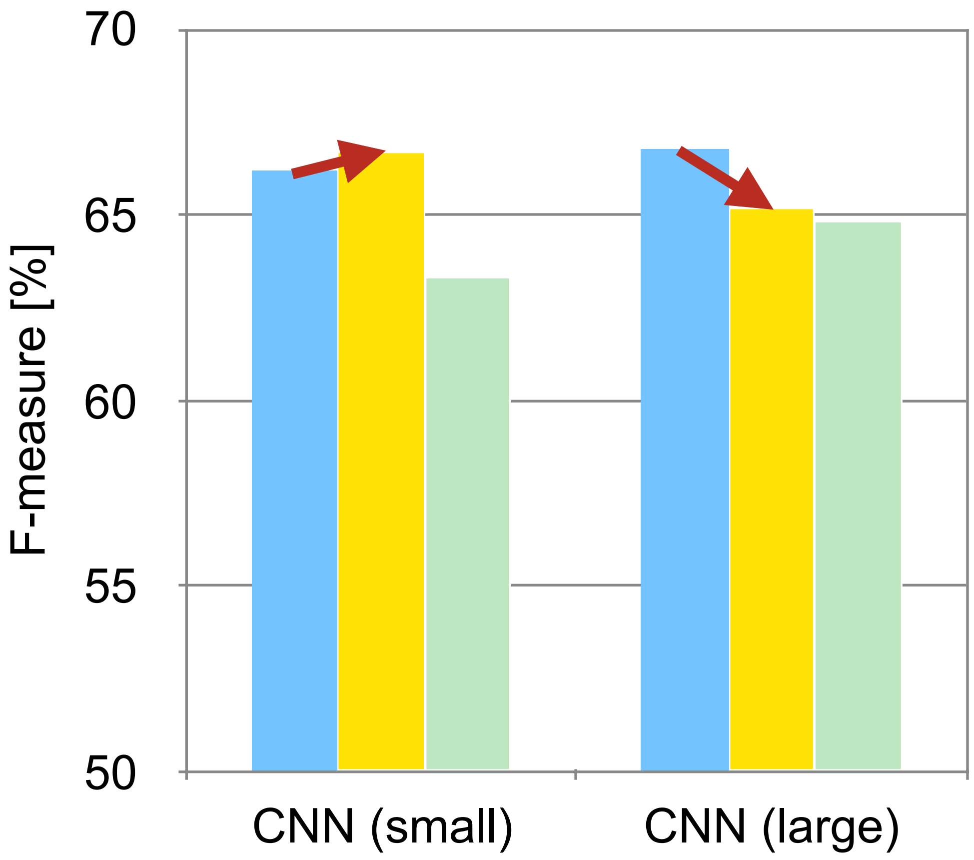
Impact of **beats for CNNs**:



RESULTS: CNNs

Impact of **beats for CNN**s:

- BF inconsistent

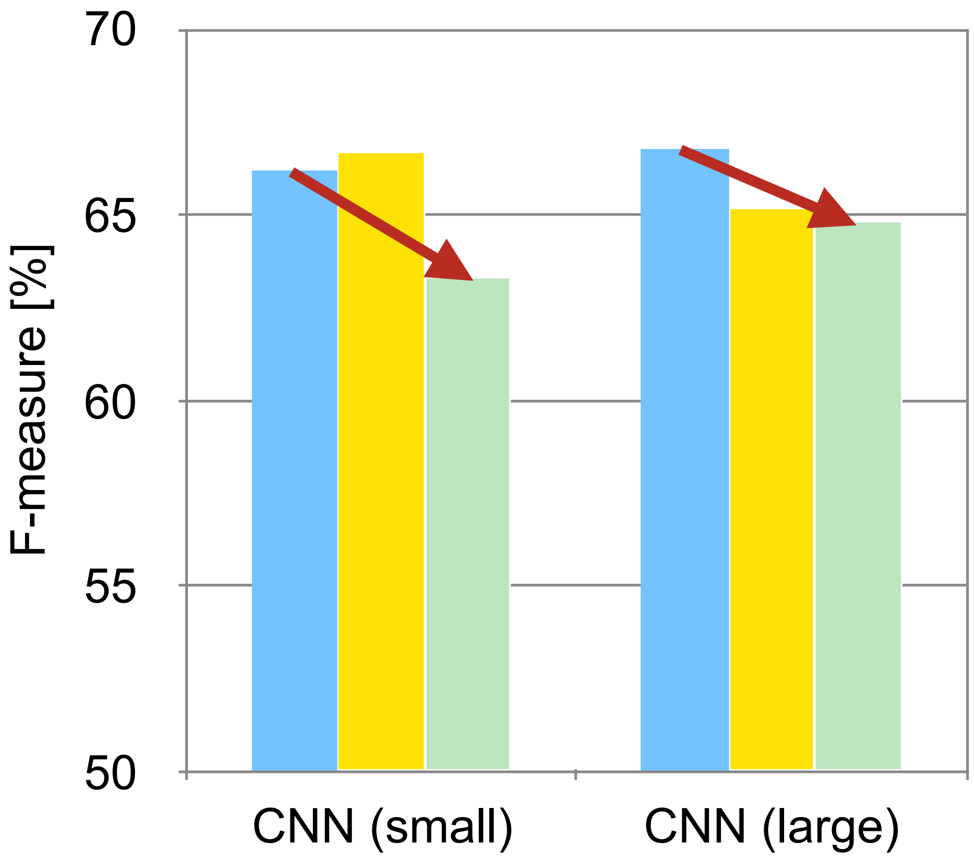


DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS: CNNs

Impact of **beats for CNN**s:

- **BF** inconsistent
- **MT** declines for both models

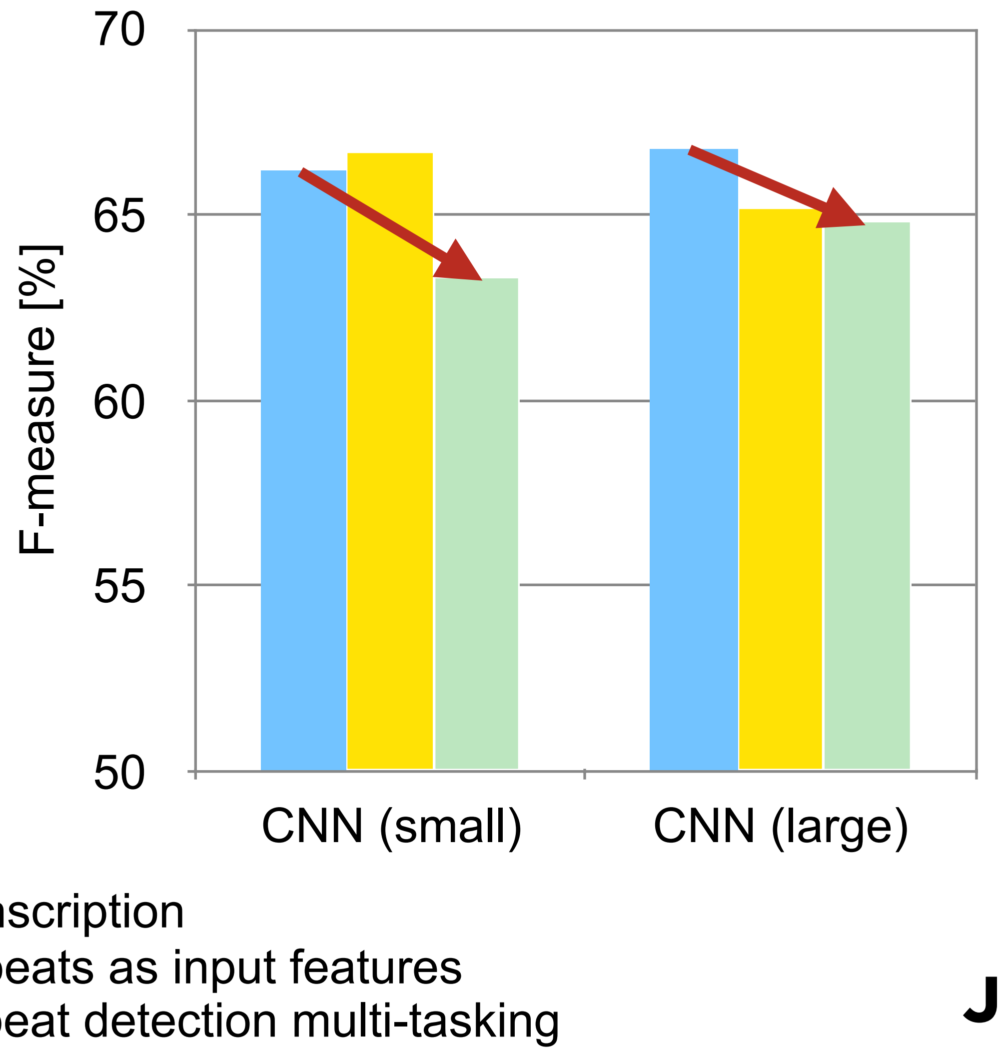


DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

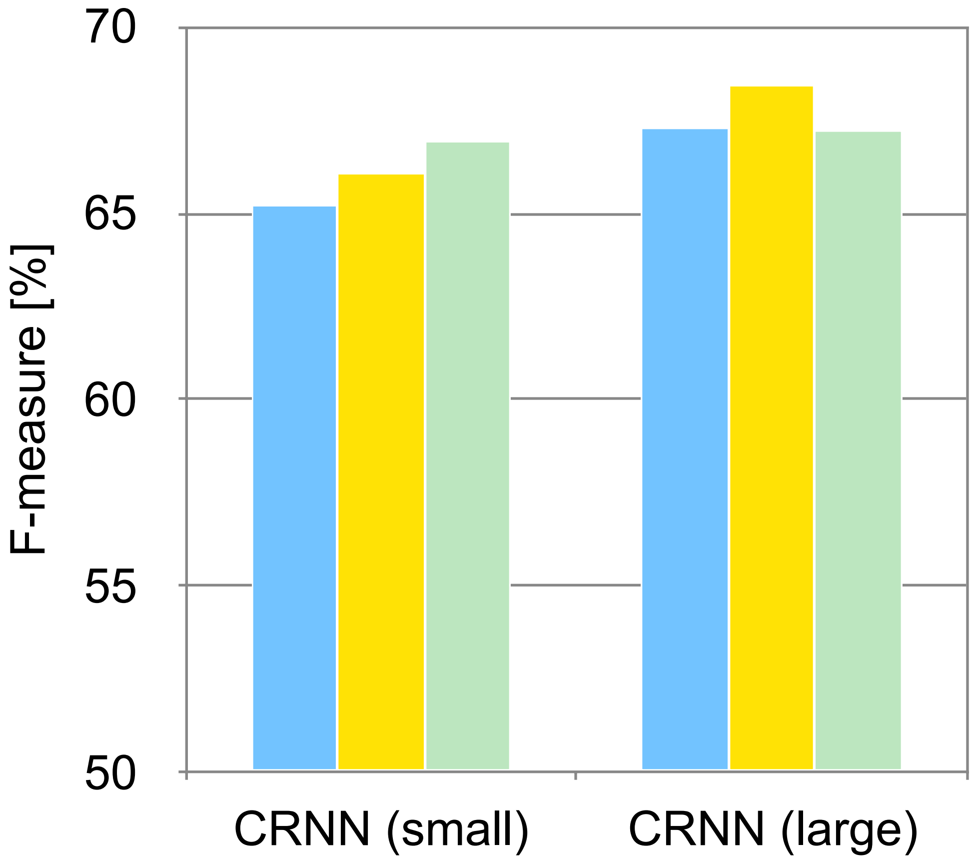
RESULTS: CNNs

Impact of **beats for CNN**s:

- **BF** inconsistent
- **MT** declines for both models
- Expected: CNNs have too little context for beats



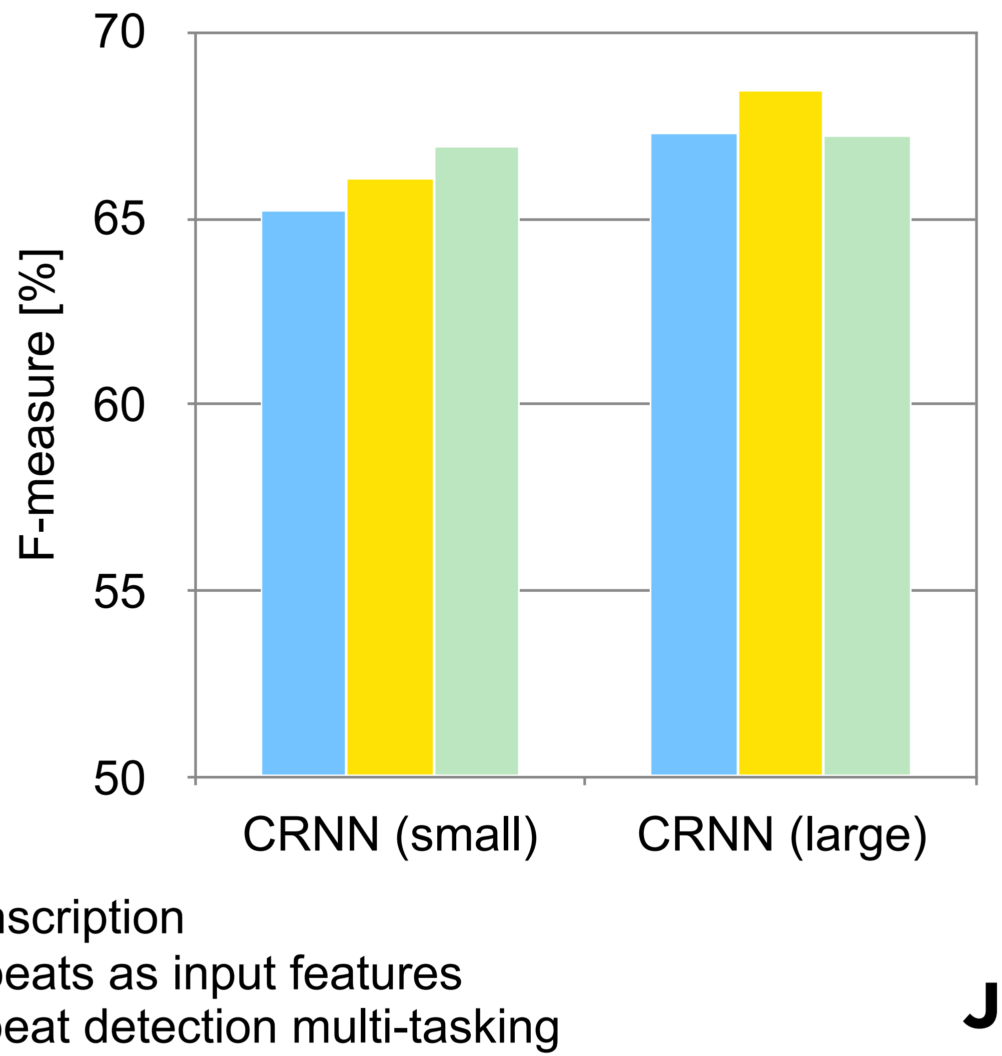
RESULTS: CRNNs



DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS: CRNNs

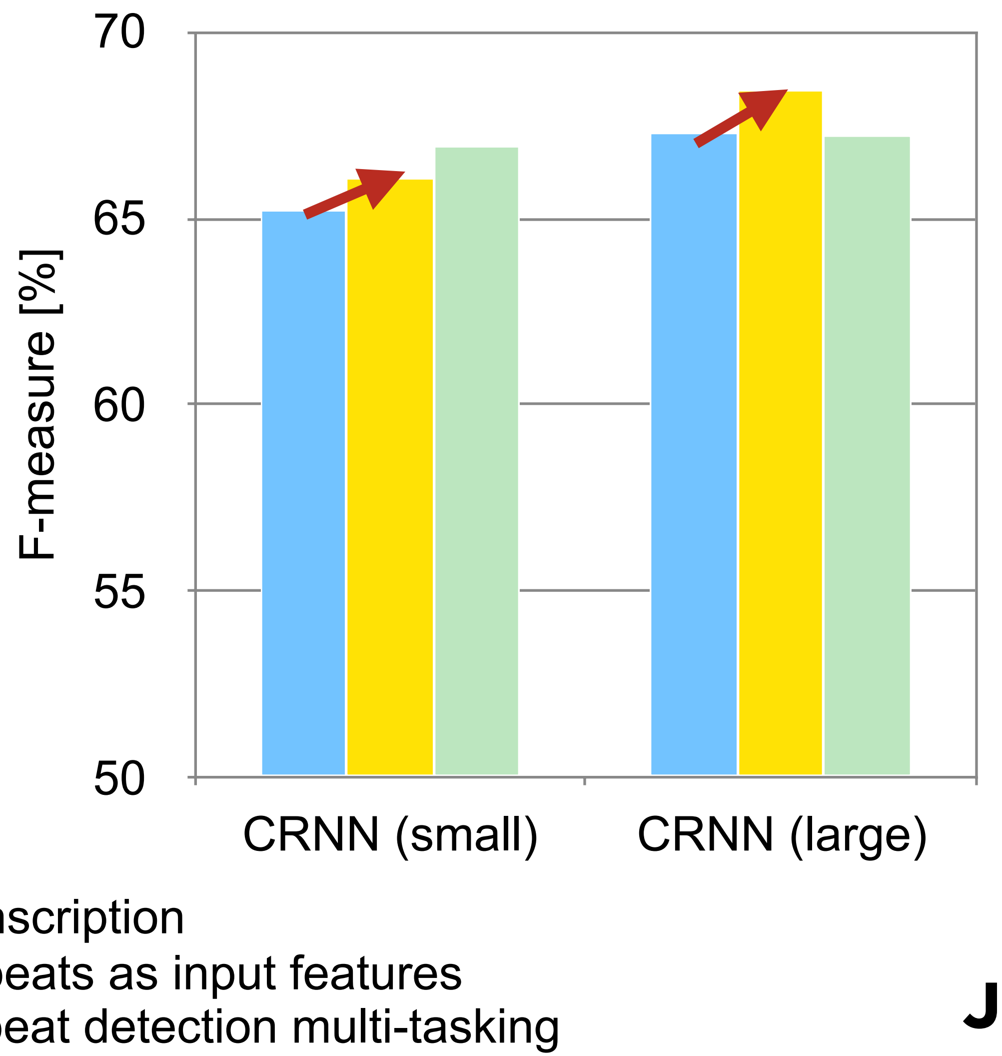
Impact of **beats for CRNNs**:



RESULTS: CRNNs

Impact of **beats for CRNNs**:

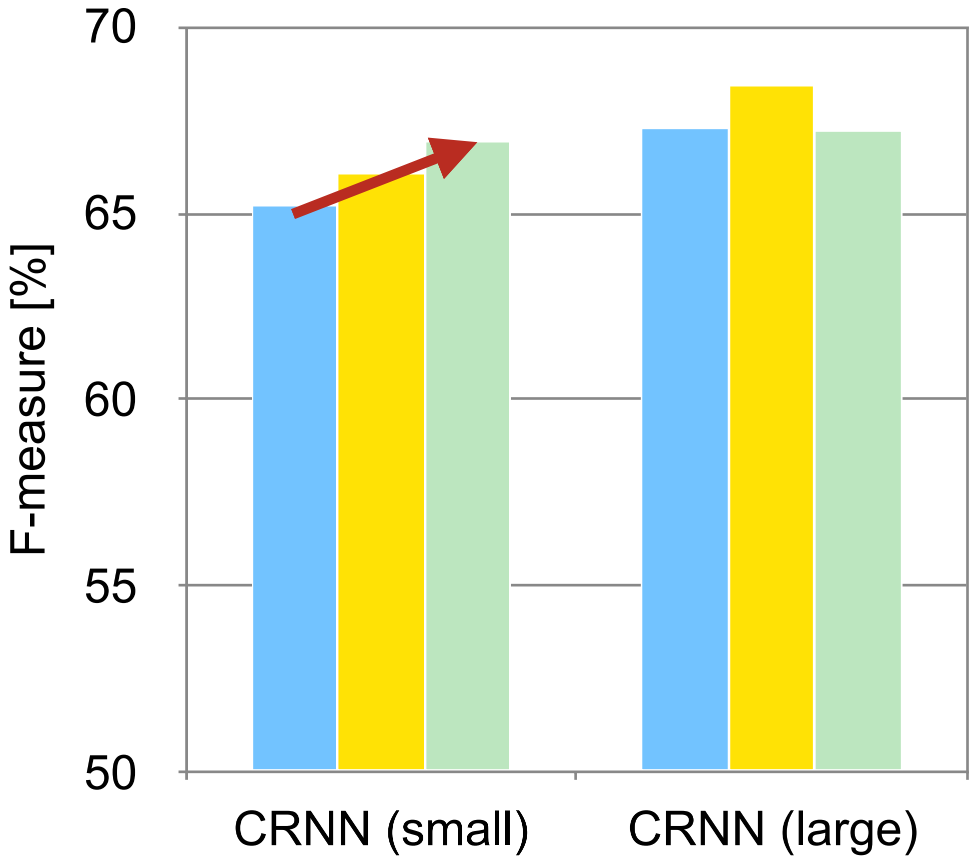
- BF improves for both models ✔



RESULTS: CRNNs

Impact of **beats for CRNNs**:

- **BF** improves for both models ✔
- **MT** improves for small models ✔

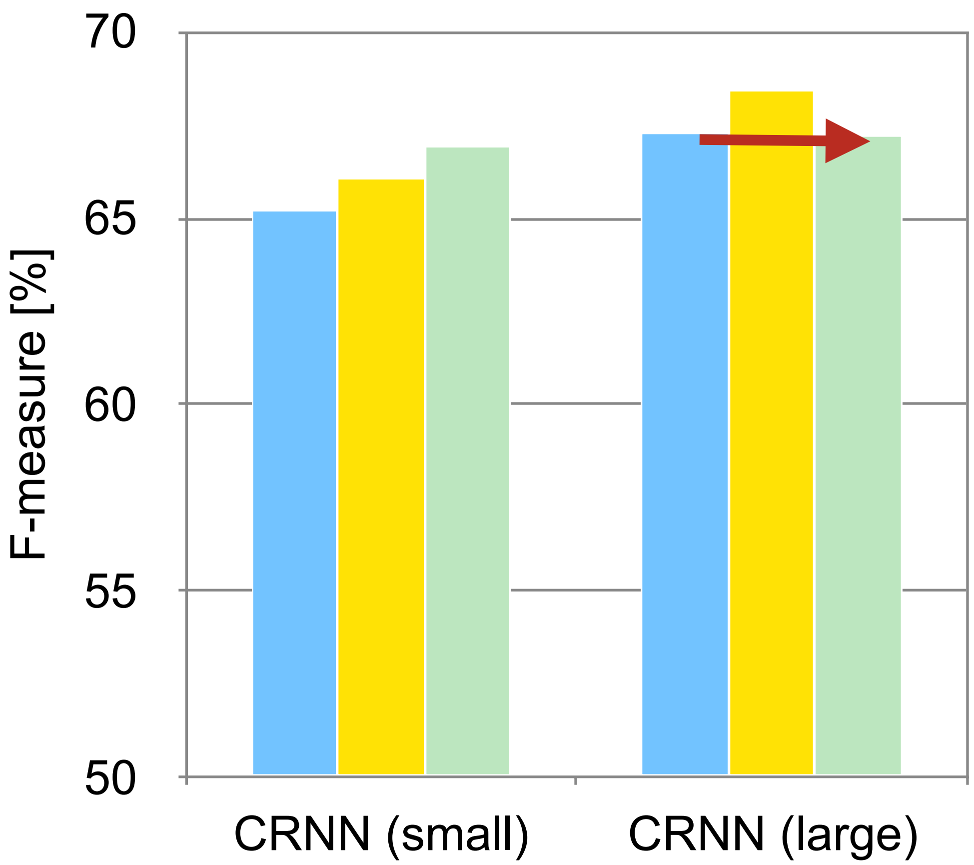


DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS: CRNNs

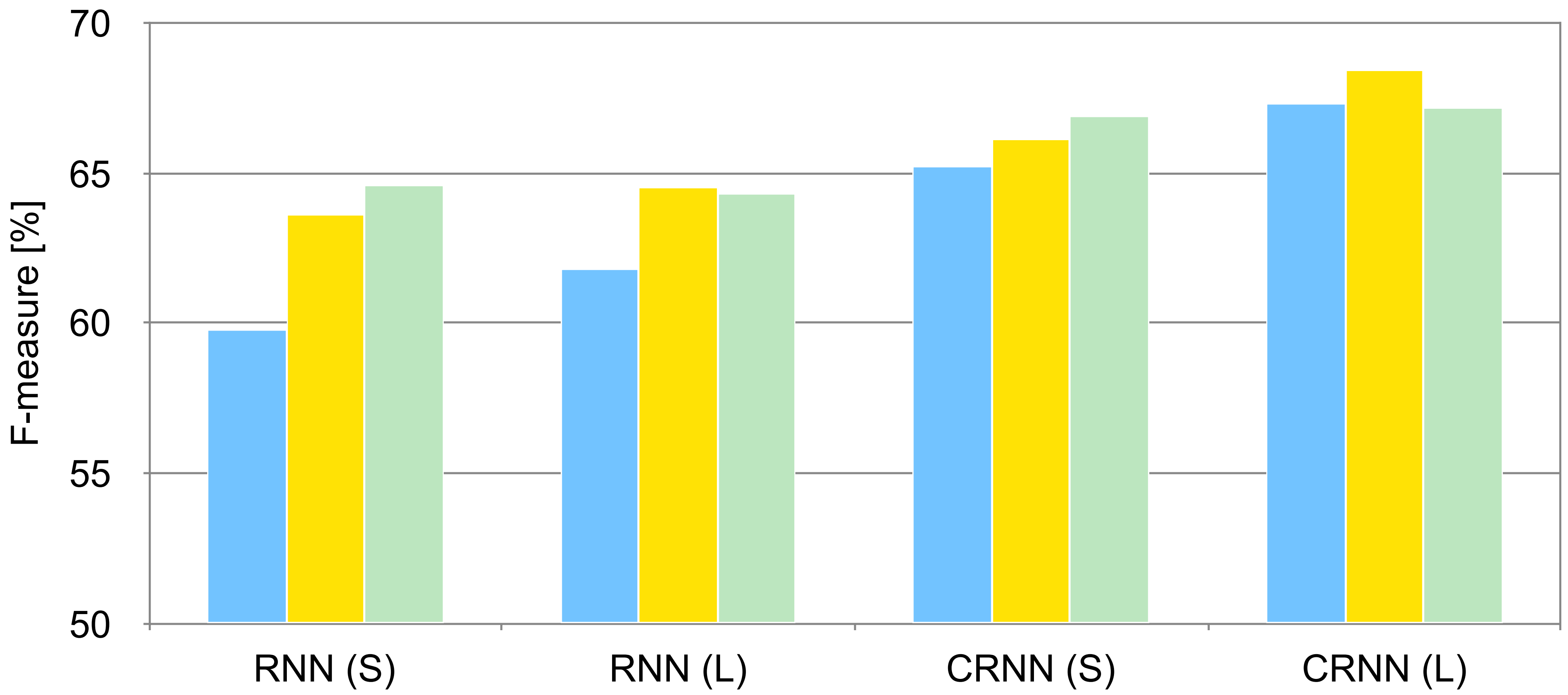
Impact of **beats for CRNNs**:

- BF improves for both models ✔
- MT improves for small models ✔
- MT equal for large model ?



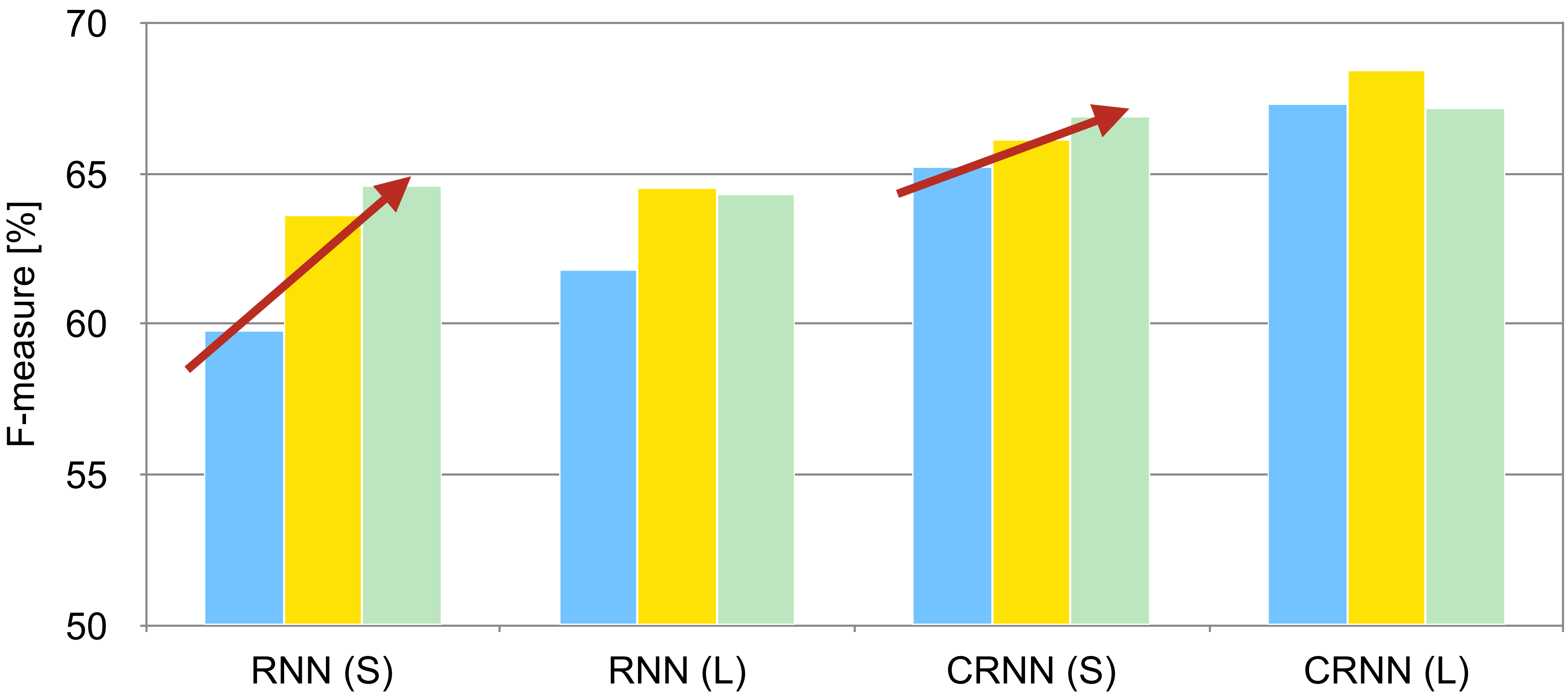
DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS FOR RECURRENT ARCHITECTURES



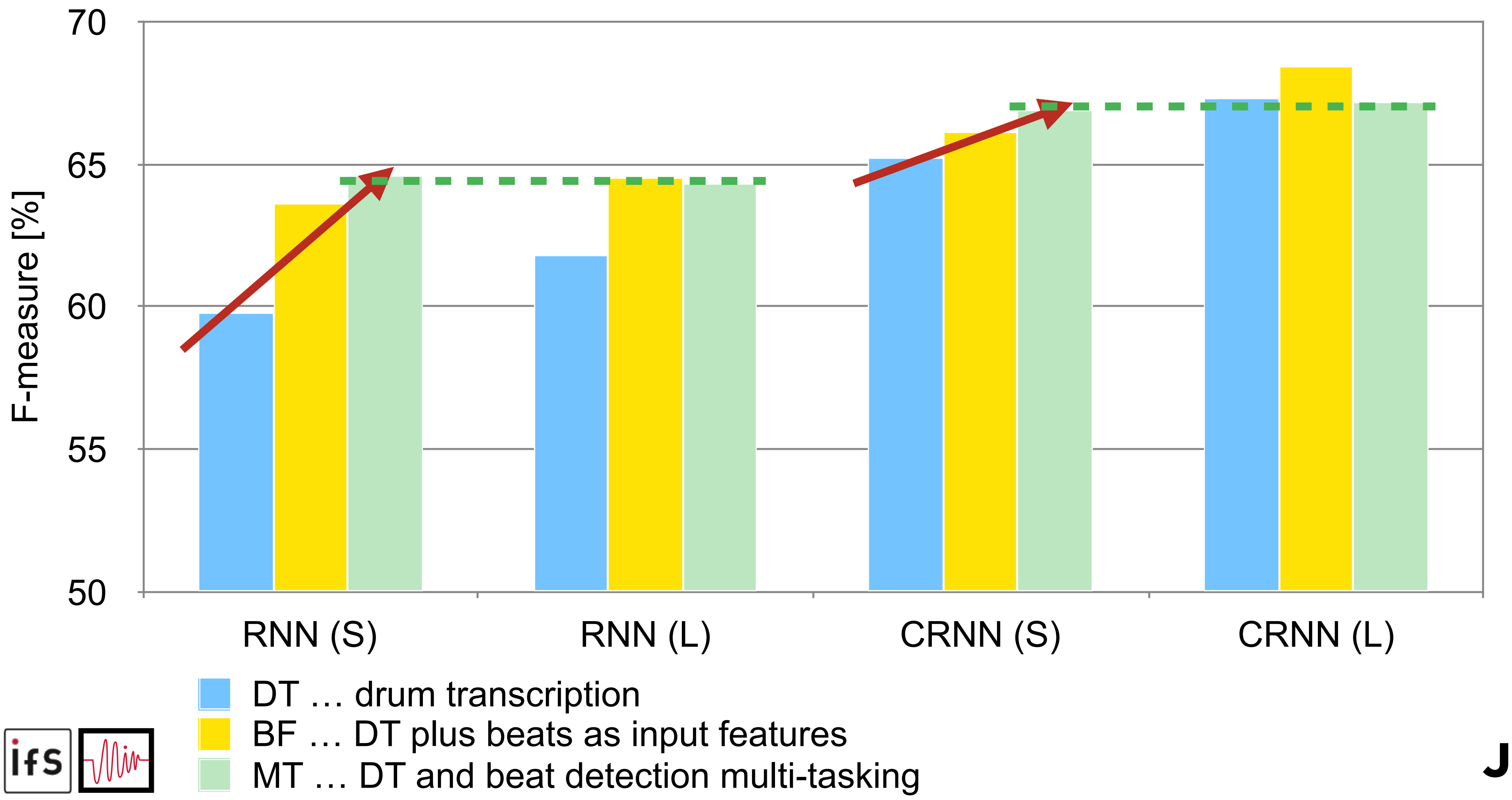
DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS FOR RECURRENT ARCHITECTURES

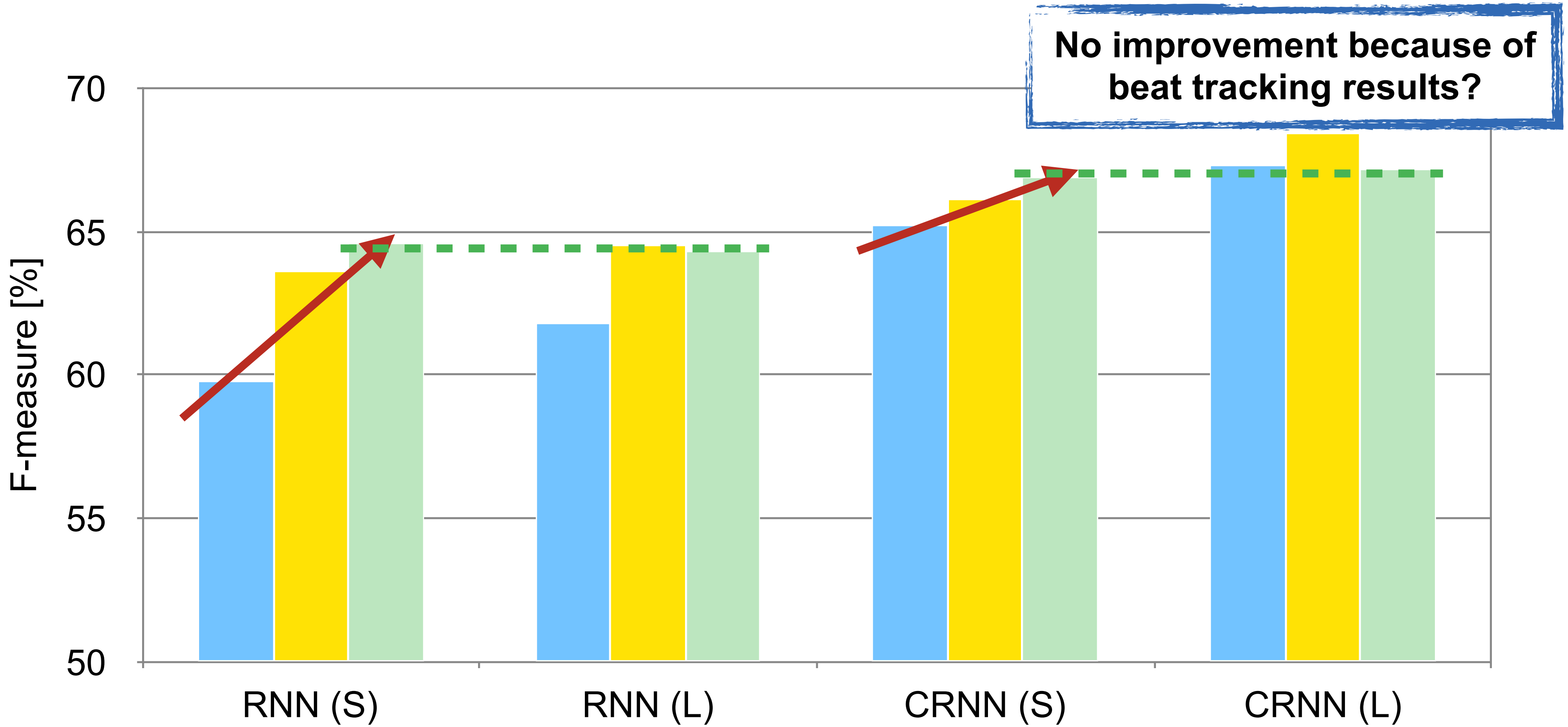


DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

RESULTS FOR RECURRENT ARCHITECTURES



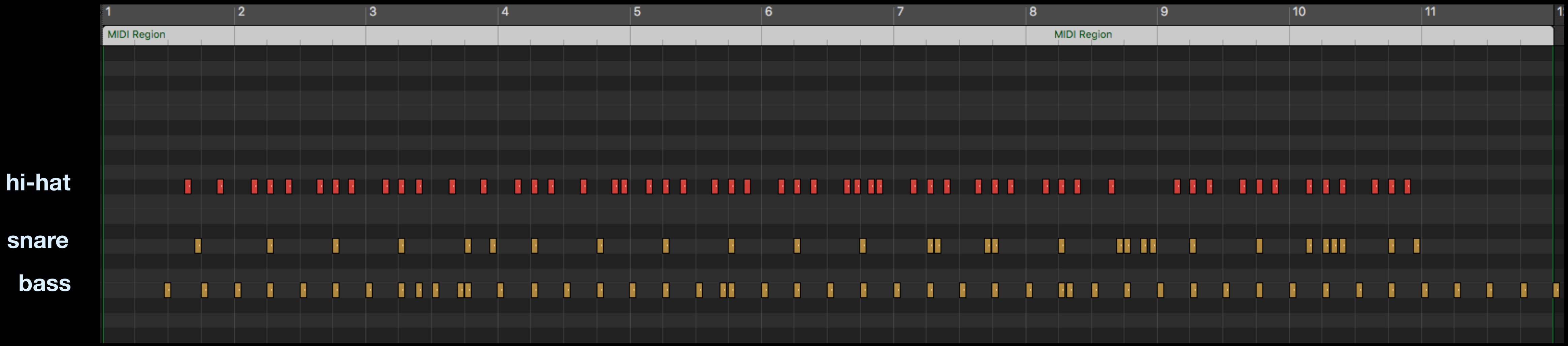
RESULTS FOR RECURRENT ARCHITECTURES



DT … drum transcription
BF … DT plus beats as input features
MT … DT and beat detection multi-tasking

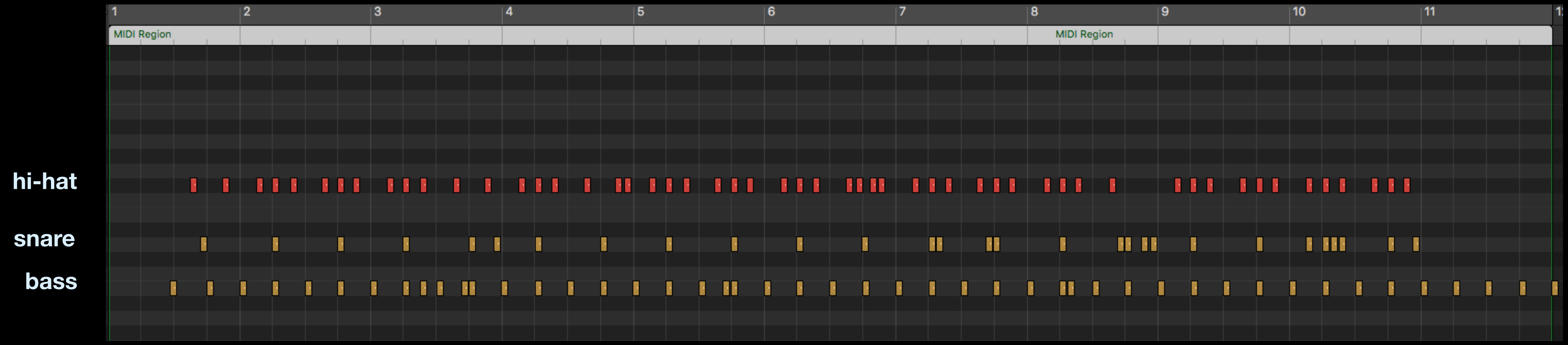
HOW DOES IT SOUND?

three instruments + beats



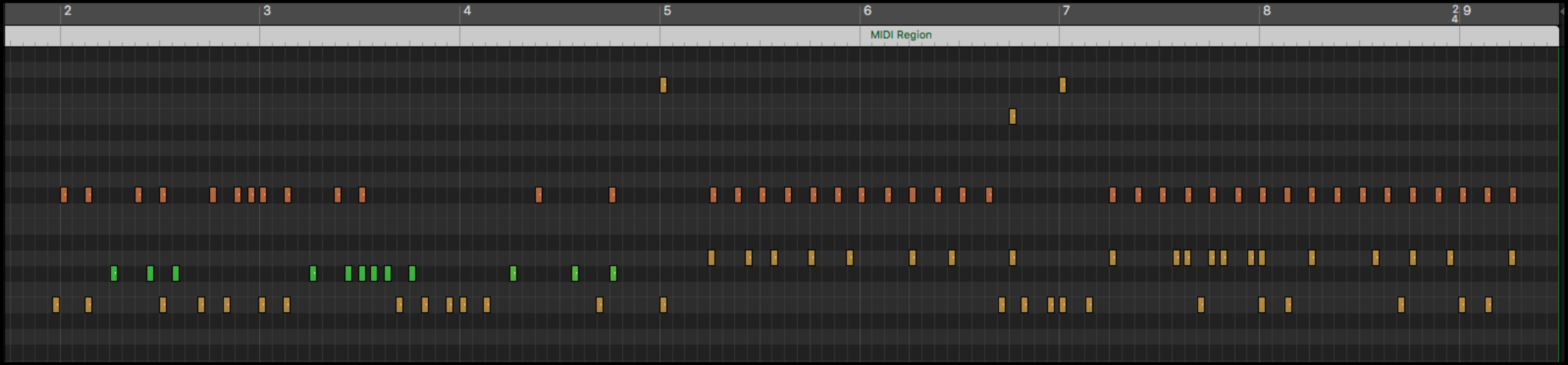
HOW DOES IT SOUND?

three instruments + beats



HOW DOES IT SOUND?

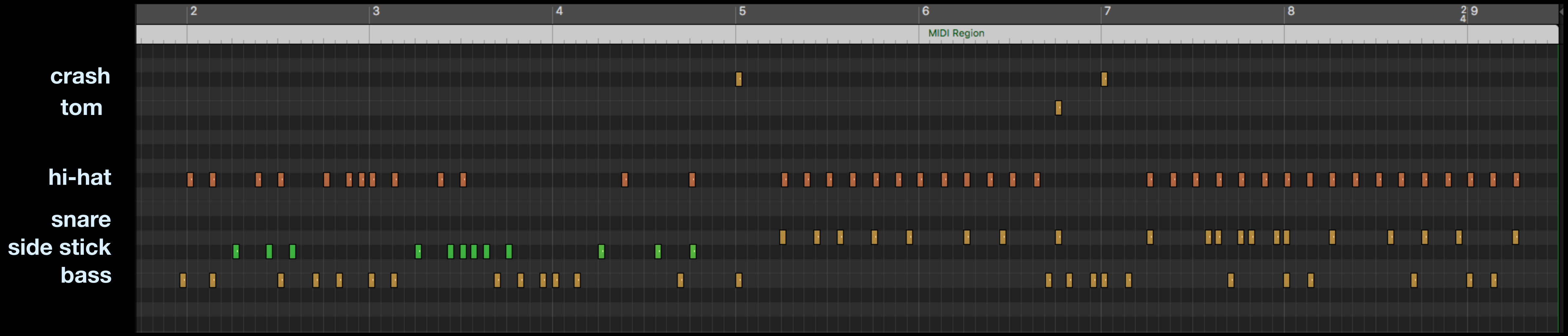
eight instruments + beats



crash
tom
hi-hat
snare
side stick
bass

HOW DOES IT SOUND?

eight instruments + beats



CONCLUSIONS

CONCLUSIONS

- **Deep learning** for automatic drum transcription

CONCLUSIONS

- **Deep learning** for automatic drum transcription
- **CRNNs** can outperform RNNs and CNNs, especially on complex data
  - Modeling of acoustic and rhythmic properties ➡ better generalization!

CONCLUSIONS

- **Deep learning** for automatic drum transcription
- **CRNNs** can outperform RNNs and CNNs, especially on complex data
  - Modeling of acoustic and rhythmic properties ➡ better generalization!
- Leverage **multi-task learning** effects to increase performance
  - All instruments under observation within **one model**
  - Beats and downbeats for additional **meta data** for transcripts

CONCLUSIONS

- **Deep learning** for automatic drum transcription

- **CRNNs** can outperform RNNs and CNNs, especially on complex data
  - Modeling of acoustic and rhythmic properties ➡ better generalization!

- Leverage **multi-task learning** effects to increase performance
  - All instruments under observation within **one model**
  - Beats and downbeats for additional **meta data** for transcripts

- CRNN **best overall results** @ MIREX'17 and MIREX'18 drum transcription

  MIREX system:
  http://ifs.tuwien.ac.at/~vogl/models/mirex-17.zip
  http://ifs.tuwien.ac.at/~vogl/models/mirex-18.tar.gz

CONCLUSIONS

- **Deep learning** for automatic drum transcription

- **CRNNs** can outperform RNNs and CNNs, especially on complex data
  - Modeling of acoustic and rhythmic properties ➡ better generalization!

- Leverage **multi-task learning** effects to increase performance
  - All instruments under observation within **one model**
  - Beats and downbeats for additional **meta data** for transcripts

- CRNN **best overall results** @ MIREX'17 and MIREX'18 drum transcription
  MIREX system:
  http://ifs.tuwien.ac.at/~vogl/models/mirex-17.zip
  http://ifs.tuwien.ac.at/~vogl/models/mirex-18.tar.gz

CONCLUSIONS

- **Deep learning** for automatic drum transcription
- **CRNNs** can outperform RNNs and CNNs, especially on complex data
  - Modeling of acoustic and rhythmic properties ➡ better generalization!
- Leverage **multi-task learning** effects to increase performance
  - All instruments under observation within **one model**
  - Beats and downbeats for additional **meta data** for transcripts
- CRNN **best overall results** @ MIREX’17 and MIREX’18 drum transcription

  MIREX system:
  http://ifs.tuwien.ac.at/~vogl/models/mirex-17.zip
  http://ifs.tuwien.ac.at/~vogl/models/mirex-18.tar.gz