Towards Improving Harmonic Sensitivity and Prediction Stability for Singing Melody Extraction
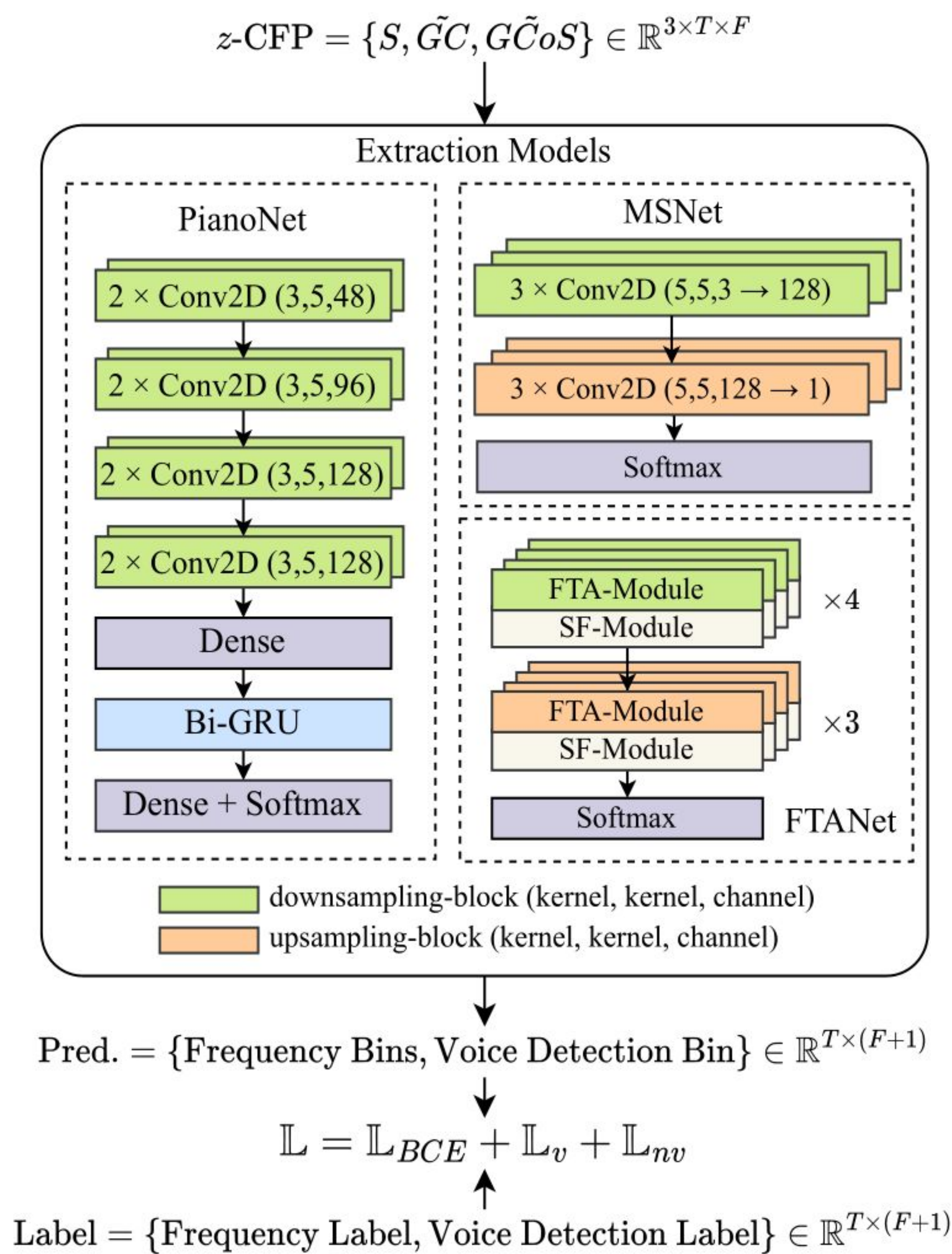
Keren Shao*, Ke Chen*, Taylor Berg-Kirkpatrick, Shlomo Dubnov

University of California San Diego

What is KKNet?

KKNet is a singing melody extraction model with:

- Feature designs on improving the harmonic sensitivity
- Loss designs on improving the prediction stability

$$z\text{-CFP} = \{S, \tilde{GC}, G\tilde{C}oS\} \in \mathbb{R}^{3 \times T \times F}$$

Extraction Models

PianoNet

- $2 \times$ Conv2D (3,5,48)
- $2 \times$ Conv2D (3,5,96)
- $2 \times$ Conv2D (3,5,128)
- $2 \times$ Conv2D (3,5,128)
- Dense
- Bi-GRU
- Dense + Softmax

MSNet

- $3 \times$ Conv2D (5,5,3 → 128)
- $3 \times$ Conv2D (5,5,128 → 1)
- Softmax

FTANet

- FTA-Module / SF-Module ×4
- FTA-Module / SF-Module ×3
- Softmax

downsampling-block (kernel, kernel, channel)
upsampling-block (kernel, kernel, channel)

$$\text{Pred.} = \{\text{Frequency Bins, Voice Detection Bin}\} \in \mathbb{R}^{T \times (F+1)}$$

$$\mathbb{L} = \mathbb{L}_{BCE} + \mathbb{L}_v + \mathbb{L}_{nv}$$

$$\text{Label} = \{\text{Frequency Label, Voice Detection Label}\} \in \mathbb{R}^{T \times (F+1)}$$

KKNet consists of the following three components:

- Z-CFP as input representation
- PianoNet as core extraction architecture
- Stability Loss Components as key converging objectives

I. Z-CFP

We propose Z-CFP, a revised representation based on CFP, as the model input.

- Z-CFP is obtained by performing the discrete z-transform on the audio power spectrum
- Z-CFP captures the trailing harmonics as important indications for melody extraction
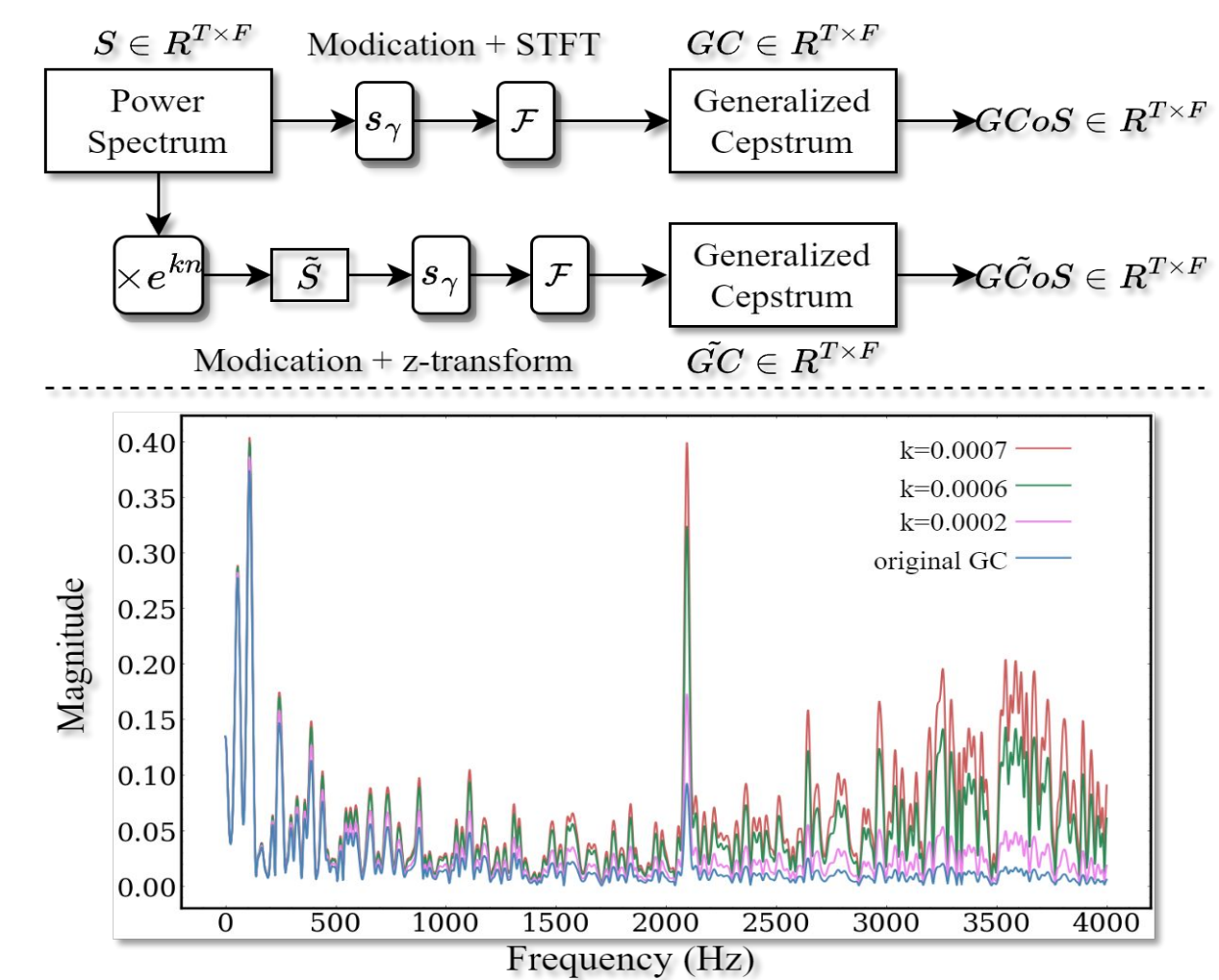
II. Stability Loss Components

We propose additional training objectives to enhance the prediction stability.

- Each component is setup as regularizers for spurious vocal and non-vocal segments
- Each component is design by a set of polynomials to identify these undesirable segments at or below the threshold duration
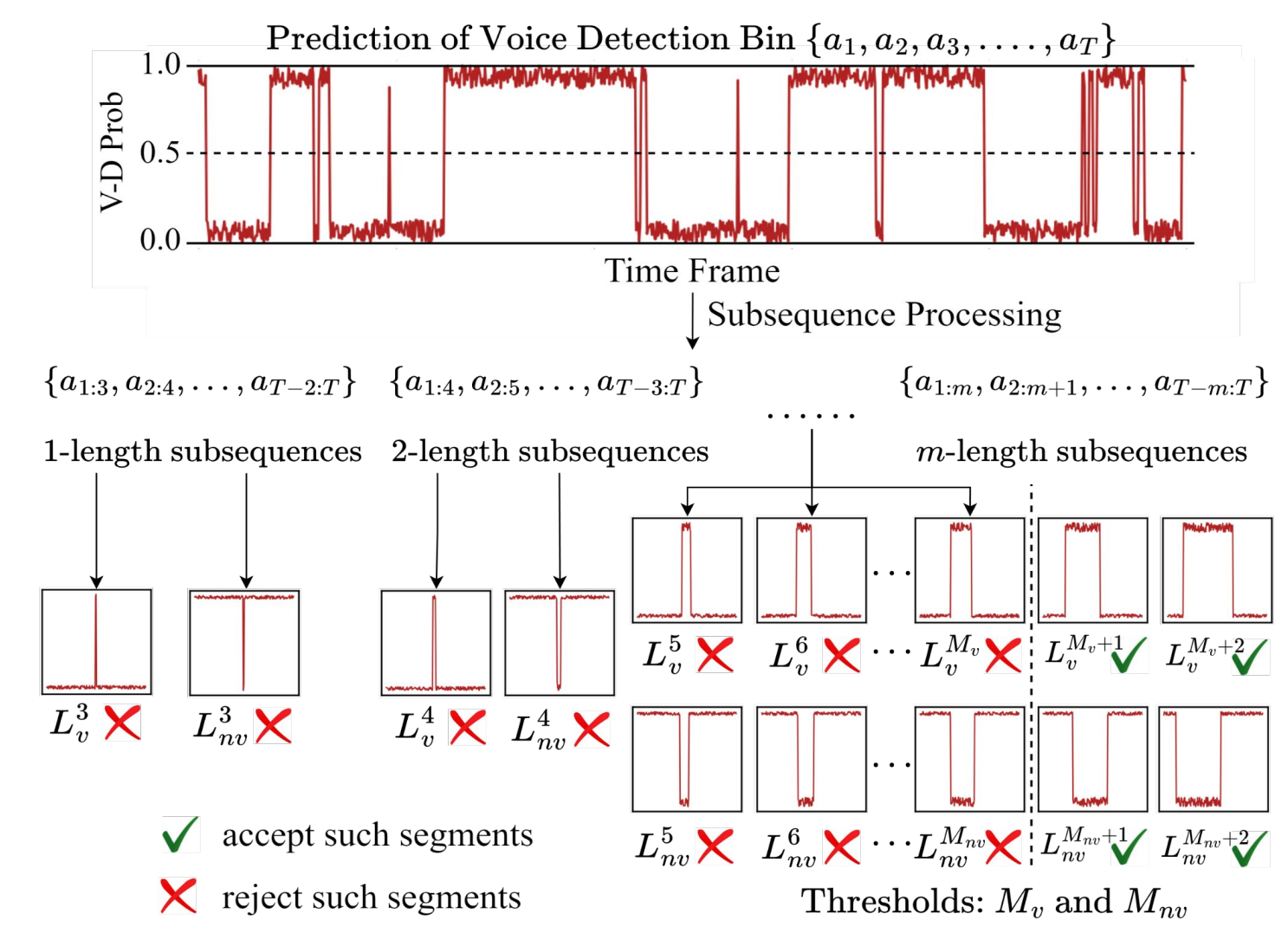
Experiments & Visualizations

- KKNet yields superior performance on three benchmark datasets of singing melody extractions
- KKNet avoids those spurious peaks whose length falls below the threshold, meanwhile maintains those long spurious segments.

Paper   Code

Overall Accuracy (OA) of Five Models on Three Datasets

MSNet SpecTNT FTANet TONet KKNet