Harmonic Analysis With Neural Semi-CRF

Qiaoyu Yang, Frank Cwitkowitz, and Zhiyao Duan

Audio Information Research Lab, University of Rochester

Introduction

- ❑ In music, harmonic analysis is the process of finding the underlying relationship among the notes and joining them together.

- ❑ The process is two-fold: recognizing harmony labels and finding their time boundaries. Most previous works only focused on the first component, which may lead to segmentation errors.

- ❑ In this paper, we introduce a novel approach to jointly detect the labels and time boundaries with neural semi-CRF (Conditional Random Field).

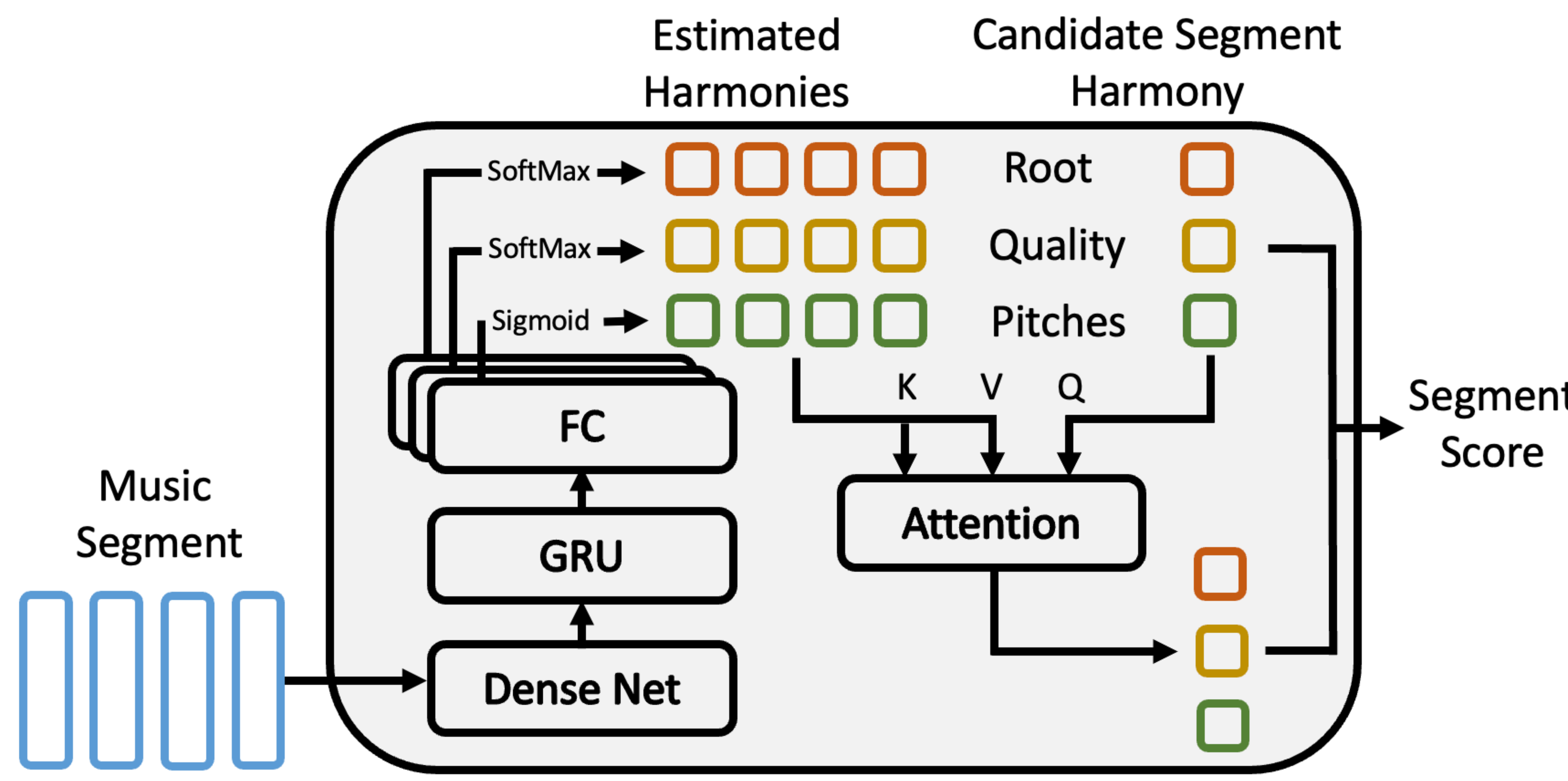
Data

- ❑ **Dataset**: Four classical music datasets of different styles are included in the experiments. In total, there are 321 pieces and 44K harmony labels

- ❑ **Input Representations**: MIDI-like symbolic music are sliced into fixed-length frames (eighth note). The pitch information in each frame is processed and transformed to a 24-dimensional vector that encodes the pitch-class activations as well as the bass note.

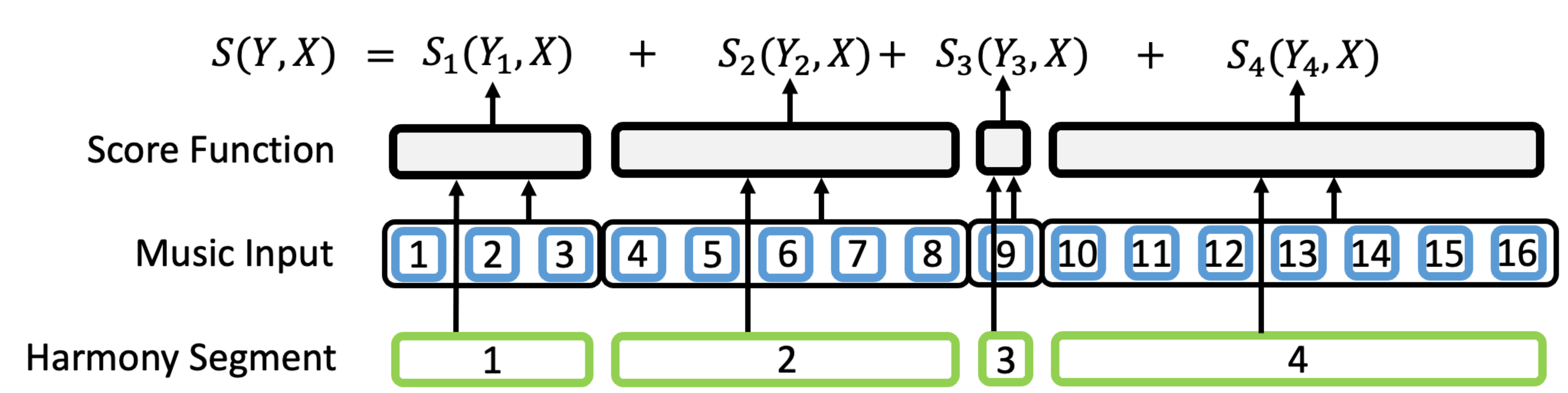
Method

Neural Score Function



1. **Frame-level estimation:** CRNN-based encoder to compute the probability distribution of harmony types at the frame-level.
2. **Attention-based score function:** For each candidate harmony region, attention is used to aggregate multiple frame-level estimations to a single distribution of harmony types.
3. **Absence Score:** An additional score component is used to explicitly penalize incomplete chord profiles and extra chordal notes in the estimations.

Neural Semi-CRF



Score function: $S(X,\ Y) = \sum_{i=1}^{m} S_i(X, Y_i)$

Training Objective: Given X and Y, find the model parameters θ that minimizes $\mathrm{L}(\theta) = -\log\left(P_\theta(Y|X)\right)$

Inference: Given X, find Y that maximizes $P(Y|X)$

Results

- ❑ Our proposed model outperforms previous methods that lack time awareness or neural feature extraction, in both accuracy of harmony labels and the quality of harmony regions

Model	Root	Quality	Majmin	Overall
CRNN	0.735	0.714	0.865	0.634
frog	0.733	0.542	0.815	0.459
RuleSCRF	0.684	0.645	0.847	0.600
Harana	**0.744**	**0.743**	**0.886**	**0.651**

Model	Under Seg	Over Seg	Overall
CRNN	0.681	0.738	0.639
frog	0.681	0.724	0.624
RuleSCRF	0.666	0.741	0.625
Harana	**0.722**	**0.747**	**0.649**

- ❑ Through ablation studies, we demonstrate the importance of all the architecture components.

Model	Root Acc	Quality Acc	Overall Acc	Under Seg	Over Seg	Overall Seg
Harana	**0.744**	0.743	**0.651**	**0.722**	0.747	0.649
Harana - no semi-CRF	0.732	0.715	0.634	0.678	0.740	0.639
Harana - no Attention Fusing	0.741	0.738	0.650	0.716	**0.749**	0.645
Harana - no Absence Score	0.743	**0.746**	0.643	0.719	0.748	**0.650**

Future Work

- ❑ Support Audio Input
- ❑ Design more efficient training strategies to alleviate the time complexity of semi-CRF
- ❑ Enable real-time processing

Acknowledgement

This work is partially supported by National Science Foundation grants No. 1846184 and 2222129. Frank Cwitkowitz would like to thank the synergistic activities funded by NSF grant DGE-1922591.