

Novel Methods for Time-Event Detection and Source Separation

Ph.D Viva Voce

Jilt Sebastian
(CS13D020)

Supervisor: Prof. Hema A. Murthy



Speech and Music Technology Lab (SMTL)
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai

4th July, 2019

Outline

- Introduction
- **Problem 1:** Analysis of High-Resolution Property of Group Delay Functions
 - Single pole Minimum Phase Systems
 - Multi-pole Minimum Phase Systems
- **Problem 2:** Time-Event Detection (TED) Tasks
 - **Task1:** Pitch Estimation
 - **Task2:** Percussive Onset Detection
 - **Task3:** Spike Estimation from Neuronal Signals
- **Problem 3:** Source Separation Systems
 - Modified Group Delay Feature
 - Signal-to-Signal Neural Networks
 - Hybrid Systems:
 - Percussive Onset Detection from Musical Mixtures
 - Gender ID under Noisy Conditions
- Summary

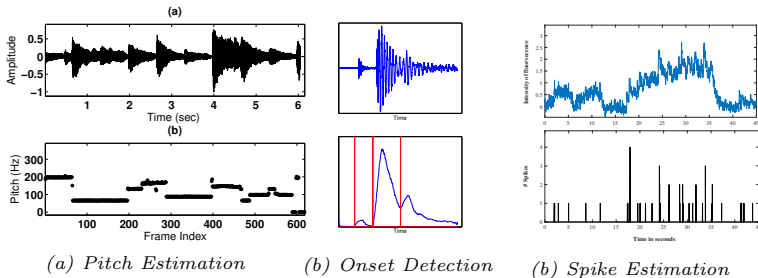


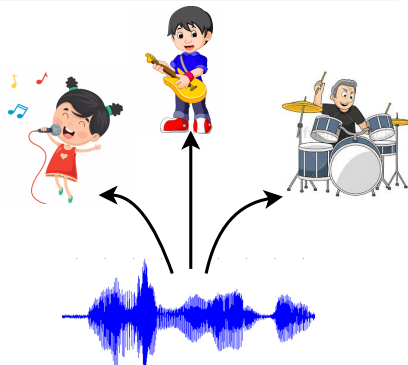
Figure 1: Various TED tasks

- Methods:
 - Spectral or Temporal
 - Signal processing-based or Supervised methods

- **Types:** Single and multi-channel, Additive and convolutive mixtures
- **State-of-the-art:** Deep learning-based methods

Methods

- Features: Magnitude spectrum, logMel
- A neural network learns source-specific time-frequency mask
- Signal is reconstructed using the mixture phase
- **Tasks:** speech separation, singing voice separation, speech enhancement



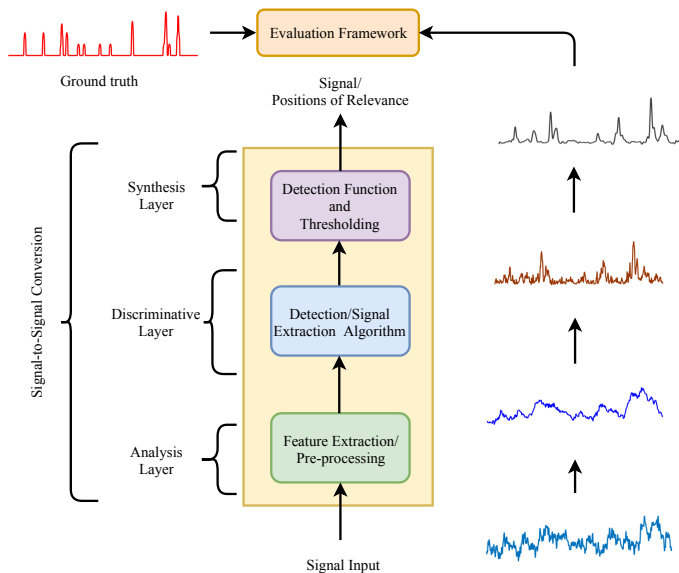


Figure 2: Time-Event Detection (TED) and Signal Extraction tasks

Novel Methods

- Mainly focus on Group delay (GD), a phase-based feature
- **Motivation:** High resolution and additive properties of GD

Applications of TED and Source Separation

- **TED:** As a preprocessing step
 - Pitch Estimation: Text-to-speech synthesis, melody-based recommendation systems
 - Onset Detection: For discovering the rhythmic structure in MIR applications
 - Spike Estimation: Behavioral and cognitive brain analysis
- **Source Separation**
 - Musical Source Separation: Karaoke extraction and preprocessing for other MIRs
 - Speech Enhancement: Automatic speech recognition, noise-robust speech technology systems

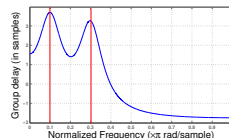
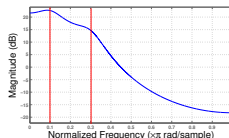
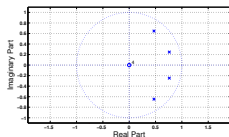
1. Background

- Spectral analysis of signals, phase-based representations

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega}$$

- GD can be directly obtained from the input signal

GD exhibits high resolution and additive properties



- **Limitation:** Noise-like behavior for non-minimum phase signals
- Modifications to GD:
 - Convert signal to minimum phase equivalent-then apply GD processing¹
 - Propose a modified version of group delay (MOD-GD)²

¹Minimum-phase signal derived from the root cepstrum, Nagarajan et. al, IEEE Letters, 2003

²Algorithms for Processing Fourier Transform Phase of Signals, Konda A. G. Murthy, PhD Dissertation

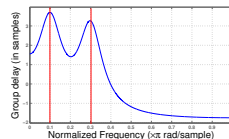
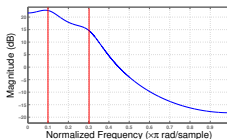
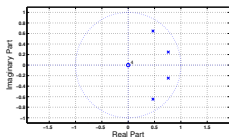
1. Background

- Spectral analysis of signals, phase-based representations

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega}$$

- GD can be directly obtained from the input signal

GD exhibits high resolution and additive properties



- **Limitation:** Noise-like behavior for non-minimum phase signals
- Modifications to GD:
 - Convert signal to minimum phase equivalent-then apply GD processing¹
 - Propose a modified version of group delay (MOD-GD)²

¹ Minimum-phase signal derived from the root cepstrum, Nagarajan et. al, IEEE Letters, 2003

² Algorithms for Processing Fourier Transform Phase of Signals, Hema A. Murthy, PhD Dissertation

Analysis³

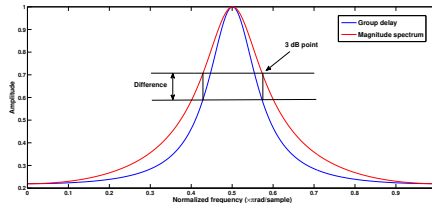
- **Previous work:** Approximated GD as a squared magnitude response around the resonance
- **Observation:** Lower bandwidth \implies more peakedness
- **Analysis:** For single pole systems, n-dB bandwidth of GD is always less than that of magnitude spectrum

Proof:

- n-dB bandwidth of both magnitude spectrum and GD is computed for a single pole system
- GD strength at n-dB bandwidth of magnitude spectrum is always lesser than that of the magnitude spectrum

Numerical analyses

- Single pole: Kurtosis and spectral flatness measures
- Multi pole: Acceleration measure



³Collaborative work with Manoj Kumar P. A.

Analysis³

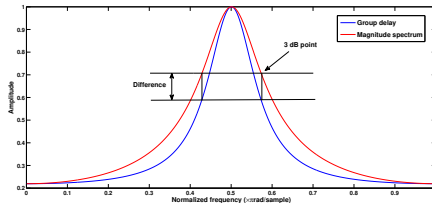
- **Previous work:** Approximated GD as a squared magnitude response around the resonance
- **Observation:** Lower bandwidth \implies more peakedness
- **Analysis:** For single pole systems, n-dB bandwidth of GD is always less than that of magnitude spectrum

Proof:

- n-dB bandwidth of both magnitude spectrum and GD is computed for a single pole system
- GD strength at n-dB bandwidth of magnitude spectrum is always lesser than that of the magnitude spectrum

Numerical analyses

- Single pole: Kurtosis and spectral flatness measures
- Multi pole: Acceleration measure



³Collaborative work with Manoj Kumar P. A.

Analysis³

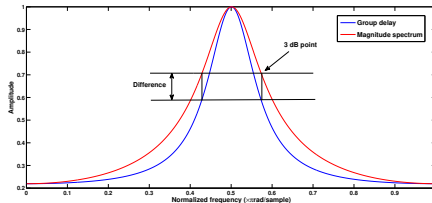
- **Previous work:** Approximated GD as a squared magnitude response around the resonance
- **Observation:** Lower bandwidth \implies more peakedness
- **Analysis:** For single pole systems, n-dB bandwidth of GD is always less than that of magnitude spectrum

Proof:

- n-dB bandwidth of both magnitude spectrum and GD is computed for a single pole system
- GD strength at n-dB bandwidth of magnitude spectrum is always lesser than that of the magnitude spectrum

Numerical analyses

- Single pole: Kurtosis and spectral flatness measures
- Multi pole: Acceleration measure



³Collaborative work with Manoj Kumar P. A.

Related work:

- Pitch histograms are modeled as multi-pole systems ⁴

Analysis:

- Extend the generalized single pole proof to multiple resonator systems
- Factorize the system function to form additive components
- Signal is represented as addition of individual responses in both of the domains
- Multi-pole systems:

Cascade and Parallel connection of resonators

⁴Automatic tonic identification in Indian classical music, Ashwin Bellur, [Masters Thesis](#)

Related work:

- Pitch histograms are modeled as multi-pole systems ⁴

Analysis:

- Extend the generalized single pole proof to multiple resonator systems
- Factorize the system function to form additive components
- Signal is represented as addition of individual responses in both of the domains
- Multi-pole systems:

Cascade and Parallel connection of resonators

⁴Automatic tonic identification in Indian classical music, Ashwin Bellur, [Masters Thesis](#)

1. Cascade Connection

Consider an all-pole system with transfer function:

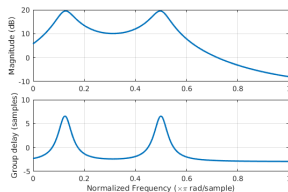
$$H(z) = \prod_{i=1}^n \frac{1}{(1 - z_i z^{-1})(1 - z_i^* z^{-1})} \quad (1)$$

Assuming that the system is obtained as a product of rational one-pole/two-pole systems, by partial fractions, the frequency response can be represented as:

$$H(z) = \sum_{i=1}^n \frac{A_i z^{-1} + B_i}{(1 - z_i z^{-1})(1 - z_i^* z^{-1})} \quad (2)$$

where, A_i and $B_i \forall i = 1 \dots n$ are constant coefficients.

- ❶ Each term corresponds to a pair of complex conjugate poles
- ❷ For every single pole, the magnitude spectrum has a lower resolution than the GD spectrum
- ❸ Fourier transform is distributive over addition



Resolving power for series connection

2. Parallel Connection

System response is the addition of individual Z-transforms. Considering a two-pole system,

$$H(z) = \frac{\alpha_1}{1 - a_1 z^{-1}} + \frac{\alpha_2}{1 - a_2 z^{-1}} \quad (3)$$

Computing the LCM, this can be converted to a cascade of resonators format

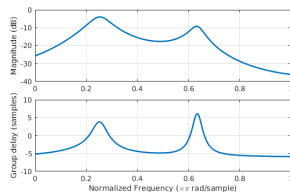
$$H(z) = (\alpha_1 + \alpha_2) \frac{1 - C_1 z^{-1}}{(1 - a_1 z^{-1})(1 - a_2 z^{-1})} \quad (4)$$

where C_1 is a constant.

Excluding the constant $(\alpha_1 + \alpha_2)$, Eqn. (4) can be written as,

$$GD(H(z)) = GD(1 - C_1 z^{-1}) + GD\left(\frac{1}{1 - a_1 z^{-1}}\right) + GD\left(\frac{1}{1 - a_2 z^{-1}}\right) \quad (5)$$

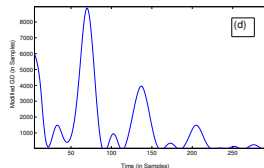
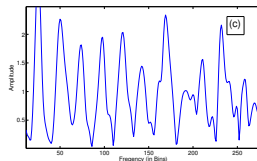
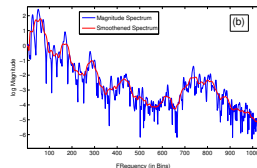
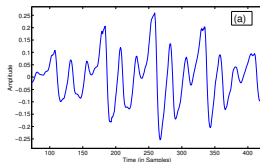
- ❶ The overall GD of the system is the summation of GD of single pole/zero systems
- ❷ GD has high resolution for a two pole system
- ❸ Multi-pole system is an addition of two/one pole systems



Resolving power for parallel connection

Task 1: Pitch estimation

- Fundamental frequency of vibrations of vocal folds
- Estimated using modified-group delay from speech utterances

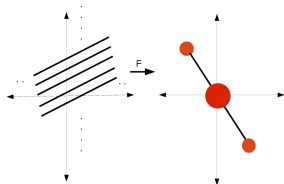


Algorithm: (a) A frame of speech, (b) log Mag. & smoothed spectrum, (c) flattened spectrum, and (d) modified GD spectrum of (c)

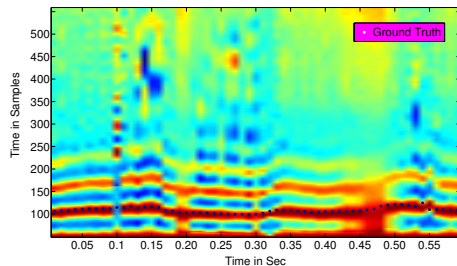
- Datasets: Synthetic and Natural (Keele)
- Very low error per frame, large contribution of error from outliers \implies consider pitch dynamics

Grating Compression Transform (GCT)

- Image processing: 2D Fourier transform of localized T-F regions
- GCT captures pitch dynamics! - Angle in anti-clockwise direction is proportional to pitch slope
- Better 2-D representation than spectrogram for obtaining pitch estimates: MODGDgram representation
- Proposed method: MOD-GD+GCT



Schematic of GCT



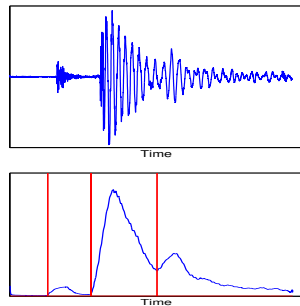
Pitch estimate plotted on modgdgram

- Better than MODGD and Mag+GCT
- Comparable with advanced algorithms

Task 2: Percussive Onset Detection (Music)

Onset Detection

- Beginning of a musical note
- **Challenge:** Variations in tempo, loudness and spectral characteristics
- **Steps:** Detection function extraction & peak picking



(top) Audio with strokes. (bottom) Possible detection function with onsets marked

Percussive Onset Detection⁵

- Focuses on Carnatic music: 5 instruments
- Detection function extraction: Envelope estimation (AM-FM demodulation) + minimum-phase group delay, Peak picking: Hard-threshold
- Created percussive onset detection dataset
- Improvement over traditional features and comparable with the state-of-the-art (CNN⁶)

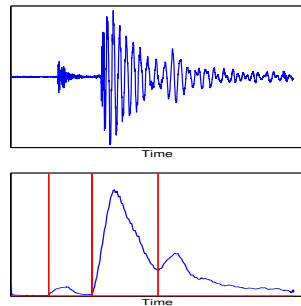
⁵Collaborative work with Manoj Kumar P. A., 2015

⁶Bock et. al., 2014

Task 2: Percussive Onset Detection (Music)

Onset Detection

- Beginning of a musical note
- **Challenge:** Variations in tempo, loudness and spectral characteristics
- **Steps:** Detection function extraction & peak picking



(top) Audio with strokes. (bottom) Possible detection function with onsets marked

Percussive Onset Detection⁵

- Focuses on Carnatic music: 5 instruments
- Detection function extraction: Envelope estimation (AM-FM demodulation) + minimum-phase group delay, Peak picking: Hard-threshold
- Created percussive onset detection dataset
- Improvement over traditional features and comparable with the state-of-the-art (CNN⁶)

⁵Collaborative work with Manoj Kumar P. A., 2015

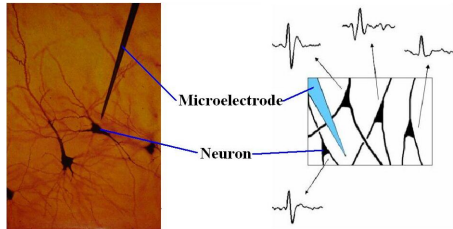
⁶Bock et. al., 2014

Neurons

- Neuronal activities in the brain are observed as spikes
- Essential for functional and behavioral analysis

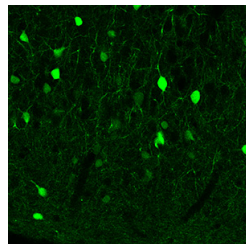
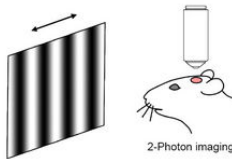
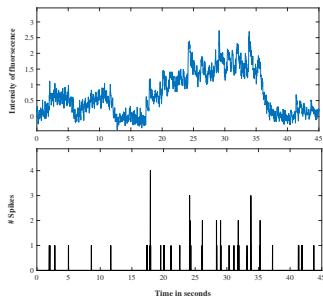
Electrophysiology

- Direct detection via electrodes
- High-temporal resolution
- Invasive, expensive setup
- Poor spatial resolution



Two-Photon Calcium Imaging

- Subject (mice) is either genetically encoded or injected with Ca^{2+} indicators
- Records a population of neurons
- Limited by slow dynamics of fluorescence signals



Generative model-based:

- Assume a model for the generated signal
- Various deconvolution-based techniques
- Limited by the assumptions about the model

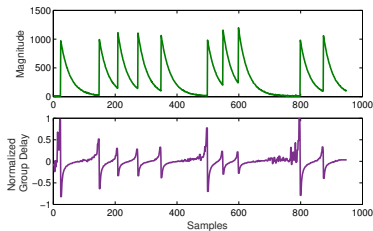
Supervised learning-based:

- Either using features or raw-fluorescence signal
- Neural network-based techniques
- Need simultaneous recordings for training
- Limited performance on unseen data

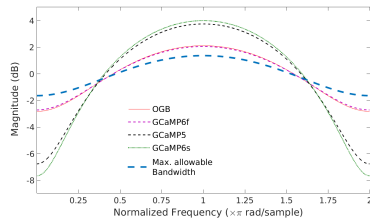
Performance measures

- **Linear correlation (Pearson) coefficient** between true and estimated spike information
- **Area under ROC** as a binary predictor for the presence of spikes
- **F-measure** Harmonic mean of precision and recall
- **Rank** measures the strength and direction of association between two ranked variables

- Motivated by success in other TED tasks
- Converts the signal to minimum-phase and compute group delay
- **Steps:** Detection function estimation and spike train extraction
- Illustrate post processing power of GDspike

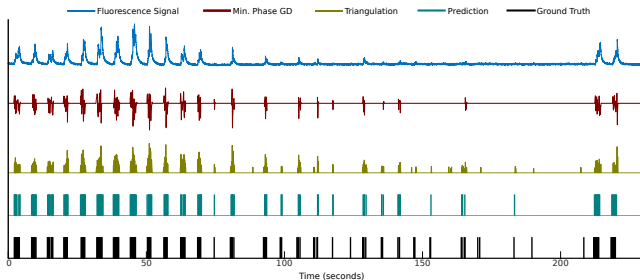
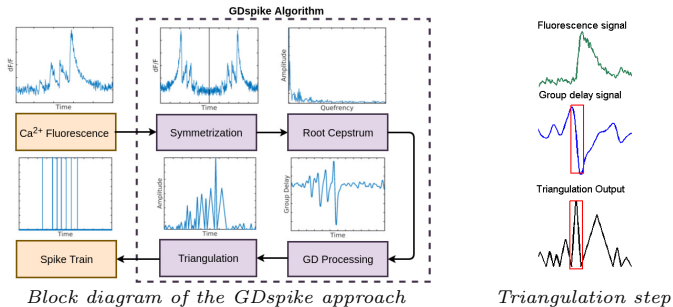


Synthetic calcium signal and its GD representation



Bandwidth of various indicators

Group delay spike: Procedure



Experiment

- **9 Datasets:** 5 are publicly available and 4 are provided by authors of MLspike
- Various indicators, brain regions and scanning rate
- Comparison with MLspike, Vogelstein and STM algorithms⁷

Results

Performance of GDspike

Algo.	Recall	Prec.	F-measure	corr.	AUC
Vogelstein	0.410	0.631	0.433	0.196	0.665
STM	0.452	0.786	0.519	0.335	0.813
MLspike	0.671	0.568	0.543	0.128	0.662
GDspike	0.547	0.640	0.489	0.214	0.680

GDspike post-processing performance

Sl. No	Algorithm	Dataset	Recall	Prec.	F-measure	Corr.	AUC
1	GDspike	Weizman	0.14	0.95	0.23	0.21	0.59
	MLspike		0.90	0.51	0.63	0.07	0.50
	ML+GD		0.36	0.80	0.48	0.28	0.81
2	GDspike	Marselie	0.22	0.85	0.31	0.15	0.64
	MLspike		0.65	0.58	0.54	0.070	0.64
	ML+GD		0.43	0.78	0.48	0.16	0.73
3	GDspike	invitro	0.20	0.94	0.36	0.21	0.67
	MLspike		0.50	0.62	0.41	0.08	0.58
	ML+GD		0.36	0.71	0.64	0.24	0.67

- Comparable to generative approaches, inferior to data driven method
- Dataset-wise thresholding improves the performance
- Post-processing improves the MLspike results

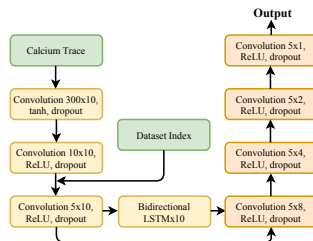
⁷ Deneux et. al., Nature 2016, Vogelstein et. al, 2010, Theis et. al, *Neuron* 2016

Challenge

- Performance of the best algorithms were still far from ground truth
- Challenge set a standardized dataset and evaluation framework
- Resulted in 26% relative improvement of primary evaluation measure

CNN+LSTM

- State-of-the-art data driven method in spikefinder contest
- Estimate the frame-wise spike estimate from a contextual window of input
- CNN receives Ca^{2+} signal and dataset index and 3 sec context is used to get one prediction
- Learning objective: maximize correlation coefficient

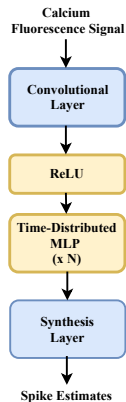


⁹Community-based benchmarking improves spike rate inference from two-photon calcium imaging data, Berens et. al, PLOS Computational Biology, 2018

Motivation

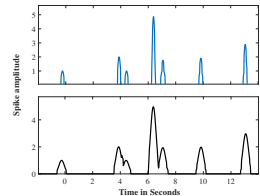
- Success of sequence-to-sequence models in speech processing tasks
- Attempts to end-to-end source separation

- S2S receives one recording of raw-fluorescence signal and outputs the entire spike information signal
- The layers operate on one second of input data, keeping the temporal information
- Automatic short-time processing
- Synthesis layer to reconstruct the spike information

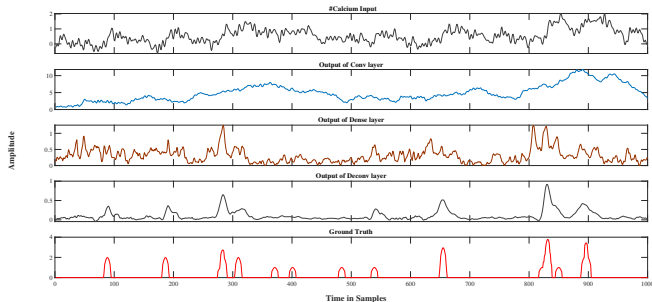


Learning Targets

- Sparse discrete signal: S2S is faster than the baseline
- Improve efficiency by convolving with a Gaussian window
- Maximizes the correlation coefficient between the target and estimated signal



Layer-wise output



- **Spikefinder Challenge Dataset**

- Various brain regions, scanning rates, scanning methods and indicators
- 10 training datasets (5 publicly available) and 5 test datasets

- **Evaluation measure:** Correlation coefficient, AUC and rank (40 ms bin width)

- **Baseline methods:** State-of-the art supervised method based on *spikefinder challenge*⁹

Configuration of S2S

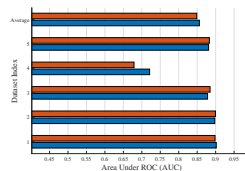
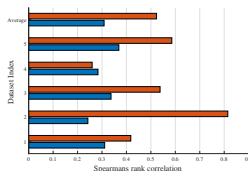
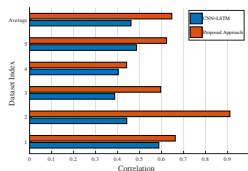
- Convolution and synthesis filters (N_{filt}): 30
- Filter width (W_{len}) and shift (W_{shift}): 100, 1
- Number of dense layers: 0 to 3
- Number of parameters: 8,790 (for 3 dense layers)
- Train-validation split is 0.7-0.3
- Adam Optimizer. Training stops when validation error reaches 1e-6 or early stopping with a patience factor of 6

⁹Berens et. al, PLOS Computational Biology, 2018

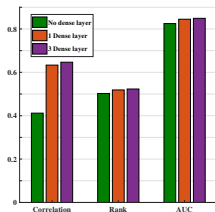
- Achieves state-of-the-art results in *spikefinder* contest:

Team Name	Train correlation	Test correlation	Δ correlation	Rank	AUC
Team 1 MLspike new	0.4823	0.4382	0.0810	0.2878	0.846
Team 2 conv6	0.4727	0.4378	0.0806	0.3319	0.846
Team 3 DeepSpike	0.4730	0.4347	0.0775	0.3338	0.851
Team 4 Purgatorio	0.5370	0.4325	0.0753	0.3258	0.815
Team 5 Embedding of CNNs	0.4900	0.4291	0.0719	0.2822	0.821
Team 6 Suite2p	0.4752	0.4188	0.0617	0.3071	0.821
Baseline STM	0.4024	0.3572	-	0.2664	0.821
Proposed S2S	0.6325	0.6404	0.0079	0.5208	0.847

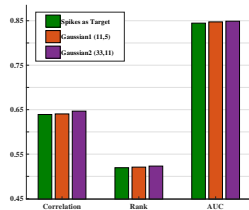
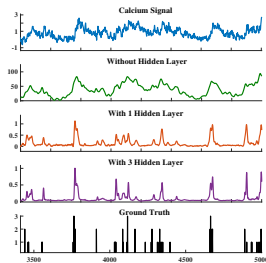
- Comparisons with best supervised model (CNN+LSTM)



Comparison of different configurations



Hidden layers



Training targets

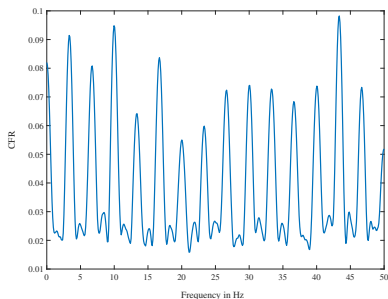
Table 1: Performance on unseen data

Configuration	Training data	Correlation	AUC	rank
1 Dense layer	GCaMP	0.608	0.510	0.832
1 Dense layer	GCaMP + OGB	0.639	0.519	0.845
3 Dense layers	GCaMP	0.622	0.518	0.843
3 Dense layers	GCaMP +OGB	0.640	0.521	0.847

- 3 Dense layer with a Gaussian train target provides the best performance.
- S2S is a computationally efficient and reliable model
- Provides progress in spike inference field: An improvement of 46% in primary evaluation measure.

Observations

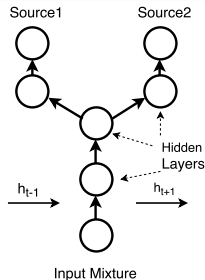
- Convolution layer filters have response over the entire frequency range
- Filters in the synthesis layer have interesting Cumulative Frequency Response (CFR)
- Conv. layer smoothens and denoises the input, dense layers emphasize the spike positions, synthesize layer further denoises the output



CFR of synthesis filters

Musical Source Separation

- In a mixture, Time-Frequency (T-F) bins are more resolved in GD than in magnitude spectrum
- Uses MOD-GD feature instead of magnitude spectrum (state-of-the-art)
- Recurrent Neural Network (RNN) learns the T-F mask¹⁰
- **Tasks:** Singing voice separation, vocal-violin separation



- Joint learning of the sources
- Discriminative training
- Output activation also depends on the past output

Objective function:

$$\|\hat{y}_{1t} - y_{1t}\|_2^2 + \|\hat{y}_{2t} - y_{2t}\|_2^2 - \gamma(\|\hat{y}_{1t} - y_{2t}\|_2^2 + \|\hat{y}_{2t} - y_{1t}\|_2^2)$$

¹⁰ Joint optimization of masks and deep recurrent neural networks for monaural source separation, Po Sen Huang et. al, IEEE transactions on Audio, Speech and Language Processing, 2015

Proposed: MODGD Feature for Musical Source Separation

- Learning source specific masks:

$$\text{Mixture} \otimes \text{Mask}_1 \equiv \text{Source}_1$$

- Source_1 (magnitude spectrum) is then combined with the mixture phase.

MODGDgram to learn T-F mask

$$m_1(t, f) = \begin{cases} 1, & s_1(t, f) - s_2(t, f) \geq \theta \\ 0, & s_1(t, f) - s_2(t, f) < \theta \end{cases}$$

- Task1: Singing Voice Separation

	Train	Dev	Test	Total
Data	171	4	825	1000
Artists	2			17

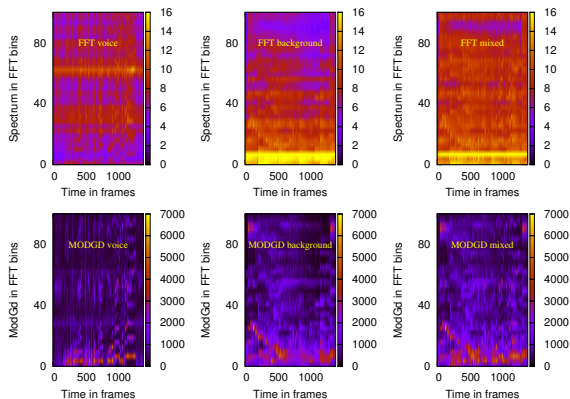
Table 2: MIR-1K Dataset.

- Task2: Vocal-Violin Separation

	Train	Dev	Test	Total
Clips	54	3	20	77

Table 3: Carnatic Music Dataset.

Feature Comparison



- BSS Evaluation metrics: Signal to Distortion (SDR), Interference (SIR) and Artifacts ratios (SAR)¹¹
- Better discrimination is observed in the MODGD domain.

¹¹Vincent et. al, "Oracle estimators for the bench-marking of source separation algorithms." Signal Processing, 2006

- **Singing Voice Separation**

Comparison with the state-of-the-art

Table 4: Results with 2-DRNN

Feature	Hidden units per layer	GNSDR	GSIR	GSAR
ModGD	500	7.15	13.46	9.11
Spectrum	500	5.74	12.15	7.62
ModGD	1000	7.50	13.73	9.45
Spectrum	1000	7.45	13.08	9.68

Table 5: Results with DRNN architectures

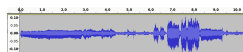
Architecture	Feature	GNSDR	GSIR	GSAR
1-DRNN	Spectrum	7.21	12.76	9.56
	ModGD	7.26	12.93	9.42
2-DRNN	Spectrum	7.45	13.08	9.68
	ModGD	7.50	13.73	9.45
3-DRNN	Spectrum	7.09	11.69	10.00
	ModGD	6.92	12.27	9.26
stacked DRNN	Spectrum	7.15	12.79	9.39
	ModGD	7.31	13.45	9.30

- MODGD with 500 nodes/layer \approx Mag. spec. with 1000 nodes/layer
- Relative improvement of **4.9 dB** for SIR over Mag. spectrum
- Better SIR is achieved for *all* the configurations
- **Vocal-Violin Separation:** Similar improvement

Feature	GNSDR	GSIR	GSAR
ModGD	9.42	13.72	11.76
Spectrum	9.38	13.55	11.80

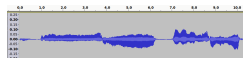
Table 6: 1-DRNN performance in the Carnatic music data set.

Demo: Vocal-Violin Separation



Vocal

Play



Violin

Play



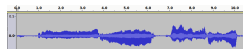
Mixture

Play



Separated Vocal

Play

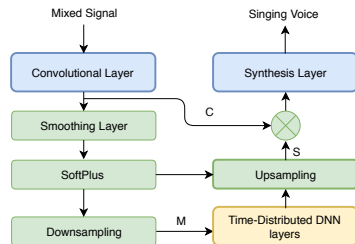


Separated Violin

Play

- **Mask-based approaches:** Hand-crafted features and mixture phase for source reconstruction
- Motivated by end-to-end approaches to speech and speaker recognition systems
- S2S proposed for spike estimation can be used as a general framework

- Encoder-Decoder framework
- Time-Distributed DNN Layers
- Use of signal-based loss functions



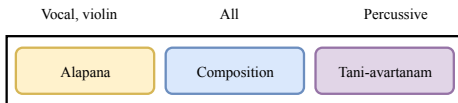
- Singing voice separation: MIR-1K Dataset.
- Convolutional filter width: 100 ms, hop size:10 ms
- S2S-2 forward transform is made as the inverse of reverse transform

Table 7: Performance measures with proposed approach.

System	Dev Set			Test Set		
	GNSDR	GSIR	GSAR	GNSDR	GSIR	GSAR
Baseline: DNN-1	5.60	21.40	6.20	5.00	15.87	5.99
Baseline: DNN-2	-	-	-	7.45	13.08	9.68
Proposed: S2S-1	9.20	24.98	9.57	7.65	22.21	8.18
Proposed: S2S-2	9.62	24.75	9.97	7.72	22.71	8.22

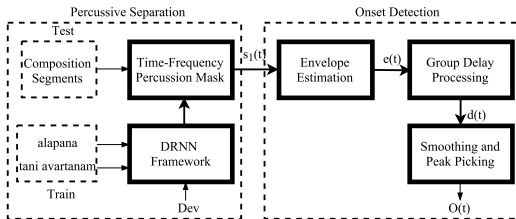
- S2S has 2.65 dB absolute improvement over traditional source separation systems
- Signal based similarity measure and learned bases: Improved SIR

1. Percussive Onset detection on composition items of Carnatic Music



Two-Stage Approach

- Stage 1: Percussive separation using DRNN
- Stage 2: Solo percussive onset detection



Block diagram of the hybrid approach

- Existing onset detector algorithms uses solo track
- Separation Baselines: Harmonic-Percussive Separation (HPS) algorithm¹²
- Onset Baselines: Directly on the mixture, on the separated track, CNN on the mixture & on the separated track, and Oracle performance

BSS evaluation metrics

Concert	DRNN			HPS		
	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

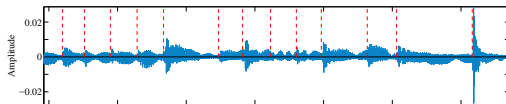
Comparison of F-measures

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

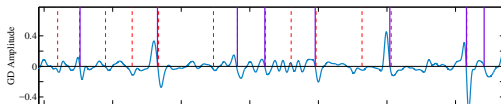
- The degradation with respect to the solo source is 4.6%, the improvement compared to the direct onset detection is 18.4%.
- Performance is improved for *all* the datasets upon separation
- The average F-measure is 11.8% higher than CNN baseline¹³

¹²Harmonic/Percussive Separation Using Median Filtering, Fitzgerald, DAFx 2010

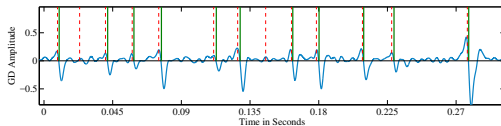
¹³Improved musical onset detection with Convolutional Neural Networks., Sebastian Bock, ICASSP 2014



(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

An example excerpt from SS dataset

Demo

Composition mixture

Play

Separated percussion

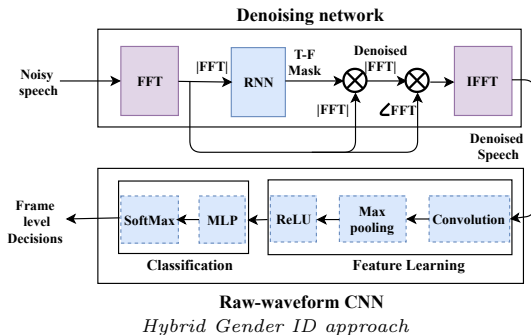
Play

Ground truth percussion

Play

Introduction

- Gender ID is limited by type & degree of noise and limited data
- **State-of-the-art:** i-vector based SVM¹⁴ has Unweighted Average Recall (UAR) around 75%
- Proposed method uses a two-stage pipeline: Denoising and gender ID
- Denoiser: Inspired from RNN-based source separation



¹⁴Gender identification using mfcc for telephone applicationsa comparative study, Ahmad Jamil, 2016

- Datasets:
 - Conversations in callfriend corpus
 - DEMAND noise corpus
- Denoiser:
 - 17% of the dataset is used for training
 - Training with 0 dB mixture, test with 0 and -5 dB
- Raw-waveform CNN: 30% of the dataset is used for training
- Baseline system: SVM trained with i-vectors
 - UBM-GMM: 2048 mixtures, 400 dimensional i-vector extractor, trained using AMI meeting corpus (down-sampled to 8 kHz)
 - SVM classifier with Radial Basis Function kernel

Raw-waveform CNN

- End-to-end learning with categorical cross entropy (CE)
- Has better performance than feature-based classifiers ¹⁵

Two raw-waveform architectures

Parameters	CNN1	CNN2
number of conv. layers	3	2
L1 width/shift (in samples)/# filters	30/10/80	150/10/80
L2 width/shift (in frames)/# filters	7/1/60	7/1/60
Max pooling size/shift	3/1	3/1
number of hidden units	1024	100
Total number of parameters	433,114	184,042

¹⁵ On learning to identify genders from the raw-speech, Selen et al, Interspeech 2018

Denoiser performance

Measure	Binary Mask		Soft Mask	
	-5 dB	0 dB	-5 dB	0 dB
GNSDR	18.09	11.12	18.24	19.50
GSIR	23.57	20.30	20.03	19.50
GSAR	14.28	13.05	14.66	14.15

Gender ID performance (UAR%)

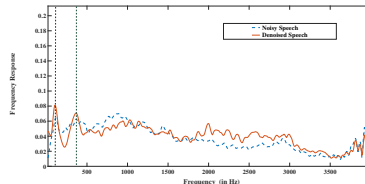
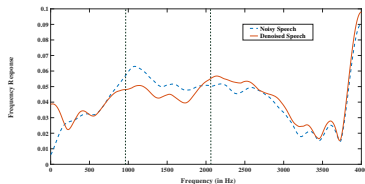
System	Noisy		Denoised	
	-5 dB	0 dB	-5 dB	0 dB
Baseline	76.84	81.95	79.83	83.34
CNN2	83.86	89.13	88.00	91.53
CNN1	87.47	91.31	90.17	93.01


- Absolute improvement of 11.06% and 13.33% over the baseline system for 0 dB and -5 dB SNR conditions
- Denoiser performs equally well for unseen noise and language conditions
- Demo:
 - Noisy conversation [Play](#)
 - Noise in the conversation [Play](#)
 - Enhanced Speech [Play](#)

Cumulative Frequency response

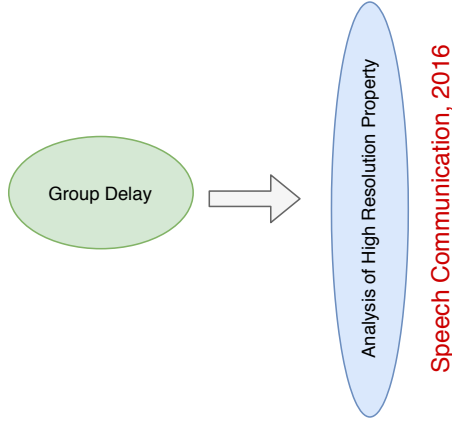
- Convolution layer filters learn gender-specific responses
- Filters learned for noisy speech is similar to the cleaned ones
- Capturing vocal tract information in CNN1 and fundamental frequency in CNN2

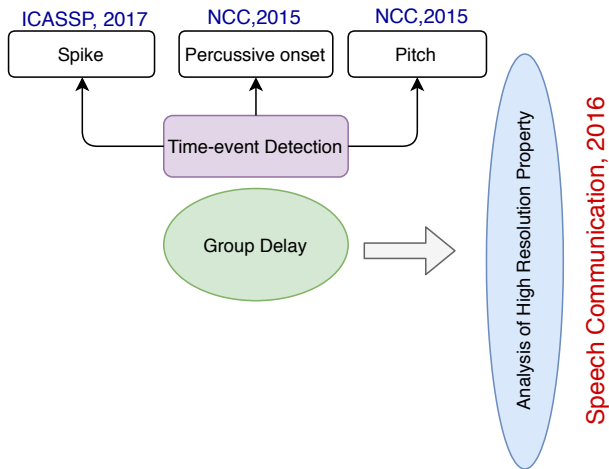
Cumulative Frequency Response of Layer1 filters in CNN1 (left) and CNN2 (right).

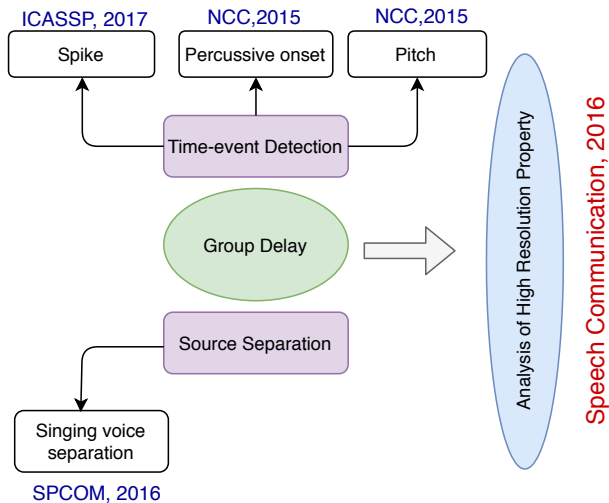




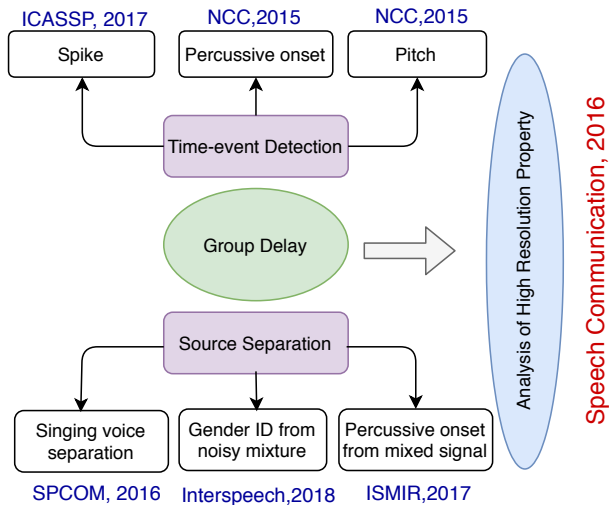
Group Delay



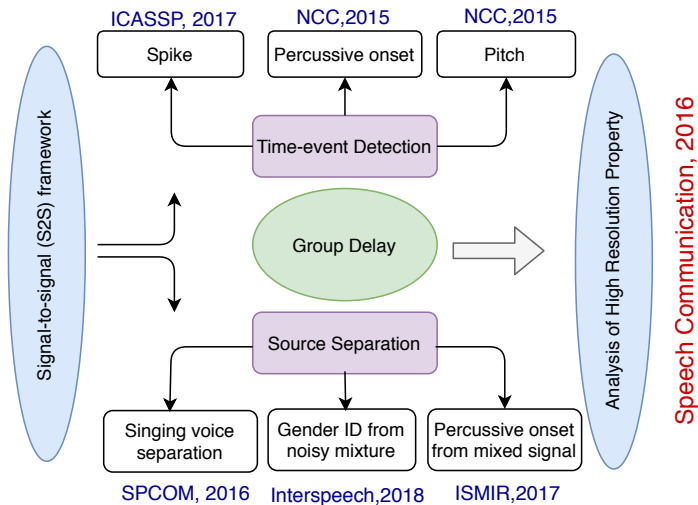




Summary of the proposed works



Summary of the proposed works



Many Thanks to

- My Guide and Mentor: Prof. Hema A. Murthy
- Doctoral Committee and the examiners
- CompMusic Project and Computational Brain Research Project: Prof. Mriganka Sur and Sur Lab
- Prof. Shirkanth Narayanan and Mr. Manoj Kumar from SAIL lab
- Swiss Government Excellence Scholarship Project: Dr. Mathew MagiMai.-Doss and speech group at Idiap
- All members of SMT and speech Labs

Journal Article(s):

- ④ Jilt Sebastian, Mari Ganesh Kumar, Mriganka Sur and Hema A. Murthy "A High Resolution-based Algorithm for Spike Estimation from Fluorescence Signals", IEEE transactions on signal processing. Volume 67, Issue 11 (2019), DOI:10.1109/TSP.2019.2908913.
- ② Jilt Sebastian, Manoj Kumar P. A. and Hema A. Murthy, "Analysis of High Resolution Property of Group Delay function with Applications to Audio Signal Processing", Journal of Speech Communication, Elsevier, pages: 42-53 (2016), URL: <https://doi.org/10.1016/j.specom.2015.12.008>.
- ③ **In preparation:** Jilt Sebastian, Mathew-Magimai.-Doss, Mriganka Sur and Hema A. Murthy "Signal-to-signal networks for improved spike estimation from calcium imaging data", PLOS Computational Biology.

Conferences Paper(s):

- ④ (Not related to the thesis) Jilt Sebastian and Piero Pierucci, "Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts", accepted for publication at Interspeech 2019, Graz, Austria.
- ② Jilt Sebastian, Manoj Kumar P. A., D. S. Pavankumar, Mathew Magimai.-Doss, Hema A. Murthy and Shrikanth Narayanan, "Denoising and Raw-waveform Networks for Weakly-Supervised Gender Identification on Noisy Speech", Interspeech 2018, Hyderabad, India, URL: <http://dx.doi.org/10.21437/Interspeech.2018-2321>.
- ③ (Not related to the thesis) Bogdan Vlasenko, Jilt Sebastian, D. S. Pavankumar and Mathew Magimai.-Doss, "Implementing Fusion Techniques for the Classification of Paralinguistic Information", Interspeech 2018, Hyderabad, India, URL: <http://dx.doi.org/10.21437/Interspeech.2018-2360>.

- ④ Jilt Sebastian and Hema A. Murthy, "*Onset Detection in Composition Items of Carnatic Music*", ISMIR 2017, Suzhou, China, URL: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/91_Paper.pdf.
- ⑤ Jilt Sebastian, Mari Ganesh Kumar M, Y. S. Sreekar, Rajeev Vijay Rikhye, Mriganka Sur and Hema A. Murthy, "*GDspike: An Accurate Spike estimation Algorithm from Noisy Calcium Flourescence Signals*", ICASSP 2017, New Orleans, United States, URL: 10.1109/ICASSP.2017.7952315.
- ⑥ Jilt Sebastian and Hema A. Murthy, "*Group Delay Based Music Source Separation Using Deep Recurrent Neural Networks*", SPCOM 2016, Bangalore, India, URL: 10.1109/SPCOM.2016.7746672.
- ④ Jilt Sebastian, Manoj Kumar P. A. and Hema A. Murthy, "*Pitch Estimation From Speech Using Grating Compression Transform on Modified Group-Delay-gram*" , NCC 2015, Mumbai, India, URL: 10.1109/NCC.2015.7084899.
- ⑥ Manoj Kumar P. A., Jilt Sebastian and Hema A. Murthy, "*Musical Onset Detection on Carnatic Percussion Instruments*", NCC 2015, Mumbai, India, URL: 10.1109/NCC.2015.7084897.

Thank You!

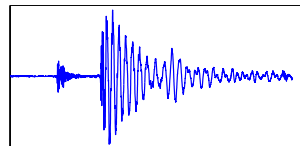
APPENDIX

Feature Extraction

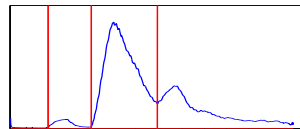
- Pre-processing: source separation
- Remove pitch information
- Transform raw audio into *detection function*

Peak picking

- Zero crossings, peak/valley detection
- Post processing : smoothening



Time



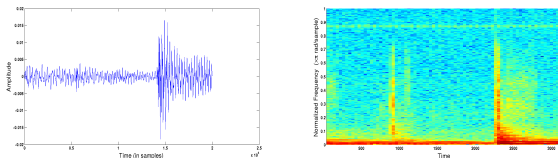
Time

(top) Audio with non-pitched and pitched strokes. (bottom) Possible detection function with onsets marked

Percussive Onset Detection on Carnatic music

- Focuses on percussion in Carnatic music: 5 instruments
- Challenge: Variations in tempo, loudness and spectral characteristics
- Feature Extraction: Envelope estimation (AM-FM demodulation) + Minimum phase group delay, Peak picking: Hard-threshold

Onset Detection Algorithm



Waveform (left) and Spectrogram (right) of a mridangam clip containing a silent stroke

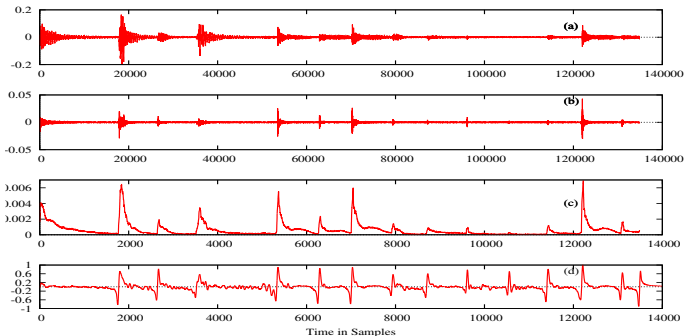


Figure 3: Working of proposed algorithm. (a) Music signal, (b) derivative of music signal, (c) envelope estimated using Hilbert transform, and (d) minimum phase group delay computed on the envelope.

Table 8: Instrument wise details of datasets.

Instrument	Total length (min:sec)	Strokes
Mridangam	18:41	5982
Ghatam	4:14	2616
Kanjira	3:11	1377
Morsing	6:35	2184
Thavil	4:39	2904
Ensemble	5:00	2529

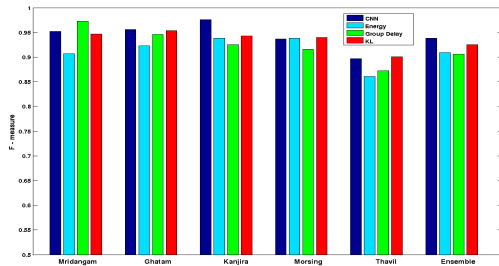


Figure 4: Convolutional Neural Network (state-of-art), Energy, Group delay and Kulback - Leibler based onset detection algorithms

- Both CNN and GD algorithms report fairly good F-measures on all instruments
- Proposed algorithm stands out in performance for the mridangam dataset

- The average correlation between discrete spike train and ground truth is 0.349 for MLspike and 0.262 for GDspike
- GDspike is 13 times faster than MLspike

Performance of GDSpike for evaluation data sets.

Data set	Recall	Precision	F-measure	Corr.	AUC
GCaMP6s	0.58	0.72	0.58	0.177	0.74
GCaMP6f	0.50	0.73	0.52	0.168	0.78
GCaMP5k	0.71	0.36	0.38	0.331	0.77
jRGECO1a	0.42	0.40	0.32	0.08	0.62
jRCaMP1a	0.26	0.63	0.29	0.214	0.57
OGB (Weizmann)	0.14	0.95	0.23	0.208	0.59
OGB (Marselie)	0.22	0.85	0.31	0.147	0.64
OGB (invitro)	0.20	0.94	0.36	0.208	0.67
GCaMP6s (Budapest)	0.69	0.60	0.56	0.394	0.76

F-measure with the dataset-wise thresholds

Dataset	Global	Dataset-wise 1	Dataset-wise 2
GCaMP6s	0.58	0.60	0.53
GCaMP6f	0.52	0.52	0.56
GCaMP5k	0.38	0.41	0.39
jRGECO1a	0.32	0.25	0.39
jRCaMP1a	0.29	0.28	0.38
OGB (Weizmann)	0.23	0.47	0.39
OGB (Marselie)	0.31	0.34	0.47
OGB (invitro)	0.36	0.59	0.51
GCaMP6s (Budapest)	0.56	0.65	0.47
Average	0.39	0.46	0.45

- Global: Mean + Standard Deviation averaged across all the datasets
- dataset-wise 1: 3-fold cross validation on 60% of each data set to determine X (Mean + X \times Standard Deviation), testing on held out data (40%)
- dataset-wise 2: Using 20% of the data to determine X (with maximum F-measure) and testing on 80% of the data