

# Score-informed Syllable Segmentation for Jingju A Cappella Singing Voice with Mel-frequency Intensity Profiles

Rong Gong<sup>‡</sup>, Nicolas Obin<sup>♦</sup>,  
Georgi Dzhambazov<sup>‡</sup> and Xavier Serra<sup>‡</sup>

<sup>‡</sup> Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

<sup>♦</sup> IRCAM, CNRS, UPMC-Sorbonne Université, Paris, France



# Structure

- Background – jingju singing evaluation

- Syllable segmentation

- Intensity-based onset detection function



- Score-informed on sequence decoding

- [update] Convolutional neural networks (CNNs) based onset detection function

# Background

## Jingju singing



- Jingju (Beijing opera): a traditional Chinese art form
- Characters are played by specific **role-types** – **patterns of performance**.
- **Four** main role-types: *sheng*, *dan*, *jing*, *chou*
- Most important role-types in terms of singing: ***laosheng* (male) and *dan* (female)**

# Background

## Learning by imitation

Tutor's demonstrative singing



Trainee's tentative singing



Tutor's **evaluation, verbal feedback or demonstrative singing**



Trainee's revised singing



A laosheng aria class in The National Academy of Chinese Theatre Arts

Minimum evaluation temporal units:

**syllable** and phoneme:

"**syllable** attack should be more stressed" or  
"**syllable** tail needs to be pronounced more clearly"

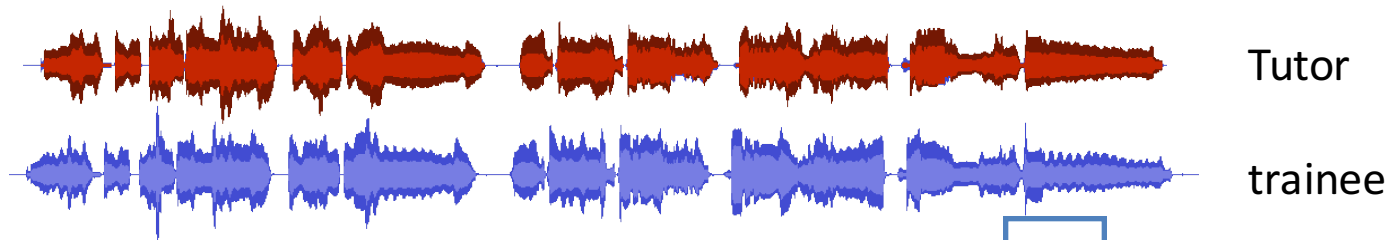
... ..



# Background

## Automatic jingju singing evaluation

Sources: singing phrase audio recordings



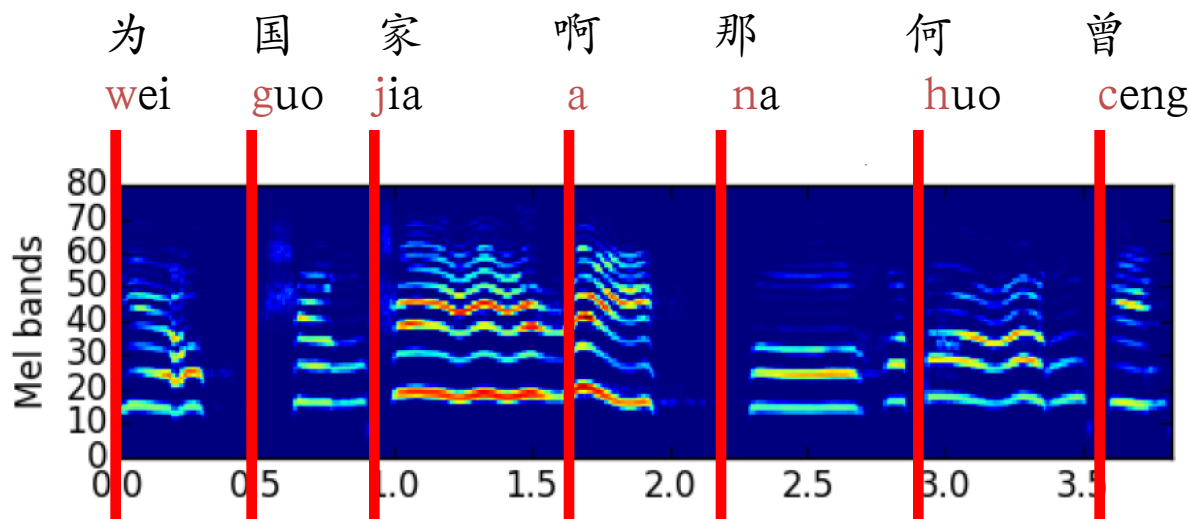
Syllable segmentation:  
Capture the temporal details

Measure similarity  
Between tutor's and trainee's singings

# Syllable segmentation

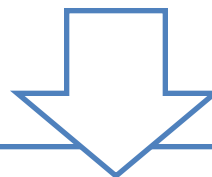
## Defining the problem

- Jingju singing
  - Chinese language
  - Mono-syllabic, written character
- syllables can start by
  - Consonants – **w**ei **g**uo **j**ia **n**a **h**uo **c**eng
  - Vowel - **a**
- Segmentation
  - Determining the syllable **boundaries** (**onset** and **offset**)

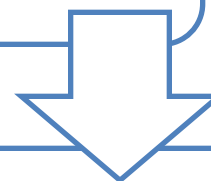


# Method

Source: singing phrase audio recording



Syllable onset detection function



Score-informed  
syllable onset sequence decoding

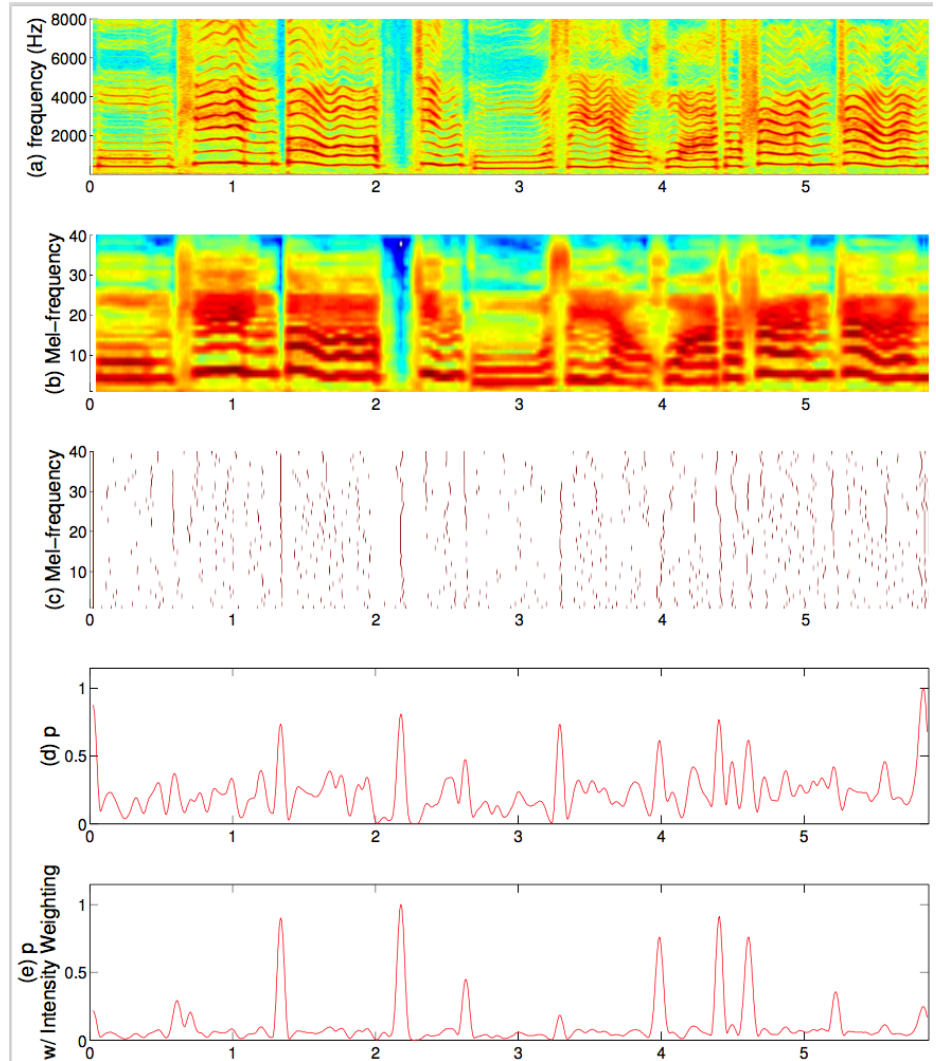
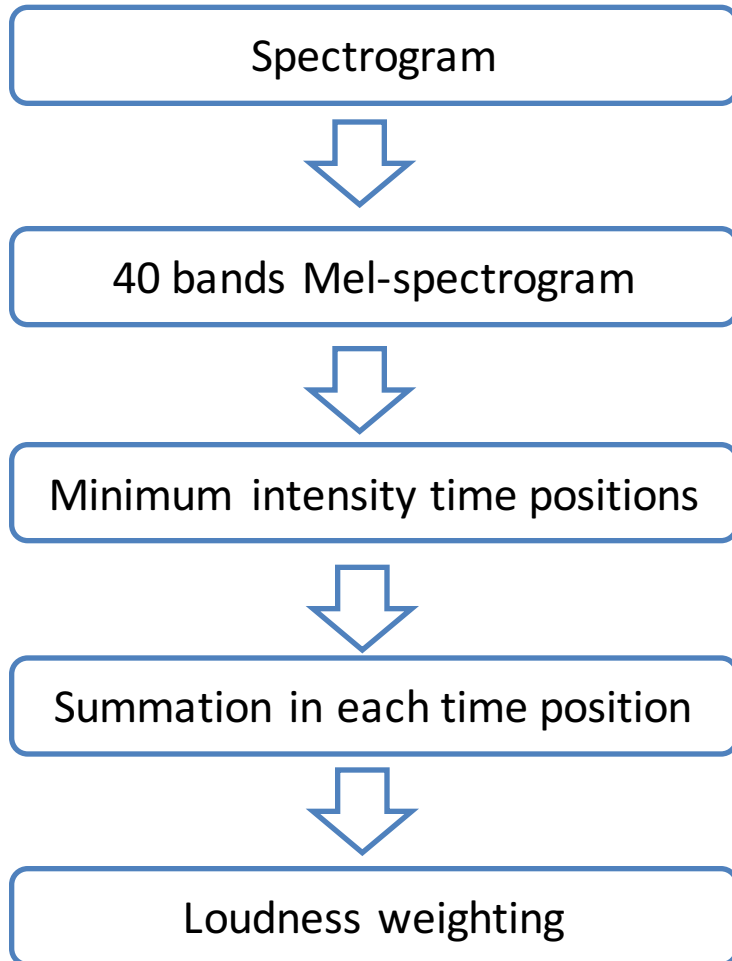
# Method

## Syllable onset detection function

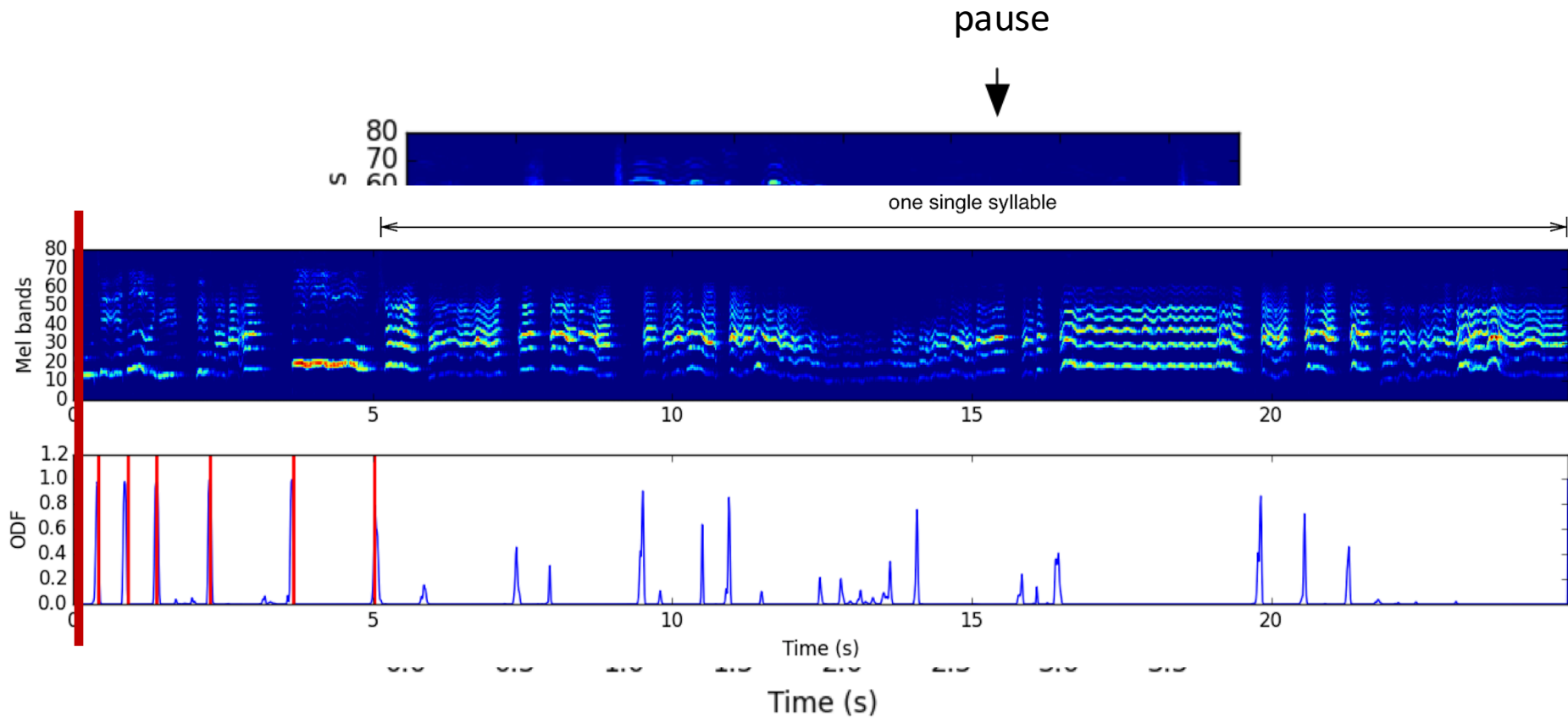
- Hand-crafted
- Intensity/energy-based
- Assumption: a syllable starts more likely by **consonants**, where the energy/intensity should be **low**.

# Method

## Syllable onset detection function



# Method Challenge

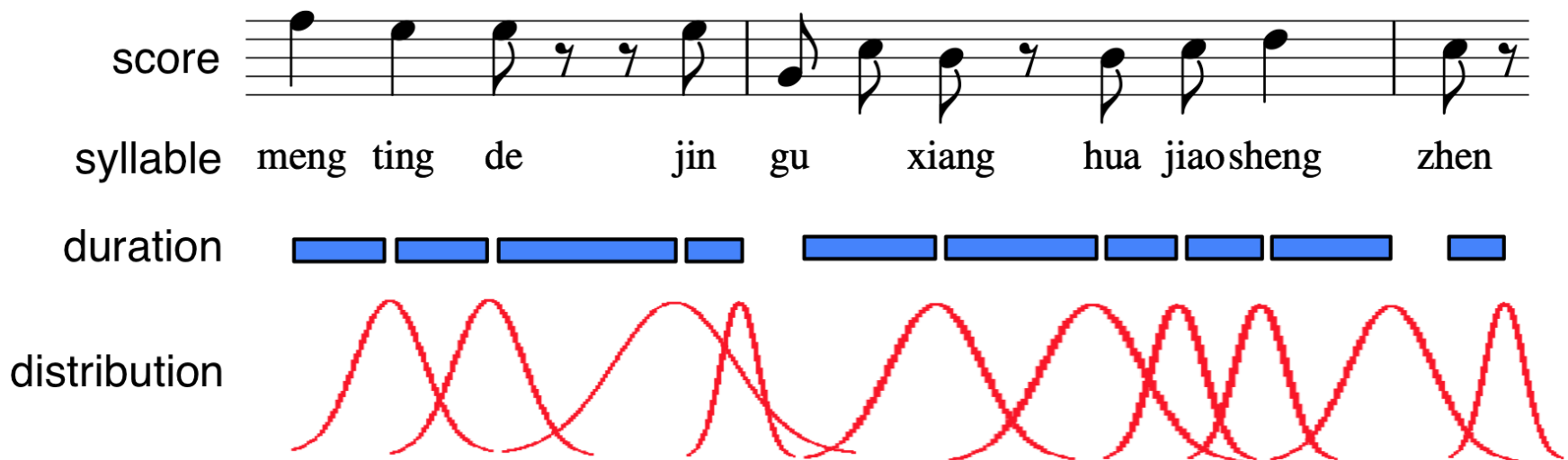




# Method

## Solution – Score-informed onset sequence decoding

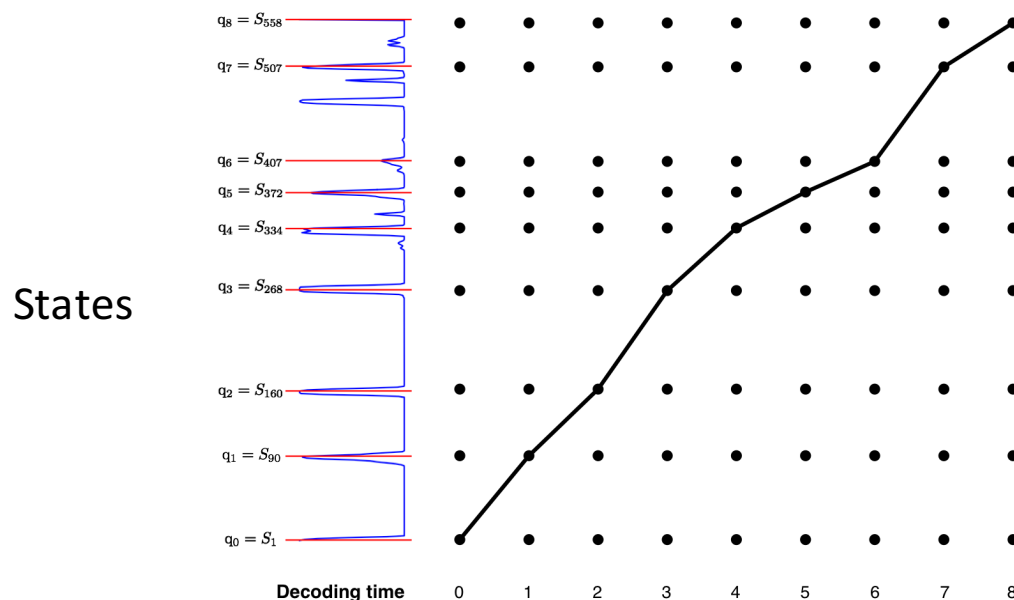
- **Score:** *a priori* information of the syllable duration.
- **Duration distribution:** Gaussian, highest probability on its mean.



# Method

## Solution – Score-informed onset sequence decoding

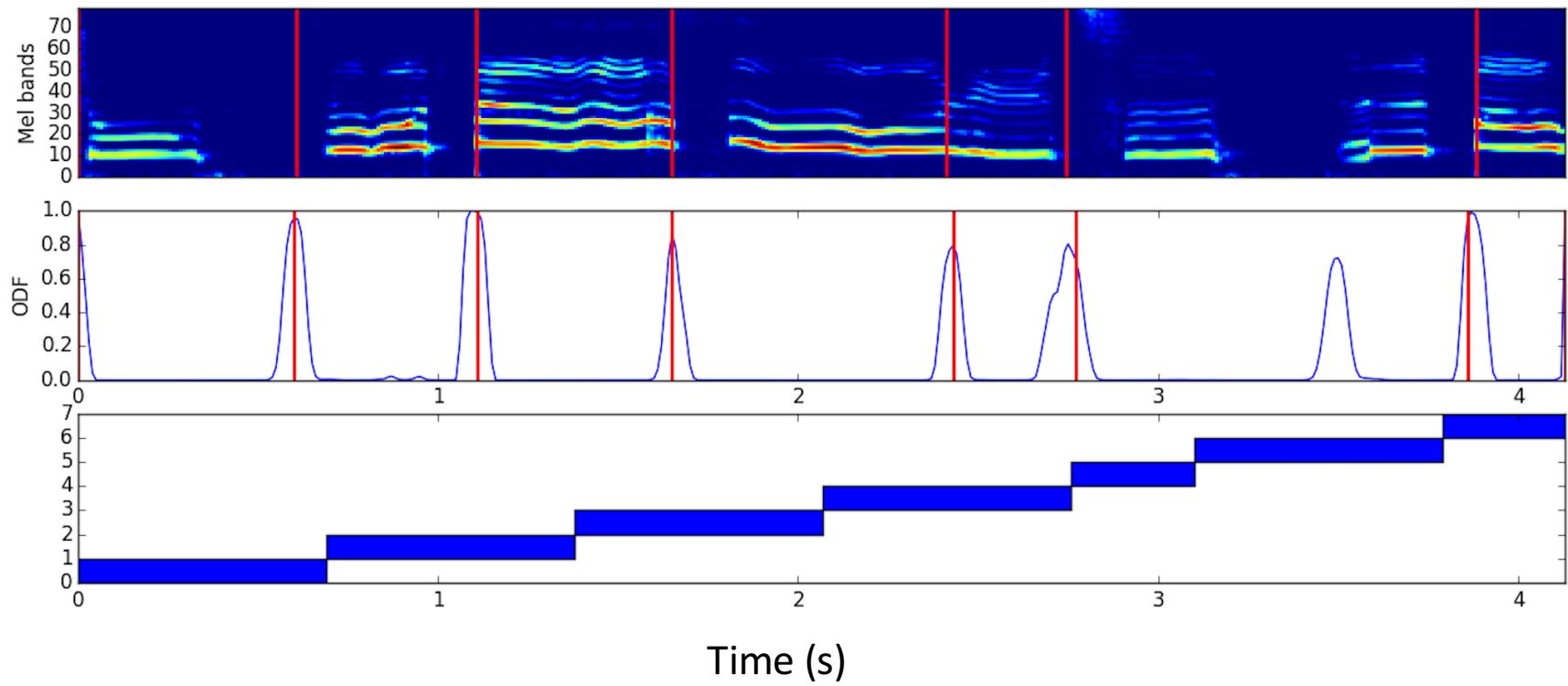
- **Topology:** left-to-right HMM
  - **States:** all timestamps
  - **Observation probabilities:** onset detection function
  - **Transition probabilities:** *a priori* syllable duration distributions



Viterbi algorithm ->  
Most-likely onset sequence

Decoding time = syllable number in score

# Results

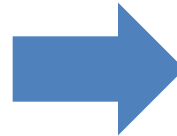


# Update to this paper

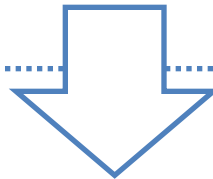
Source: singing phrase audio recording



Hand crafted  
onset detection function  
**Not robust**



**Convolutional neural  
networks** onset detection  
function



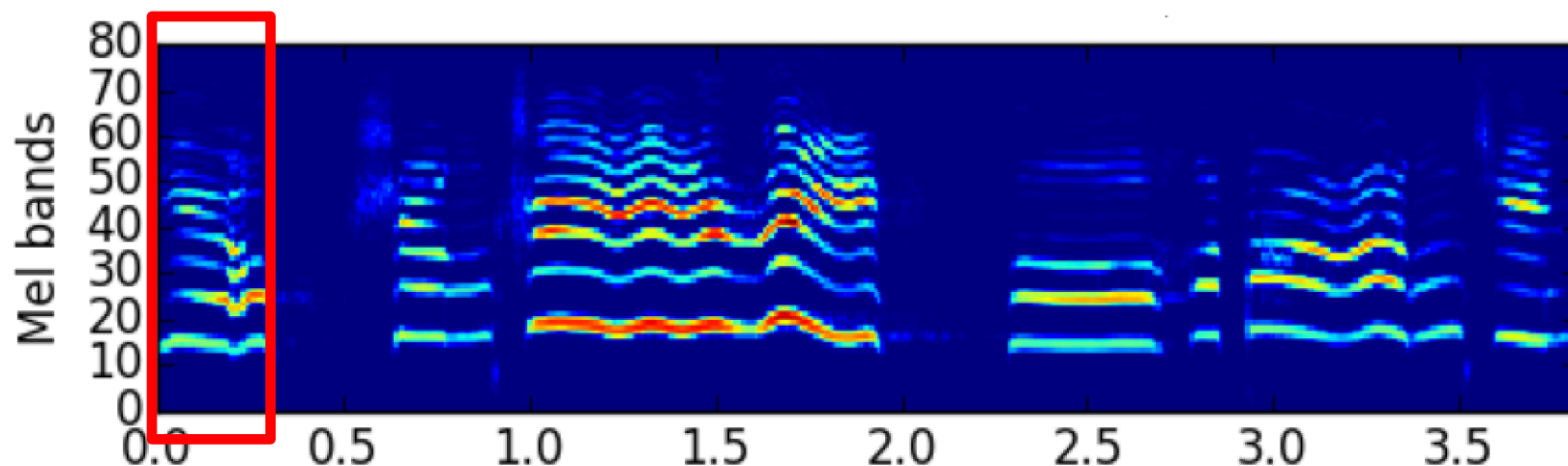
Score-informed  
syllable onset sequence decoding

This update has been written in another paper and submitted

# Method (update)

## CNNs - Input

shift 1 frame to right



Input: Context window = 21 frames = 0.21 s

Output: **onset probability** on the center of the input

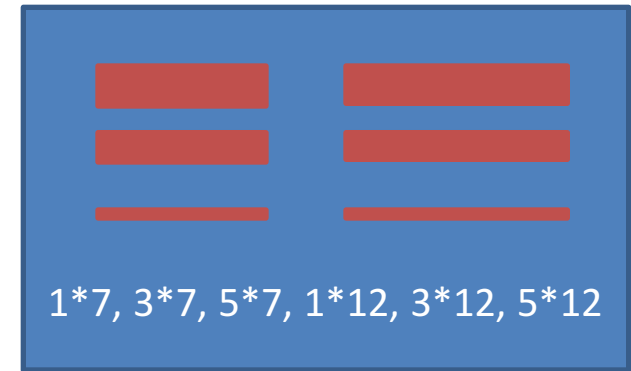
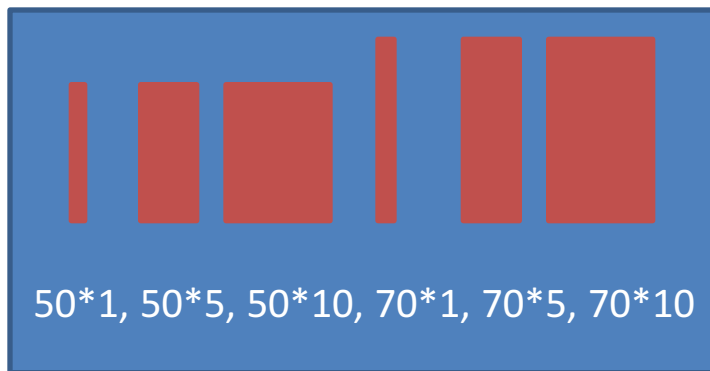
# Method (update)

## CNNs – Timbral/Temporal

Input

Mel spectrograms  
80\*21 (Mel bands \* window size)

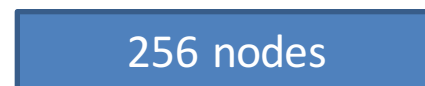
Conv layer 1  
6 timbral/  
temporal filters



Conv layer 2  
1 square filter



Fully-connected



2 nodes softmax

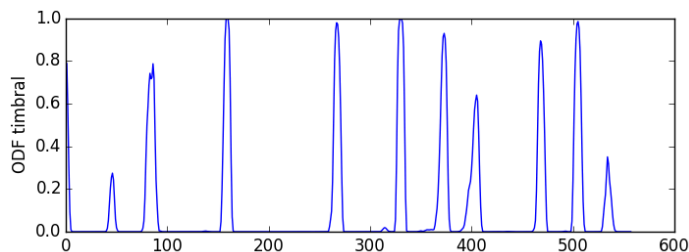


onset non-onset



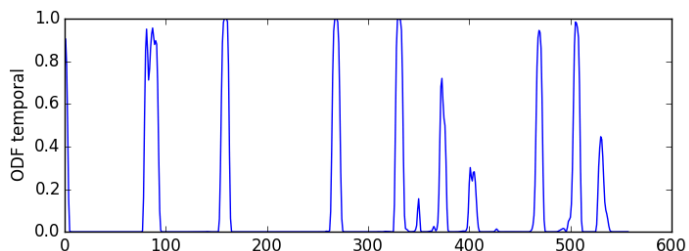
# Method (update)

## Fusion of onset detection function



Timbral CNNs ODF

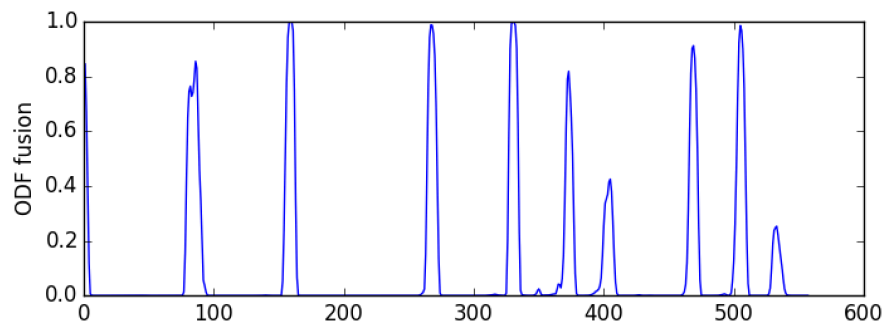
$$\text{ODF\_timbral}^{(0.5)} * \text{ODF\_temporal}^{(0.5)}$$



Temporal CNNs ODF

//

Fused ODF



# Evaluation

## Dataset

- Two role-types: *dan* and *laosheng*
- 39 arias, 291 phrases, 2641 syllables
- **Audio** from Queen Mary and MTG dataset
- Onset **annotation** by MTG researchers

## Metrics

- Onset and offset tolerance: 50ms

# Evaluation Results

Onset detection functions	F-measure (%)
Fusion CNNs	<b>86.37</b>
Timbral CNNs	84.83
Temporal CNNs	83.28
Hand crafted (FMA paper)	53.86

Onset and offset tolerance: 50ms

# Future work

- Incoherence between scores and recordings
  - Syllable insertion/deletions
  - Prolongation of the last syllable in a phrase
  - Incorporating domain knowledge

# Conclusion

- Background – jingju singing evaluation
- Syllable segmentation
  - Intensity-based onset detection function
  - Score-informed onset sequence decoding
  - CNNs based onset detection function

# Reproducibility

- Code
  - <https://github.com/ronggong/jingjuSyllabicSegmentaion>
- Dataset
  - <http://doi.org/10.5281/zenodo.345490>

• Thank you!