Four-way Classification of Tabla Strokes with Models Adapted from Automatic Drum Transcription

Rohit M. A., Amitrajit Bhattacharjee, Preeti Rao

Dept. of Electrical Engineering
I.I.T. Bombay, India

The Tabla

- Pitched, percussive hand-drums
  - Bayan - bass (F0 ∈ 80-100 Hz)
  - Dayan - treble (F0 ∈ 200-400 Hz)

- In performance
  - Solo – playing improvisation, compositions
  - Accompaniment – cyclic stroke pattern (*theka)*

Bayan
(Bass)

Dayan
(Treble)

Tabla solo
https://youtu.be/ckE8GH5tI2A?t=4095

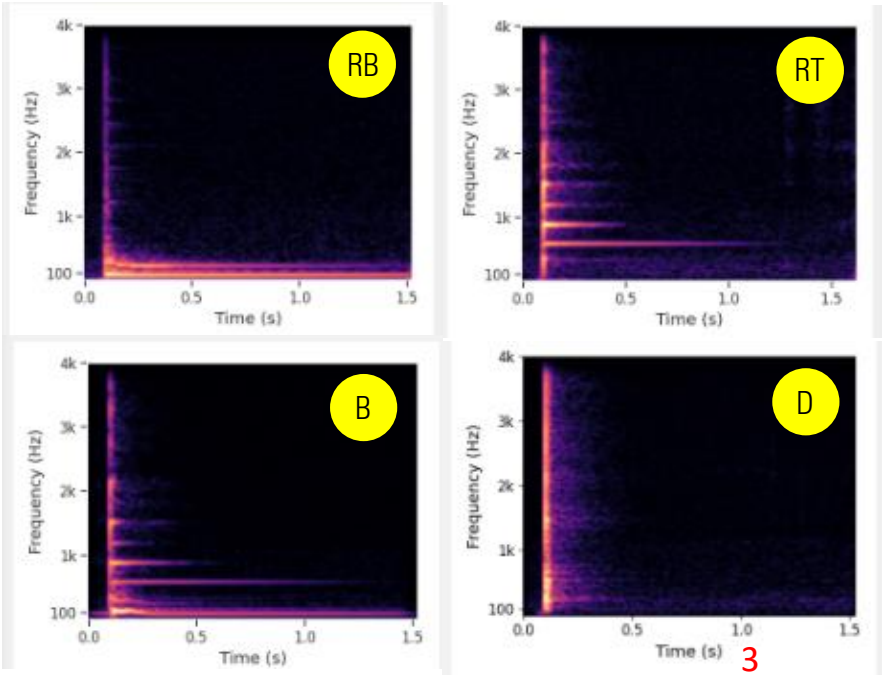
Percussion accompaniment
https://youtu.be/oACbmNkih0I?t=103

Tabla Strokes and Categories

- Tabla strokes
  - 10-15 in number, identified by '*bols*' (syllables like Na, Tin, Ghe, Dha, etc.)
  - May involve single drum or both simultaneously
  - Resonant (R) – sustained, harmonic
  - Damped (D) – transient, percussive

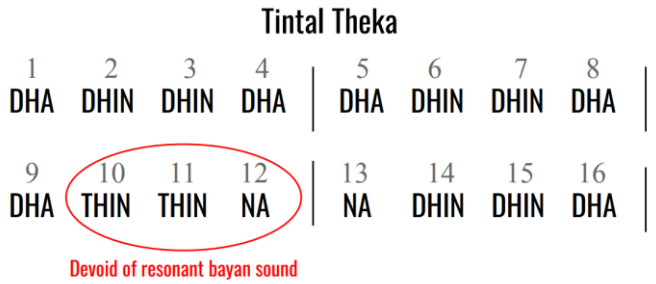
- 4 stroke categories

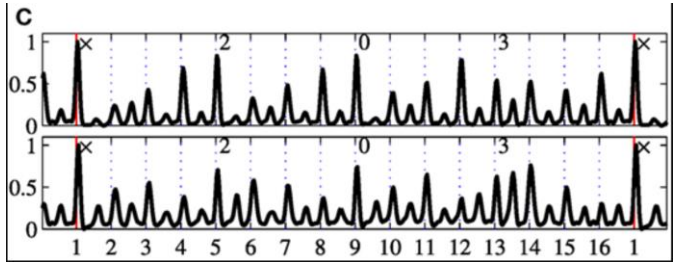
Category	Bass drum	Treble drum	Bols
Resonant Bass (RB)	R	D / Nil	Ghe, Dhe, Dhi
Resonant Treble (RT)	D / Nil	R	Na, Tin, Tun
Resonant Both (B)	R	R	Dha, Dhin
Damped (D)	D / Nil	D / Nil	Ti-Ta, Te-Re, Ke,



Relevance of Stroke Categories

- Musicologically motivated
  - Mark salient positions in theka
  - Are tied to expressive tabla playing elements (loudness dynamics, pitch modulation) [2,3]





- Aiding computational musicology
  - Analysis of played strokes requires expensive manual labelling
  - Automatic stroke classification can enable corpus-level analysis

[2] M. Clayton, "Theory and practice of long-form non-isochronous metres," Music Theory Online 2020
[3] A. Srinivasamurthy et al. "Aspects of tempo and rhythmic elaboration in hindustani music: A corpus study." Frontiers in Digital Humanities 2017

Existing Methods do not Generalize

- Previous work – tabla bol classification

- Poor instrument-independent classification accuracies
  - Highly variable test set accuracies on unseen tabla (15 - 95% [4])
  - Lack of sufficient data with diversity in playing & instrument characteristics

[4] P. Chordia, "Segmentation and recognition of tabla strokes," ISMIR 2005

Objectives

- 4-way tabla stroke classification system
  - Robust to instrument and playing-style changes
- Target classes
  - Damped (D)
  - Resonant Treble (RT)
  - Resonant Bass (RB)
  - Resonant Both (B)
- Target test scenario
  - Tabla accompaniment to Hindustani classical vocals

Approach

1. Build larger, diverse dataset

2. Design effective classification models
   - Exploit pre-trained models from western drums transcription

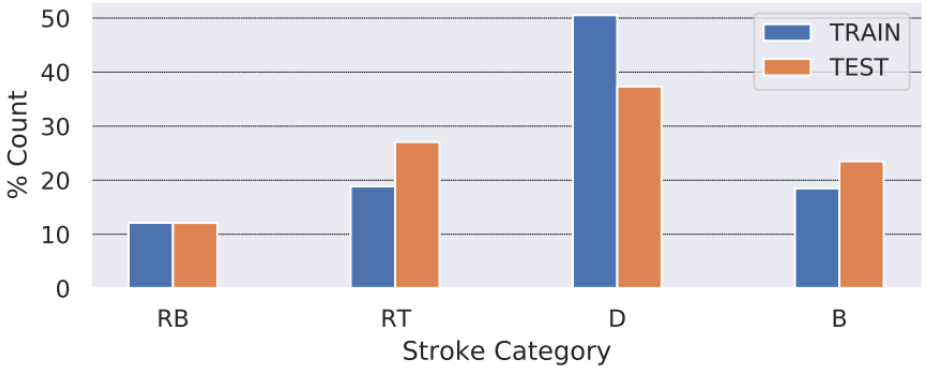
3. Explore novel data augmentation methods

Dataset

- New labelled dataset diverse in terms of
  - Instruments, players, tabla tuning, playing tempo

	# tablas	Duration	# strokes
Training [5,6]	10	76 min.	26,600
Testing [6]	3	20 min.	4,470

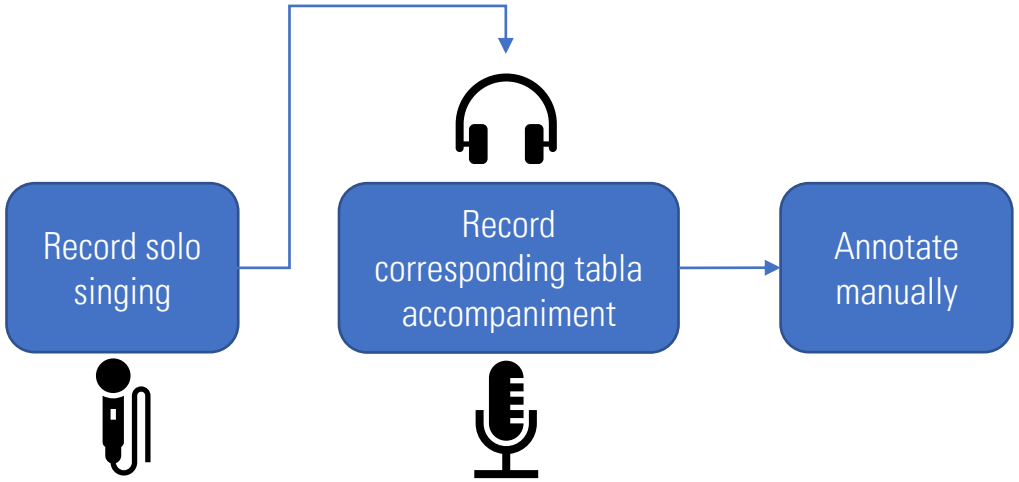
- Stroke category distribution is not uniform
  - Most strokes of D, least of RB



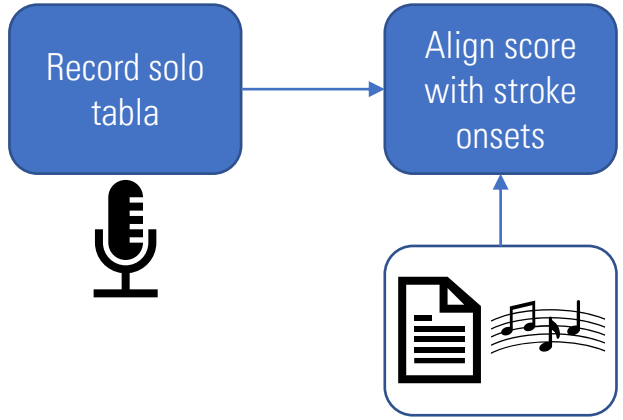
[5] R. Gowriprasad and K. S. R. Murty, "Onset detection of tabla strokes using LP analysis". SPCOM 2020.
[6] M. A. Rohit and P. Rao, "Automatic stroke classification of tabla accompaniment in hindustani vocal concert audio". JASI 2021

Data Collection and Annotation

- Testing set
  - Realistic tabla accompaniment to vocals (theka)
  - Recorded in isolation



- Training set
  - Tabla solo audios (kaida, tukda, tihai, etc.)
  - Split into 3 cross-validation folds (no tabla overlap)



Methods

1. Classification – CNN models inspired from western automatic drums transcription (ADT)
   A. 3-way drums CNN with transfer learning
   B. Bank of retrained 1-way CNNs

2. Data augmentation

Overview of CNN methods

- Input
  - Mel-spectrogram excerpt of ‘C’ channels x ‘F’ bins x ‘T’ frames
  - Channels are spectrograms computed at different resolutions
- Target
  - Binary value indicating presence of onset at frame T/2
- Output
  - Onset probability in [0,1]
  - Multi-label: single model, multiple outputs
  - 1-way: different models, one-versus-all binary output

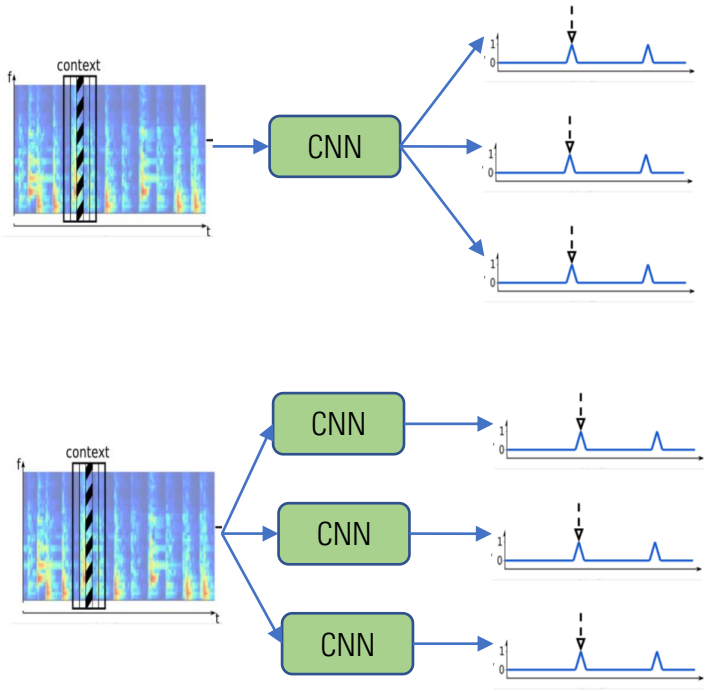
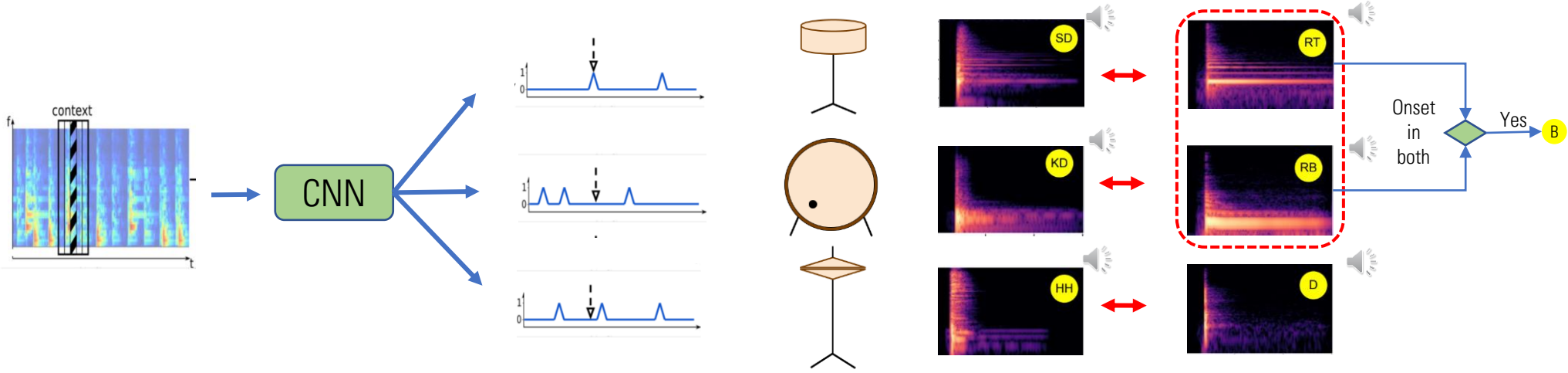


Figure inspiration: R. Vogl et al, “Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks,” ISMIR 2017

A. 3-way CNN with Transfer Learning

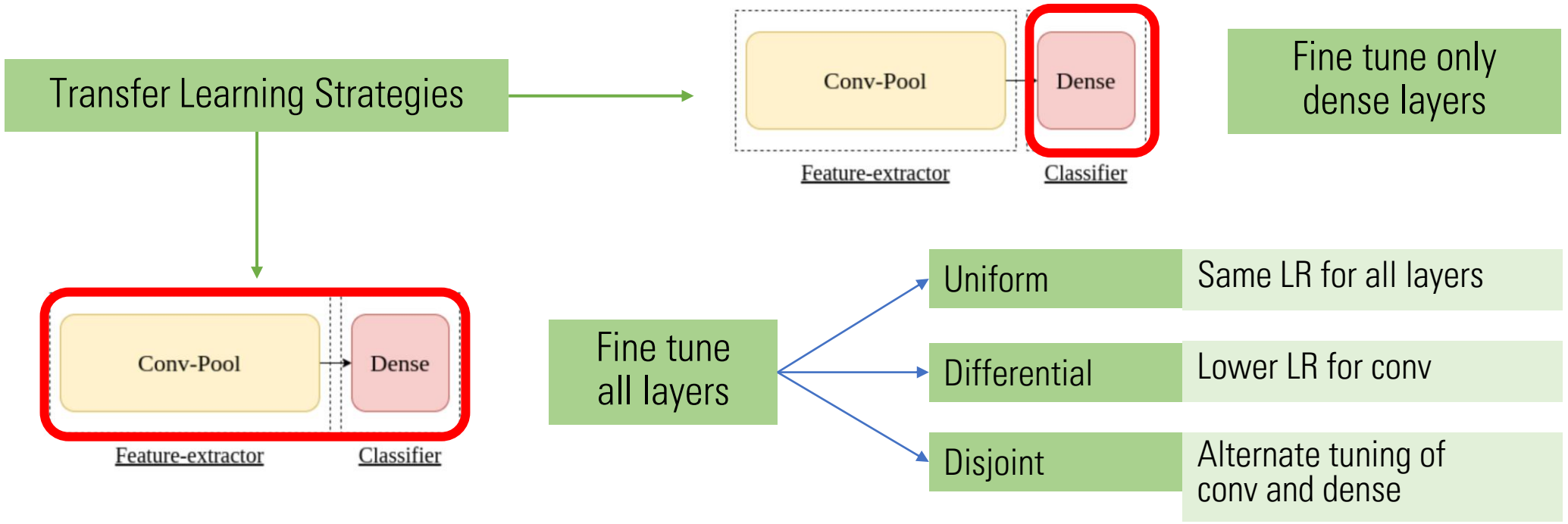
- Pre-trained CNN drums transcription (ADT) models from *madmom* [7] fine-tuned on tabla data
  - Model originally trained on MIREX drums transcription dataset (about 3x our tabla dataset)

- Motivated by correspondence between drum types and tabla stroke categories



[7] R. Vogl and P. Knees, "Mirex submission for drum transcription 2018". ISMIR 2018

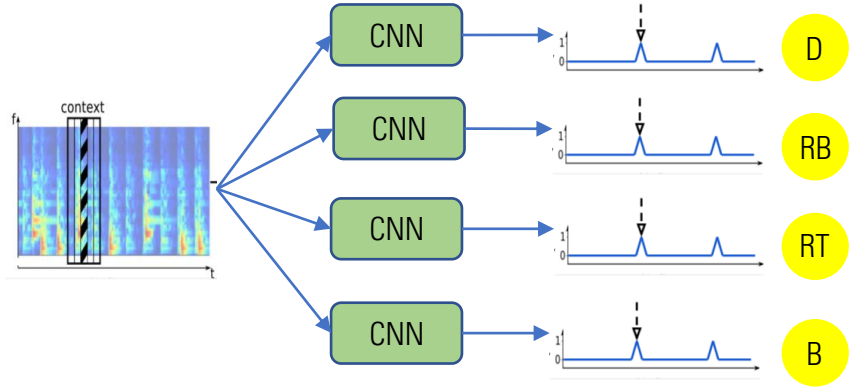
A. 3-way CNN with Transfer Learning

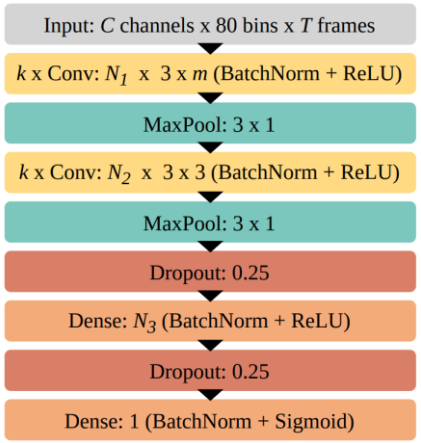


- Evaluate against re-training same model architecture on tabla data

B. Bank of 1-way CNNs

- Bank of four 1-way CNNs
  - Allows optimizing model separately for each category
- We start with a baseline architecture from ADT [8]
  - Hyperparameters tuned for each category to account for data imbalance
  - Include variations to input representation and model capacity
- Trained from scratch for each category
  - No pre-trained ADT model available
  - Drums dataset used for 3-way CNN also not public





Variant	Hyperparameter values
Baseline	C=3, T=15 (150 ms), k=1, m=7, N_1=16, N_3=128
↑context	T=21 (210 ms)
Mid-channel	C=1 (middle)
↑conv filters	N_1=32
↑dense units	N_3=256
↑conv filt. + ↑dense units	N_1=32, N_3=256
2x conv layers	k=2, m=3

[8] C. Jacques and A. Röbel. "Automatic drum transcription with convolutional neural network". DAFx 2018

Methods

1. Classification – CNN models inspired from western automatic drums transcription (ADT)
   A. 3-way drums CNN with transfer learning
   B. Bank of retrained 1-way CNNs

2. Data augmentation

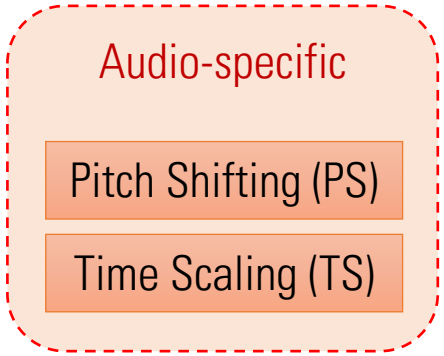
Data Augmentation - Overview

- Diversity expected in tabla dataset
  - **Instrument characteristics** – tuning, timbre, resonance & decay levels
  - **Playing style** – tempo, expressive dynamics
  - **Recording conditions** – spectral levels, balance, decay level

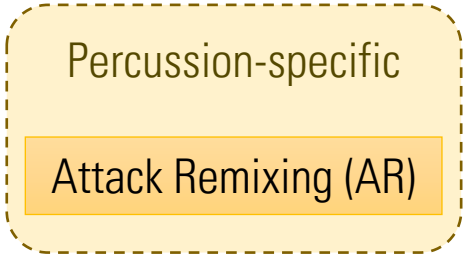
- We explore individual methods and combinations
  - Each method applied to time-domain training set audio, generates 4 variations
  - Evaluated using 1-way CNN models

Augmentation Methods From Literature

- Pitch-shifting (PS) & time-scaling (TS)
  - Capture tuning and tempo variations



- Attack remixing (AR)
  - Modify relative levels of signal attack and decay
  - Used previously in ADT [9]



[9] C. Jacques and A. Röbel. "Data augmentation for drum transcription with CNNs". EUSIPCO 2019

Tabla-specific Methods – Spectral Filtering

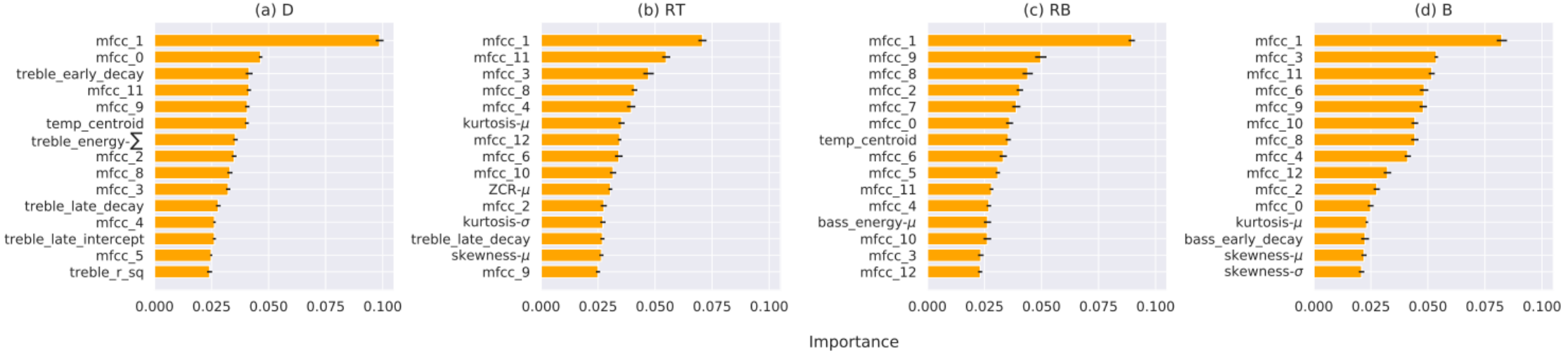
- Spectral filtering commonly used in MIR [10]
  - Filter applied over randomly chosen spectral bands
- Given specific bands of activity in tabla, we filter bass & treble regions
  - Capture recording conditions, resonance characteristics
- We also identify and modify features that vary across instruments & not stroke categories
  - Perturbing attributes irrelevant to discrimination task shown to be effective for augmentation [11]
  - For our task, these can be instrument-specific low-level acoustic features

[10] J. Schluter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks". ISMIR 2015.
[11] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," IEEE ASRUW 2017.

Finding instrument-dependent characteristics

- **Tabla identification task** using a random forest classifier
  - About 50 features used, represent various spectral and temporal characteristics [12]
  - Training set has audio from 10 unique tabla-sets
  - Samples from each stroke category are used separately to fit RF models

- Resulting feature ranking highlights MFCC-1 as most important

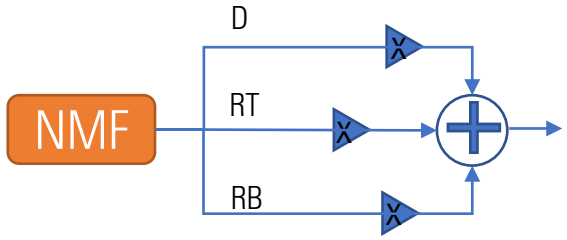


[12] M. A. Rohit and P. Rao, "Automatic stroke classification of tabla accompaniment in hindustani vocal concert audio". JASI 2021

Tabla-specific Methods – Stroke Remixing

- Expressive playing involves modifying relative stroke intensities
- Compound strokes differ in contribution of each drum

- Decompose tabla audio into components of each stroke type
  - NMF with pre-initialized and fixed bases
- Remix at different levels to simulate playing dynamics



[13] P. López-Serrano et al, "Nmf toolbox: Music processing applications of nonnegative matrix factorization," DAFx 2019.

Results

1. Transfer learning with 3-way drums CNN models
2. Re-trained 1-way CNN models (bank of 4)
3. Data augmentation
4. Overall comparison

1. Adapted 3-way drums CNN

- Disjoint tuning of both conv and dense layers better than other fine-tuning approaches
- Higher mean f-score than re-trained model of same architecture

Cross-validation F-scores

Method	Stroke category				Mean
	D	RT	RB	B	
Pre-trained (PT)	36.8	15.1	9.8	7.3	17.3
Re-trained	**81.0**	53.7	15.7	63.0	53.4
FT dense random init.	74.4	55.9	33.6	63.4	56.8
FT dense PT init.	71.7	54.8	29.4	60.9	54.2
Uniform FT all	76.3	59.7	29.5	65.3	57.7
Differential FT all	72.5	58.7	30.0	63.5	56.2
Disjoint FT all: dense rand. init.	77.2	57.4	33.0	65.9	58.3
Disjoint FT all: dense PT init.	74.8	**66.4**	**34.7**	**66.5**	**60.6**

2. Re-trained 1-way CNNs

- Separate hyperparameter tuning results in superior class-specific architectures
  - **D**: more dense layer units
  - **RT**: more conv layer filters
  - **RB and B** (data scarce): Baseline model

Cross-validation F-scores

Model	Stroke category			
	D	**RT**	**RB**	**B**
Baseline	84.6	83.2	**46.5**	**83.8**
↑context	84.3	81.4	41.9	73.0
Mid-channel	84.7	81.7	42.1	75.6
↑conv filters	84.7	**84.5**	44.7	77.6
↑dense units	**86.7**	82.9	40.1	73.6
↑conv filters+↑dense units	83.5	83.4	43.3	82.0
2x conv layers	84.3	82.4	42.4	75.9

3. Data augmentation

- Improves all cross-validation f-scores
- Combination of PS, TS, SF, SR gives highest f-scores except in RT

Cross-validation F-scores

Method	Stroke category				Mean
	D	RT	RB	B	
No aug.	86.7	84.5	46.5	83.8	75.4
Pitch-shift	87.2	**85.5**	51.2	83.9	76.9
Time-scale	88.2	85.0	50.2	82.2	76.4
Attack-remix	84.3	84.2	48.1	81.3	74.5
SF-bass	84.5	80.9	40.4	79.9	71.4
SF-treble	85.8	81.7	48.7	76.0	73.0
SF-tilt	86.3	82.7	43.8	82.0	73.7
SF-all	87.6	84.6	50.7	85.6	77.1
SR-bass	86.0	84.8	43.3	83.6	74.4
SR-treble	86.1	84.8	39.4	79.0	72.3
SR-damp.	86.2	85.3	50.1	86.5	77.0
SR-all	86.8	85.3	48.1	84.4	76.2
Combined	**88.5**	84.2	**53.6**	**87.9**	**78.5**

Overall comparison – CV and Test

- Re-trained 1-way CNNs – overall best-performing system (CV and test)
- Data augmentation further improves it, except in test set RB
- Fine-tuned drums CNN gives highest test set RB f-score

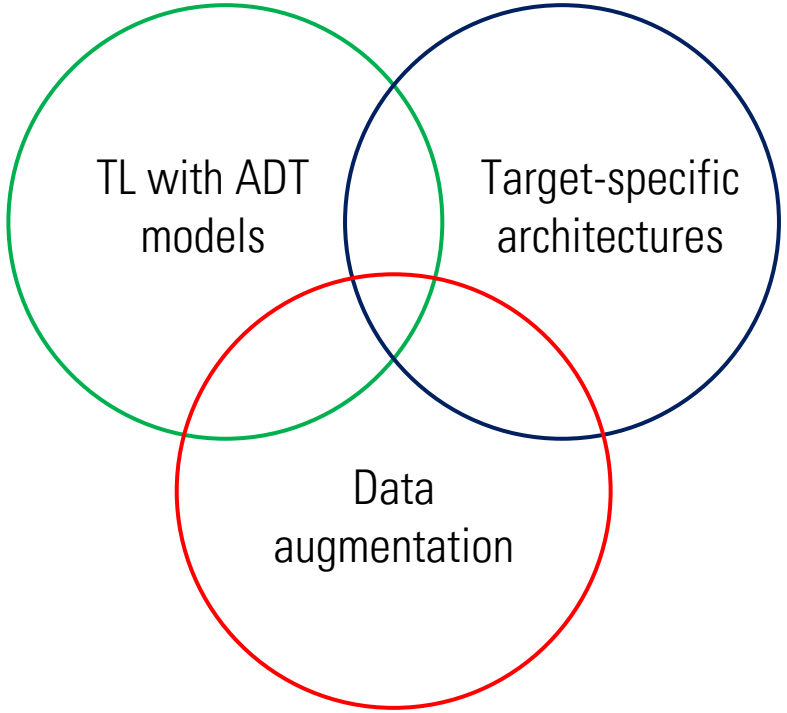
CV / test F-scores

Method	Stroke Category				Mean
	D	RT	RB	B	
RF baseline[14]	86.2 / 74.2	77.7 / 75.0	39.7 / 35.3	73.6 / 41.5	69.3 / 56.5
3-way drums CNN	74.8 / 65.4	66.4 / 77.4	34.7 / **47.5**	66.5 / 56.8	60.6 / 61.8
1-way CNNs	86.0 / 79.5	84.5 / 84.1	46.5 / 38.0	83.8 / 69.0	75.4 / 67.6
+ Data-aug	**88.5 / 83.3**	**85.5 / 84.3**	**53.6** / 34.1	**87.9 / 80.1**	**78.9 / 70.4**

[14] M. A. Rohit and P. Rao, "Automatic stroke classification of tabla accompaniment in hindustani vocal concert audio". JASI 2021

Take-aways

- Addressed 4-way tabla stroke classification into musicologically relevant categories
- Introduced diverse, realistic dataset by building on existing ones
- Showed promising results using different approaches that can be brought together



Thank you for your attention!

Questions?