ONSET DETECTION IN COMPOSITION ITEMS OF CARNATIC MUSIC

Jilt Sebastian
Indian Institute of Technology, Madras
jiltsebastian@gmail.com

Hema A. Murthy
Indian Institute of Technology, Madras
hema@cse.itm.ac.in

ABSTRACT

Complex rhythmic patterns associated with Carnatic music are revealed from the *stroke* locations of percussion instruments. However, a comprehensive approach for the detection of these locations from composition items is lacking. This is a challenging problem since the melodic sounds (typically vocal and violin) generate soft-onset locations which result in a number of false alarms.

In this work, a separation-driven onset detection approach is proposed. Percussive separation is performed using a Deep Recurrent Neural Network (DRNN) in the first stage. A single model is used to separate the percussive vs the non-percussive sounds using discriminative training and time-frequency masking. This is then followed by an onset detection stage based on group delay (GD) processing on the separated percussive track. The proposed approach is evaluated on a large dataset of live Carnatic music concert recordings and compared against percussive separation and onset detection baselines. The separation performance is significantly better than that of Harmonic-Percussive Separation (HPS) algorithm and onset detection performance is better than the state-of-the-art Convolutional Neural Network (CNN) based algorithm. The proposed approach has an absolute improvement of 18.4% compared with the detection algorithm applied directly on the composition items.

1. INTRODUCTION

Detecting and characterizing musical events is an important task in Music Information Retrieval (MIR), especially in Carnatic music, which has a rich rhythm repertoire. There are seven different types of repeating rhythmic patterns known as *tāla*s, which when combined with 5 *jāti*s give rise to 35 combinations of rhythmic cycles of fixed intervals. By incorporating 5 further variations called *gati*/*nadai*, 175 rhythmic cycles are obtained [13]. A *tāla* cycle is made up of *mātrā*s, which in turn are made up of *aksharā*s or strokes at the fundamental level. Another complexity in Carnatic music is that the start of the *tāla* cycle and of the composition need not be synchronous. Nevertheless, percussion keeps track of *rhythm*. The detection of percussive syllable locations aids higher level retrieval tasks such as *aksharā* transcription, *sama* (start of *tāla*) and *eḍuppu* (start of composition) detection and *tāla* tracking.

Various methods have been proposed for detecting onsets from music signals using a short-term signal, the linear prediction error signal, spectral magnitude or phase, energy and their combination [1, 3, 11, 14, 15]. In [2], various acoustic features are analyzed for this task and in [7], spectral methods are modified to enable onset detection. These and other algorithms are analyzed in detail in [5]. Recent efforts include the use of Recurrent (RNN) [17] and Convolutional Neural Networks (CNN) [19] for onset detection. All of the above techniques are primarily for the detection of monophonic musical onsets.

Every item in Carnatic music has, at its core, a composition. Every item in a concert is characterized by three sections. A lyrical composition section that is performed together by the lead performer, accompanying violinist and the percussion artist. This section is optionally preceded by a pure melody section (*ālāpana*) in which only the lead performer and the accompanying violinist perform. The composition section is optionally followed by a pure percussion section (*tani āvarthanam*). Onset detection and *aksharā* transcription in *tani āvarthanam*s are performed in [15], and [16] respectively. Percussive onset detection for an entire concert that is made up of 10-12 items, each associated with its own *tāla* cycle, is still challenging as the composition items are made up of ensembles of a lead vocal, violin/ensembles of the lead instrument(s) and percussion.

Onset detection in polyphonic music/ensemble of percussion either use audio features directly [4], or performs detection on the separated sources. Dictionary learning-based methods using templates are employed in the separation stage in certain music traditions [10, 22]. Harmonic/percussive separation (HPS) from the audio mixture is successfully attempted on Western music in [8] and [9]. Onset detection of notes is performed on polyphonic music in [4] for transcription. Efficient percussive onset detection on monaural music mixtures is still a challenging problem. The current approaches lead to a significant number of false positives, owing to the difficulty in detecting only the percussive syllables with varying amplitudes and the presence of melodic voices.

In a Carnatic music concert, the lead artist and all the accompanying instruments are tuned to the same base frequency called *tonic* frequency and it may vary for each

Attribution: Jilt Sebastian, Hema A. Murthy. "ONSET DETECTION IN COMPOSITION ITEMS OF CARNATIC MUSIC", 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017.

concert. This leads to the overlapping of pitch trajectories. The bases do not vary over time in the case of dictionary-based separation methods, leading to a limited performance in Carnatic music renderings. HPS model [8] does not account for the melodic component and variation of *tonic* across the concerts. The state-of-the-art solo onset detection techniques, when applied to the polyphonic music, perform poorer ($\approx$ 20% absolute) than on the solo samples [22].

In this paper, a separation-driven approach for percussive onset detection is presented. A deep recurrent model (DRNN) is used to separate the percussion from the composition in the first stage. It is followed by the onset detection based on signal processing in the final stage. The proposed approach achieves significant improvement (18.4%) over the onset detection algorithm applied to the mixture and gracefully degrades (about 4.6% poorer) with respect to onset detection on solo percussion. The proposed approach has better separation and detection performance, when compared to that of the baseline algorithms.

2. DATASETS

Multi-track recordings of six live vocal concerts ($\simeq$ 14 hours) are considered for extracting the composition items. These items contain composition segments with vocal and/or violin segments in first track and percussive segments in the second track. To create the ground truth, onsets are marked (manually by the authors) in the percussive track. These onsets are verified by a professional artist [1] . Details of the datasets prepared from various concerts are given in Table 1. The composition items consist of recordings from both male and female artists sampled at 44.1 kHz. Some of the strokes in the mridangam are dependent on the tonic, while others are not. The concerts SS and KD also include *ghatam* and *khanjira*, which are secondary percussion instruments. Recordings are also affected by nearby sources, background applauses and the perpetual *drone*.

Concert	Total Length hh:mm:ss	Comp. Segments mm:ss (Number)	No. of Strokes
KK	2:15:50	1:52 (3)	541
SS	2:41:14	0:38(4)	123
MH	2:31:47	1:16 (3)	329
ND	1:15:20	1:51 (3)	330
MO	2:00:15	7:14 (3)	1698
KD	2:20:23	5:32 (3)	1088
Total	13:41:59	18:23 (19)	4109

Table 1: Details of the dataset

Training examples for the percussion separation stage are obtained from the *ālāpana* (vocal solo, violin solo) and mridangam *tani āvarthanam* segments. These are mixed to create the polyphonic mixture. A total of 12 musical clips are extracted from four out of six recordings, to obtain the training set (17min and 5s), and the validation set (4min and 10s). Hence, around 43% of the data is found to be sufficient for training. 10% of the dataset is used for the validation of neural network parameters and the rest for testing the separation performance. The concert segments KK and ND are only used for testing the proposed approach to check the generalizability of the approach across various concerts. The composition segments shown in Table 1 column 3 (with ground truth) are used as the test data. Onset detection is then performed on the separated percussive track.

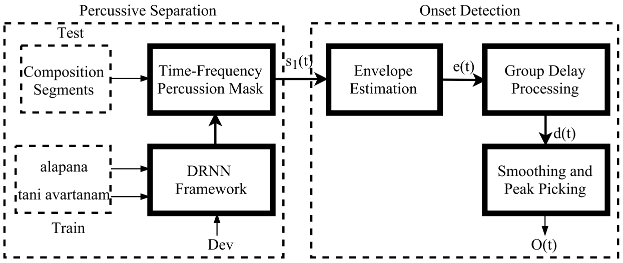


Figure 1: Block diagram of the proposed approach.

3. PROPOSED APPROACH

The proposed method consists of two stages: percussive separation stage and solo onset detection stage. Initially, the time-frequency masks specific to percussive voices (mainly mridangam) are learned using a DRNN framework. The separated percussion source is then used as input to the onset detection algorithm. Figure 1 shows the block diagram of the overall process which is explained subsequently in detail.

3.1 Percussive Separation Stage

A deep recurrent neural network framework originally proposed for singing voice separation [12] is adopted for separating the percussion from the other voices. *Ālāpana* segments are mixed with *tani āvarthanam* segments for learning the timbral patterns corresponding to each source. Figure 2 shows the time-frequency patterns of the composition mixture segment, melodic mixture and the percussive source in Carnatic music. The patterns associated with different voices are mixed in composition segments leading to a fairly complex magnitude spectrogram (Figure 2 *left*) which makes separation of percussion a nontrivial task. The DRNN architecture for percussive separation stage is shown in Figure 3. The network takes the feature vector corresponding to the composition items (x_t) and estimates the mask corresponding to the percussive ($y_t'^1$) and non-percussive ($y_t'^2$) sources. The normalized mask corresponding to the percussive source ($M_1(f)$) is used to filter the mixture spectrum and then combined with the mixture phase to obtain the complex-valued percussive spectrum:

$$\widehat{S_p}(f) = M_1(f)X_t(f) \quad (1)$$

$$S_p(t) = ISTFT(\widehat{S_p}\angle X_t) \quad (2)$$

[1] Thanks to musician Dr. Padmasundari for the verification

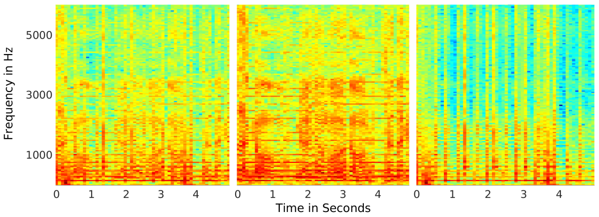


Figure 2: Spectrograms of a segment of composition ***(left)*** obtained from the mixture (KK dataset) containing melodic sources, vocal and violin ***(middle)*** and the percussive source ***(right)***.

where, ISTFT refers to inverse short-time Fourier transform, $\widehat{S_p}$ is the estimated percussive spectrum, $\angle(X_t)$ is the mixture phase at time t and, $S_p(t)$ is the percussive signal estimated for t^{th} time frame.

We use the short-time Fourier transform (STFT) feature as it performs better than conventional features in musical source separation tasks [21]. The regression problem of finding the source specific-magnitude spectrogram is formulated as a binary mask estimation problem where each time-frequency bin is classified as either percussive or non-percussive voice. The network is jointly optimized with the normalized masking function ($M_1(f)$) by adding an extra deterministic layer to the output layer. We use a single model to learn both these masks despite the fact that only percussive sound is required in the second stage. Thus, discriminative information is also used for the learning problem. The objective function (Mean Squared Error) that is minimized is given by:

$$||\widehat{y}_{1t} - y_{1t}||^2 + ||\widehat{y}_{2t} - y_{2t}||^2 - \gamma(||\widehat{y}_{1t} - y_{2t}||^2 + ||\widehat{y}_{2t} - y_{1t}||^2) \quad (3)$$

where $\widehat{y}_t$ and y_t are the estimated and original magnitude spectra respectively. The γ parameter is optimized such that more importance is given to minimizing the error for the percussive voices than maximizing the difference with respect to the other sources. This is primarily to ensure that the characteristics of percussive voice are not affected significantly by separation, as the percussive voice will be used later for onset detection. The recurrent connections are employed to capture the temporal dynamics of the percussive source which are not captured using the contextual windows. The network has a recurrent connection at the second hidden layer and is parametrically chosen based on the performance on development data. The second hidden layer output is calculated from the current input and output of the same hidden layer in the previous time-step as:

$$h^2(x_t) = f(W^2h^2(x_t) + b^2 + V^2h^2(x_{t-1})) \quad (4)$$

where, W and V are the weight matrices, V being the temporal weight matrix and the function $f(\cdot)$ is the ReLU activation [12].

A recurrent network trained with *Ālāpana* and *tani āvarthanam* separates the percussion from the voice by generating a time-frequency percussive mask. This mask is used to separate the percussive voice in the composition segment of a Carnatic music item. The separated signal is used for onset detection in the next stage (Figure 1).

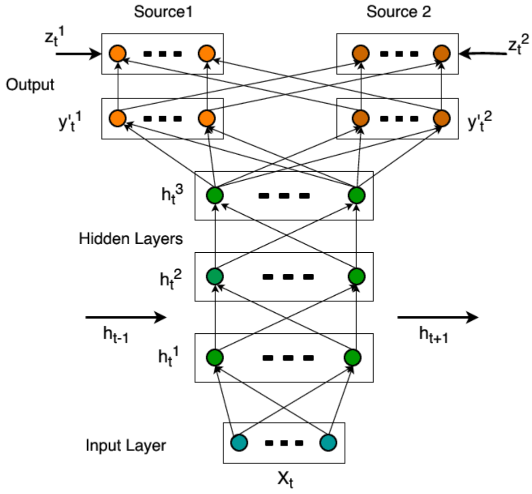


Figure 3: Percussive separation architecture [2]

3.2 Onset Detection Stage

The separated percussive voice is used as the source signal for the onset detection task. Note that this signal has other source interferences, artifacts and other distortions. The second block in Figure 1 corresponds to the onset detection stage. Onset detection consists of two steps. In the first step a detection function is derived from the percussive strokes which is then used in onset detection in the second step.

It is observed that the percussive strokes in Carnatic music can be modeled by an AM-FM signal based on amplitude and frequency variations in the vicinity of an onset [15]. An amplitude and frequency modulated signal ($x(t)$) is given by,

$$x(t) = m_1(t)cos(\omega_c t + k_f \int m_2(t)dt) \quad (5)$$

where, k_f is the frequency modulation factor, ω_c is the carrier frequency and, $m_1(t)$ and $m_2(t)$ are the message signals. The changes in the frequency are emphasized in the amplitude of the waveform by finding the differences of the time-limited discrete version of the signal, $x[n]$. The envelope function $e[n]$ is the amplitude part of $x^{'}[n]$. The real-valued envelope signal can be represented by the corresponding analytic signal defined as:

$$e_a[n] = e[n] + ie_H[n] \quad (6)$$

$e_H[n]$ is the Hilbert transform of the envelope function. The magnitude of $e_a[n]$ is the detection function for the onsets. The high-energy positions of the envelope signal ($e[n]$) corresponds to the onset locations. However, these positions have a large dynamic range and the signal has a limited temporal resolution. It has been shown in [20] that minimum-phase group delay (GD) based smoothing

[2] Example redrawn from [12]

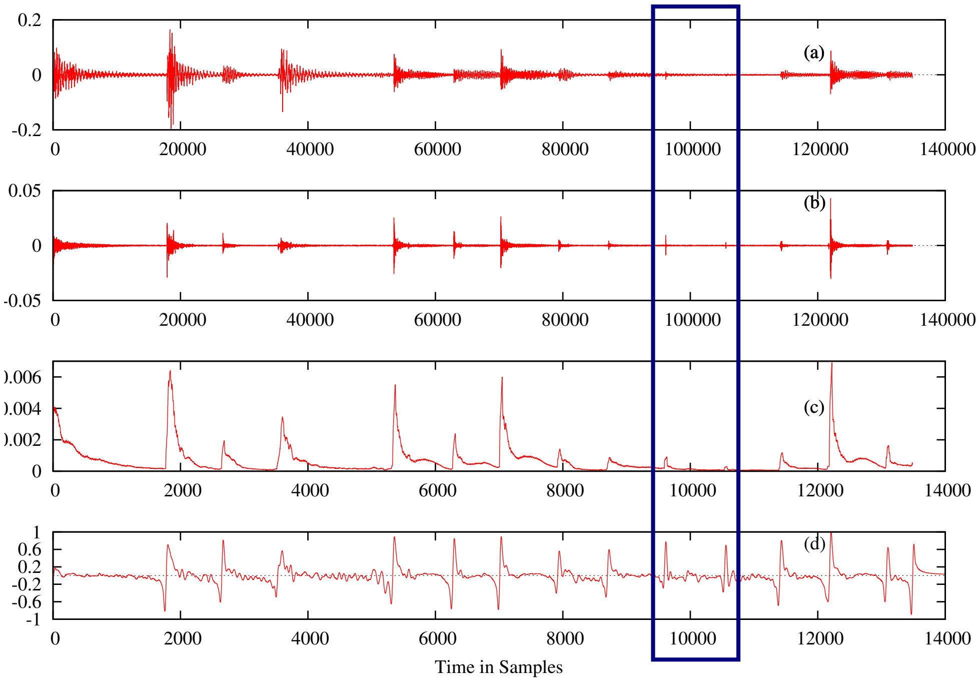


Figure 4: Solo onset detection algorithm. (a) Percussion signal (b) Derivative of (a). c) Envelope estimated on (b) using Hilbert transform. (d) Minimum phase group delay computed on (c).

leads to a better resolution for any positive signal that is characterized by peaks and valleys. The envelope is a non-minimum phase signal and it needs to be converted to a minimum phase equivalent to apply this processing.

It is possible to derive such an equivalent representation with a root cepstral representation. The causal portion of the inverse Fourier transform of the magnitude spectrum raised to a power of α is always minimum phase [18].

$$e^{'}[k] = \{s[k] \mid_{k>0},\ s[k] = IFT((e(n) + e[-n])^{\alpha})\} \quad (7)$$

Note that $e^{'}[k]$ is in root cepstral domain and k is the quefrency index. This minimum-phase equivalent envelope is then subjected to group delay processing.

The group delay is defined as negative frequency derivative of the unwrapped phase function. It can be computed directly from the cepstral domain input signal $e^{'}[k]$ as:

$$\tau(\omega) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + X_I(e^{j\omega})Y_I(e^{j\omega})}{|X(e^{j\omega})|^2} \quad (8)$$

where, $X(e^{j\omega})$ and $Y(e^{j\omega})$ are the discrete Fourier transforms of $e^{'}[k]$ and $ne^{'}[k]$ respectively. Also, R and I denote the real and imaginary parts respectively. The high resolution property of the group delay domain emphasizes the onset locations. Onsets are reported as instants of significant rise, above a threshold.

Figure 4 illustrates the different steps in the algorithm using a mridangam excerpt taken from a *tani āvarthanam* segment. It is interesting to note that in the final step, the group delay function emphasizes all the strokes approximately to an equal amplitude, and even those onsets in which there is no noticeable change in amplitude are also obtained as peaks (highlighted area in Figure 4).

4. PERFORMANCE EVALUATION

The proposed percussive onset detection approach is developed specifically for rhythm analysis in Carnatic music composition items. However, it is instructive to compare the performance with other separation and onset detection algorithms. Also, it is important to note that the proposed approach could be applied to any music tradition with enough training musical excerpts to extract the onset locations from the polyphonic mixture. The dataset for these tasks is described in Section 2. The vocal-violin channel (*ālāpana*) and the percussion channel (*tani āvarthanam*) are mixed at 0 dB SNR. The STFT with a window length of 1024 samples and hop size of 512 samples is used as the feature for training a DRNN with 3 hidden layers (1000 units/layer) and temporal connection at the 2^{nd} layer. This architecture shows a very good separation for the singing voice separation task [12]. The dataset consists of segments with varying tempo, loudness and number of sources at a given time. The challenge lies in detecting the onsets in the presence of the interference caused by other sources and the background voices.

4.1 Evaluation Metrics

Since the estimation of percussive onsets also depends on the quality of separation, it is necessary to evaluate the separated track. We measure this using three quantitative measures based on BSS-EVAL 3.0 metrics [23]: Source to Artifacts Ratio (SAR), Source to Interference Ratio (SIR) and Source to Distortion Ratio (SDR). The artifacts introduced in the separated track is measured by SAR. The suppression achieved for the interfering sources (vocal and violin) is represented in terms of SIR which is an indicator of the timbre differences between the vocal-violin mixture and percussive source. SDR gives the overall separation quality. The length-weighted means of these measures are used for representing the overall performance in terms of global measures (GSAR, GSIR and GSDR).

The conventional evaluation metric for the onset detection is F-measure, which is the harmonic mean of precision and recall. An onset is treated as correct (*True Positive*) if it is reported within a ±50ms threshold of the ground truth [6] as strokes inside this interval are usually unresolvable. Additionally, this margin accounts for the possible errors in the manual annotation. The F-measure is computed from sensitivity and precision. Since it is impossible to differentiate between simple and composite[3] strokes for mridangam, the closely spaced onsets (within 30 *ms*) are not merged together unlike in [5].

4.2 Comparison Methods

The performance of the separation stage is compared with a widely used Harmonic/Percussive Separation (HPS) algorithm [8] for musical mixtures. It is a signal processing-based algorithm in which median filtering is employed on the spectral features for separation. Other supervised percussive separation models were specific to the music traditions. We have not considered the Non negative Matrix Factorization (NMF)-based approaches since the separation performance was worse on Carnatic music, hinting the inability of a constant dictionary to capture the variability across the percussive sessions and instruments.

The onset detection performance is compared with the state-of-the-art CNN-based onset detection approach [19]. In this approach, a convolutional network is trained as a binary classifier to predict whether the given set of frames has an onset or not. It is trained using percussive and non percussive solo performances. We evaluate the performance of this algorithm on the separated percussive track and, on the mixture . The onset threshold amplitude is optimized with respect to the mixture and percussive solo channel for evaluating the performance on the separated and mixture tracks respectively for both of these algorithms.

5. RESULTS AND DISCUSSION

5.1 Percussive Separation

The results of percussive separation are compared with that of the HPS algorithm in Table 2. The large variability of the spectral structure with respect to the *tonic*, strokes and the percussive instruments (different types of mridangam as well) cause the HPS model to perform poorly with respect to the proposed approach. The DRNN separation benefits from the training whereas the presence of the melodic component with rich harmonic content adds to the interference in HPS method. This results in a poor separation of melodic mixture and percussive voice in HPS approach as indicated by an overall difference of 5.51 dB SDR with respect to DRNN approach. Although DRNN is not trained on the concerts KK and MD, separation measures are quite similar to other concerts. This is an indicator of the generalization capability of the network since each concert is of a unique *tonic* (base) frequency, and is recorded under a different environment. Separated sound examples are available online[4] .

	DRNN			HPS		
Concert	**GSDR**	**GSIR**	**GSAR**	**GSDR**	**GSIR**	**GSAR**
SS	7.00	13.70	8.61	3.39	6.73	7.93
ND	7.54	17.30	8.98	0.46	3.05	7.67
KK	7.37	13.93	8.93	0.66	2.04	10.09
MH	6.40	15.64	7.63	0.82	3.31	7.79
KR	7.37	13.93	8.93	1.32	2.43	9.09
MD	6.40	15.64	7.63	2.40	8.06	4.78
Average	**7.01**	15.02	8.45	**1.50**	4.27	7.89

Table 2: Percussive separation performance in terms of BSS evaluation metrics for the proposed approach and HPS algorithm

5.2 Onset Detection

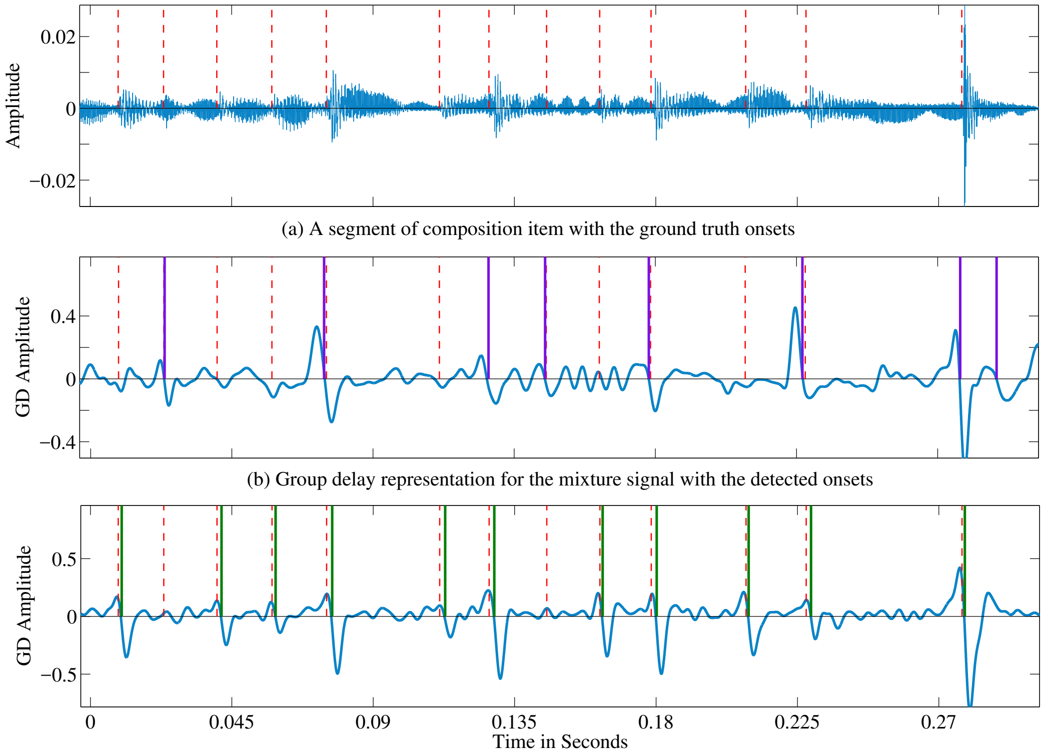
Concert	Proposed	Direct	Solo	CNN	CNN Sep.
SS	0.747	0.448	0.864	0.685	0.656
ND	0.791	0.650	0.924	0.711	0.740
KK	0.891	0.748	0.972	0.587	0.636
MH	0.874	0.687	0.808	0.813	0.567
KR	0.891	0.748	0.972	0.859	0.848
MD	0.874	0.687	0.808	0.930	0.919
Average	**0.845**	**0.661**	0.891	0.764	**0.727**

Table 3: Comparison of F-measures for the proposed approach, direct onset detection on the mixture, solo percussion channel, CNN on the mixture and on the separated percussive channel.

The accuracy of onset detection is evaluated using F-measure in Table 3. The performance varies with the dataset and the results with the maximum average F-measure is reported. The degradation in performance with respect to the solo source is only about 4.6%, while the improvement in performance compared to the direct onset detection on the composite source is 18.4%. The separation step plays a crucial role in onset detection of the composition items as the performance has improved for *all* the datasets upon separation. It should be noted that the algorithm performs really well for solo percussive source. This is reason for making comparisons with solo performances. For SS data (Table 1) with fast tempo (owing to multiple percussive voices) and significant loudness variation (Example online[4]), the direct onset method causes a large number of false positives resulting in lower precision whereas the proposed approach results in a reduced number of false positives. Figure 5 shows an example of a

[3] both left and right strokes co-occurring in the mridangam

[4] `https://sites.google.com/site/percussiononsetdetection`



(a) A segment of composition item with the ground truth onsets

(b) Group delay representation for the mixture signal with the detected onsets

(c) Group delay representation for the separated signal with the detected onsets

Figure 5: An excerpt from SS dataset illustrating the performance of the proposed approach with respect to the direct onset detection method. Red dotted lines represent the ground truth onsets, violet (b) and green (c) lines represent the onsets detected on the mixture signal and the separated percussive signal respectively.

composition item taken from the SS dataset. It compares the performance of the proposed approach with that of the onset detection algorithm applied directly on the mixture. By adjusting the threshold of onset, the number of false positives can be reduced. However, it leads to false negatives as shown in Figure 5(b). The proposed approach is able to detect almost all of the actual onset locations (5(c)).

The proposed approach is then compared with the CNN algorithm. The optimum threshold of the solo algorithm for the Carnatic dataset [15] is used to evaluate the performance. The proposed method performs better than the CNN algorithm applied on the mixture (Table 3). This is because the CNN method is primarily for solo onset detection. The performance of the baseline on the separated channel is also compared with the group delay-based method. The threshold is optimized with respect to the performance of the baseline algorithm on the mixture track. The average F-measure of the proposed approach is 11.8% better than that of the CNN-based algorithm. This is because CNN-based onset detection requires different thresholds for different concert segments. This suggests that the GD based approach generalizes better in the separated voice track and is able to tolerate the inter-segment variability. A consistently better F-measure is obtained by the GD based method across all recordings. This separation-driven algorithm can be extended to any music tradition with sharp percussive onsets and having enough number of polyphonic musical ensembles for the training. These onset locations can be used to extract the strokes of percussion instruments and perform *tāla* analysis.

6. CONCLUSION AND FUTURE WORK

A separation-driven approach for percussive onset detection in monaural music mixture is presented in this paper with a focus on Carnatic music. Owing to its tonic dependency and improvisational nature, conventional dictionary-based learning methods perform poorly on percussion separation in Carnatic music ensembles. Vocal and violin segments from the *ālāpana* and mridangam phrases from the *tani āvarthanam* of concert recordings are used to train a DRNN for the percussive separation stage. The separated percussive source is then subjected to onset detection. The performance of the proposed approach is comparable to that of the onset detection applied on the solo percussion channel and achieves 18.4% absolute improvement over its direct application to the mixture. It compares favourably with the separation and onset detection baselines on the solo and separated channels. The onset locations can be used for analyzing the percussive strokes. Using repeating percussion patterns, the *tāla* cycle can be ascertained. This opens up a plethora of future tasks in Carnatic MIR. Moreover, the proposed approach is generalizable to other music traditions which include percussive instruments.

7. ACKNOWLEDGEMENTS

This research is partly funded by the European Research Council under the European Unions Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583). Authors would like to thank the members of Speech and Music Technology Lab for their valuable suggestions.

8. REFERENCES

[1] Juan P Bello, Chris Duxbury, Mike Davies, and Mark Sandler. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters*, 11(6):553–556, 2004.

[2] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.

[3] Juan Pablo Bello and Mark Sandler. Phase-based note onset detection for music signals. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages V–441. IEEE, 2003.

[4] Emmanouil Benetos and Simon Dixon. Polyphonic music transcription using note onset and offset detection. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 37–40. IEEE, 2011.

[5] Sebastian Böck, Florian Krebs, and Markus Schedl. Evaluating the online capabilities of onset detection methods. In *Proc. of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 49–54, 2012.

[6] Sebastian Böck and Gerhard Widmer. Local group delay based vibrato and tremolo suppression for onset detection. In *Proc. of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 361–366, Curitiba, Brazil, November 2013.

[7] Simon Dixon. Onset detection revisited. In *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx)*, pages 133–137, 2006.

[8] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, pages 15–19, 2010.

[9] Derry Fitzgerald, Antoine Liukus, Zafar Rafii, Bryan Pardo, and Laurent Daudet. Harmonic/percussive separation using kernel additive modelling. In *Proc. of the 25th IET Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014)*, pages 35–40, 2014.

[10] Masataka Goto and Yoichi Muraoka. A sound source separation system for percussion instruments. *Transactions of the Institute of Electronics, Information and Communication Engineers (IEICE)*, 77:901–911, 1994.

[11] Masataka Goto and Yoichi Muraoka. Beat tracking based on multiple-agent architecture a real-time beat tracking system for audio signals. In *Proc. of 2nd International Conference on Multiagent Systems*, pages 103–110, 1996.

[12] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. *In Proc. of International Society for Music Information Retrieval (ISMIR)*, pages 477–482, 2014.

[13] M Humble. The development of rhythmic organization in indian classical music. *MA dissertation, School of Oriental and African Studies, University of London.*, pages 27–35, 2002.

[14] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 6, pages 3089–3092. IEEE, 1999.

[15] Manoj Kumar, Jilt Sebastian, and Hema A Murthy. Musical onset detection on carnatic percussion instruments. In *Proc. of 21st National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015.

[16] Jom Kuriakose, J Chaitanya Kumar, Padi Sarala, Hema A Murthy, and Umayalpuram K Sivaraman. Akshara transcription of mrudangam strokes in carnatic music. In *Proc. of the 21st National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015.

[17] Erik Marchi, Giacomo Ferroni, Florian Eyben, Stefano Squartini, and Bjorn Schuller. Audio onset detection: A wavelet packet based approach with recurrent neural networks. In *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, pages 3585–3591, July 2014.

[18] T Nagarajan, V K Prasad, and Hema A Murthy. The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation. In *Speech Communications*, pages 95–101, July 2001.

[19] Jan Schlüter and Sebastian Böck. Improved Musical Onset Detection with Convolutional Neural Networks. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6979–6983, Florence, Italy, May 2014.

[20] Jilt Sebastian, Manoj Kumar, and Hema A Murthy. An analysis of the high resolution property of group delay function with applications to audio signal processing. *Speech Communications*, pages 42–53, 2016.

[21] Jilt Sebastian and Hema A Murthy. Group delay based music source separation using deep recurrent neural networks. In *Proc. of International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IEEE, 2016.

[22] Mi Tian, Ajay Srinivasamurthy, Mark Sandler, and Xavier Serra. A study of instrument-wise onset detection in beijing opera percussion ensembles. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2159–2163, 2014.

[23] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.