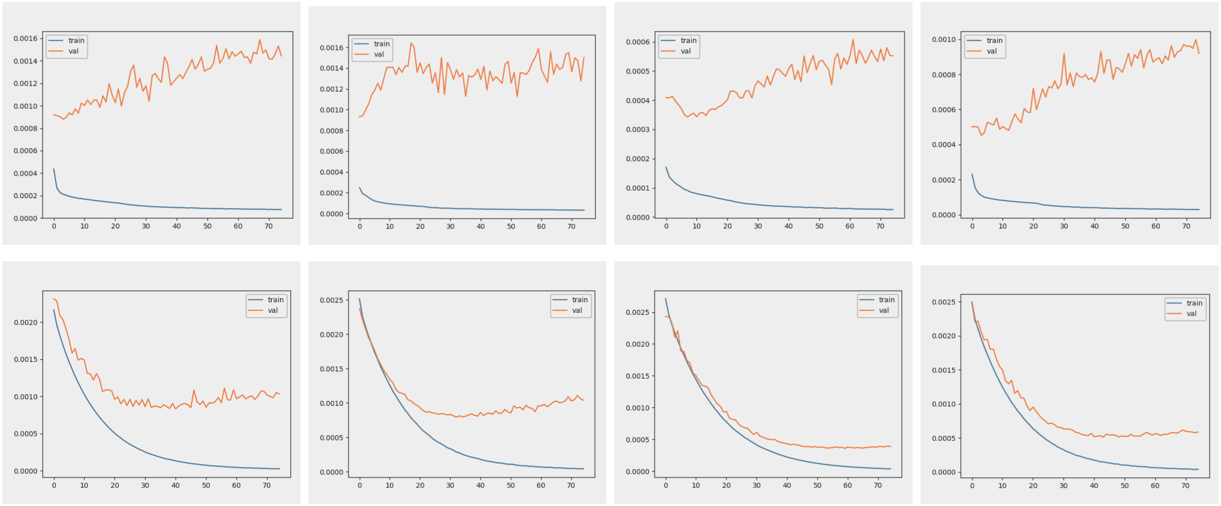
Contents

1. Arriving at the Baseline 1-way CNN

Our 1-way CNN models are based on a previous model from ADT [1]. This (previous) model consists of two convolutional layers with ReLU activations, each followed by a maxpool layer. The output of these layers is passed to a dense (fully-connected) layer with ReLU activation and a final output dense layer with sigmoid activation, with a dropout layer ($p = 0.5$) in between to combat overfitting.

Upon training this model on our training dataset, severe overfitting was observed. The following model variations were thus attempted to address this, while keeping the model performance from reducing drastically:

1. The #hidden units in the dense layer which is more likely to cause overfitting, were reduced
2. Convolutional layer filters were increased, to help lower layers learn features better
3. BatchNorm was introduced after every conv and dense layer (before every activation func.)
4. Additional dropout was added before the first dense layer
5. Dropout probability set to a lower value of 0.25 for both dropout layers



(a)Damped (b) Resonant Treble (c) Resonant Bass (d) Resonant both

***Figure 1.1**: Training (blue) & validation (orange) loss curves for one of the 3-fold CV splits. Top: Using model architecture from [1]. Bottom: After modifications.*

Figure 1.1 shows the loss curves from one of the 3-fold CV runs for each stroke category. We clearly see that our modifications (plots in the bottom row) result in reduced overfitting and better convergence of the validation losses.

2. Tabla Identification Task

For this task we utilised a random forest system similar to the one used previously for 4-way tabla stroke classification [2]. The model is trained on a set of 49 acoustic features, which include frame-wise features calculated from short-time spectra and aggregated using statistical measures, as well as other features computed from the overall signal envelope. Results from the 10-way tabla instrument identification (classification) task, performed separately using samples from each of the 4 categories appear in Table 3.1. Models were trained and evaluated using a random 3-fold cross-validation with 3 randomly seeded repetitions in each case. Figure 3.1 shows the full list of feature importances obtained from the trained models.

Stroke category	Damped	Resonant Treble	Resonant Bass	Resonant Both
Mean F-score (std. dev.)	92.82 (0.39)	94.36 (0.83)	90.15 (1.32)	95.55 (0.66)

Table 3.1: *F-scores from the 10-way tabla identification/classification task. Scores are averaged across the 3 folds and 3 randomly seeded repetitions (values in parentheses are the standard deviations).*

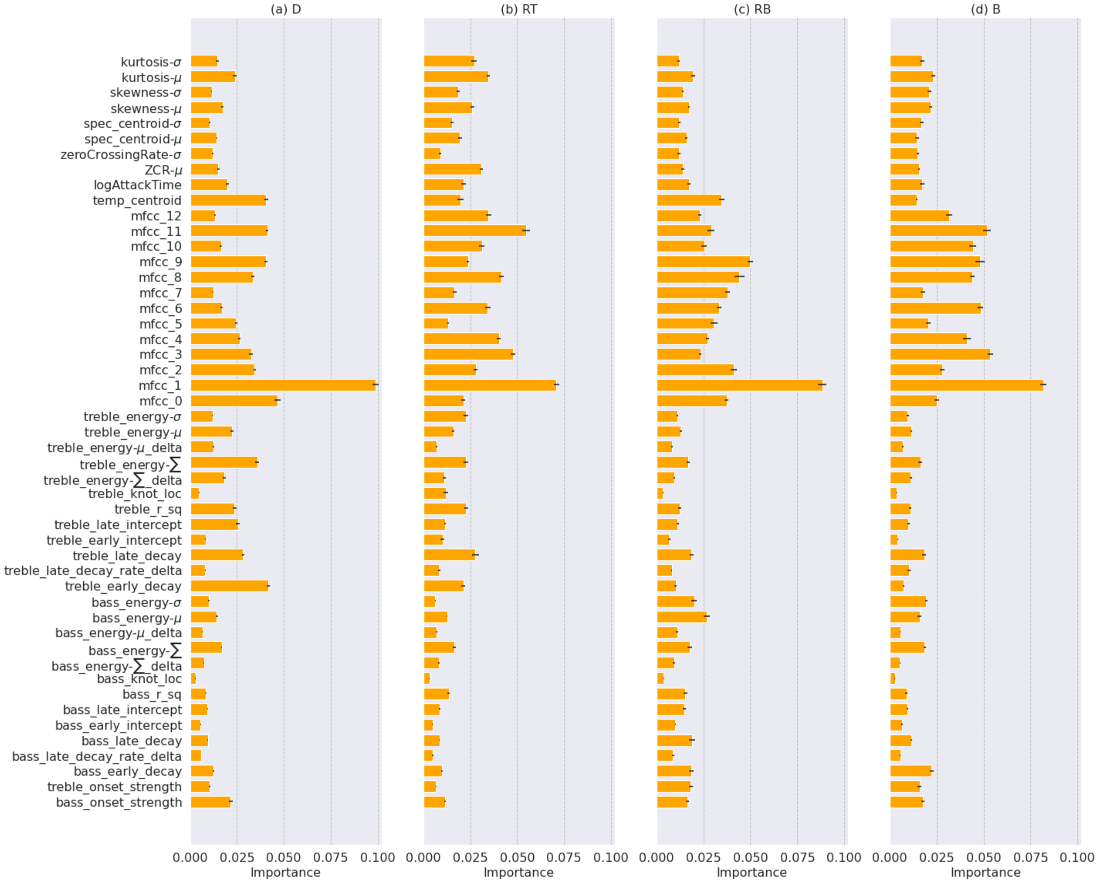


Figure 3.1: *Full list of feature importances (bars are mean, whiskers are std. dev.) from the models trained on each stroke category.*

3. Visualising NMF Templates

Templates are extracted by separately averaging the attack and decay portions of 10 instances for each of the 3 distinct stroke types. The use of separate templates for the attack and decay is motivated by the contrasting nature of their spectral characteristics in tabla strokes. It is also inspired from the previous use of such a strategy for western drums [3].

The 10 instances for template extraction are chosen such that they contain no interfering sustained sounds from previous strokes. Further, given the timbral diversity within each stroke type (due to the various bols in each category), the instances are chosen suitably to represent this variability (e.g., instances from ‘Na’ & ‘Tun’ are both included for resonant treble). Figure 4.1 is a plot of the templates obtained for one of the tabla instruments in our training dataset.

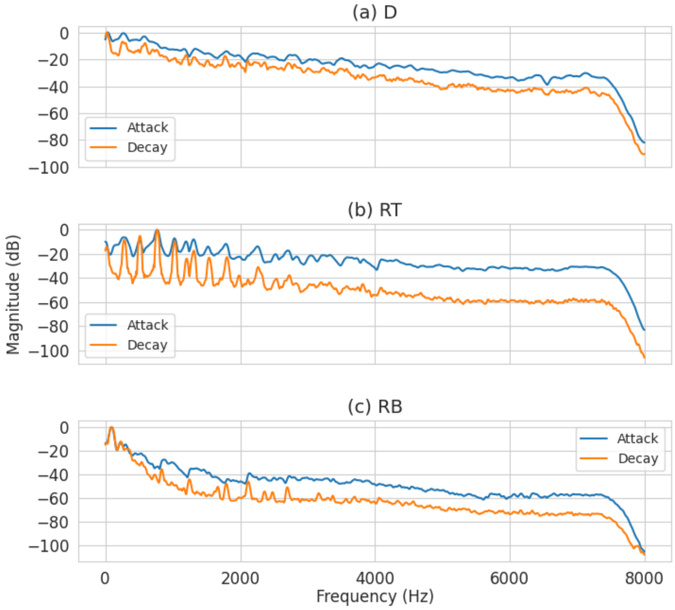


Figure 4.1: *Attack and decay templates for each stroke type as extracted for one of the training set tablas, to be used for NMF decomposition.*

Note how there is a difference of nearly 10dB in magnitude between the attack and decay templates in frequencies beyond 3-4 kHz. The attack and decay spectra appear more similar for the damped strokes, when compared to the other two. The decay template has sharp peaks coinciding with the harmonics for resonant treble, while it has a major peak at a very low frequency value in the case of resonant bass.

4. Illustrating the Augmentation Methods

Figures 5.2 - 5.4 show spectrograms (0-4kHz) of a modified version each of a 2.5 seconds long portion of a tabla audio from the training dataset. The modifications are from the augmentation methods *attack-remixing*, *spectral-filtering-all* and *stroke-remixing-all*, respectively. In the filtering method, spectral tilt modification affects the entire bandwidth shown in the plot (0-4kHz) of the percussive component. The remaining half of the spectrum is left unchanged (not shown for better visualisation).

Figure 5.1 shows the original clip. Corresponding audios are also provided in the supplementary material. The clip contains a sequence of the following 7 strokes: RT, D, D, B, B, RT, B.

The factor ‘α’ in each case controls the amount of linear scaling applied:

- In attack remixing, the percussive component is scaled by this factor before remixing with the harmonic component.
- In spectral filtering, the hann window multiplied with the corresponding spectral band in each short-time spectral slice is scaled by this factor
- In stroke-remixing, the corresponding separated stroke component is scaled by this factor before remixing with the other components.

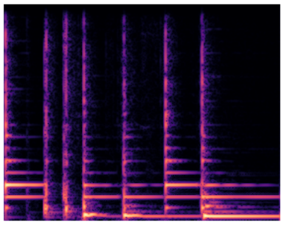
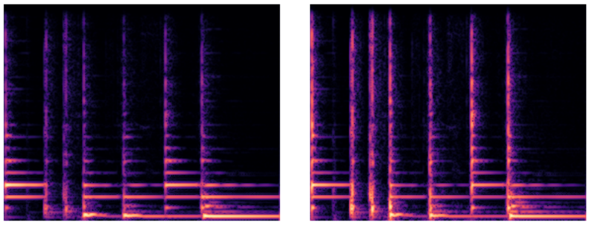
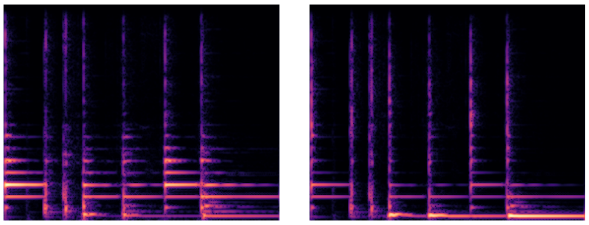


Figure 5.1: *Original*



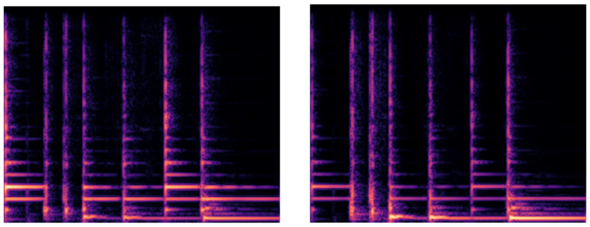
(a) (b)

Figure 5.2: *Attack remixing with α_{ar} at (a) 0.3 & (b) 3*



(a) (b)

Figure 5.3: *Spectral filtering - all with {$\alpha_{sf\text{-}bass}$, $\alpha_{sf\text{-}treble}$, $\alpha_{sf\text{-}tilt}$} at (a) {0.2, 4.0, 0.2} and (b) {4.0, 0.2, 2.0}*



(a) (b)

Figure 5.4: *Stroke remixing - all with {$\alpha_{sr\text{-}bass}$, $\alpha_{sr\text{-}treble}$, $\alpha_{sr\text{-}damp}$} at (a) {0.6, 2.0, 0.2} and (b) {2.0, 0.5, 3.0}*

1. In Figure 5.2, the effect of scaling the percussive component (attack) is apparent, with the attack being weaker in (a) and much stronger in (b).
2. In spectral-filtering (Figure 5.3), the bass & treble regions appear correspondingly scaled, while the entire attack portion of each stroke appears modified depending on the value of $\alpha_{sf\text{-}tilt}$.
3. Comparing filtering and stroke-remixing, we notice how they are different from each other. For instance, while SR-damp mainly affects the two damped strokes in the sequence, SF-tilt affects the spectral tilt in the attack portion of all strokes.

6. References

[1] C. Jacques and A. Röbel, “Automatic drum transcription with convolutional neural networks,” in Proc. of the 21th Int. Conf. on Digital Audio Effects, Aveiro, Portugal, 2018.

[2] M. A. Rohit and P. Rao, “Automatic stroke classification of tabla accompaniment in hindustani vocal concert audio,”To appear in Journal of Acoustical Society of India, April 2021.

[3] Battenberg, Eric, Victor Huang, and David Wessel. "Live drum separation using probabilistic spectral clustering based on the Itakura-Saito divergence." in Proc. of the AES 45th Conf. on Time-Frequency Processing in Audio, Helsinki, Finland. 2012.