

HKDSME: Heterogeneous Knowledge Distillation for Semi-supervised Singing Melody Extraction Using Harmonic Supervision

Anonymous Authors

1 IMPLEMENT DETAILS

For our proposed framework HKDSME, the input embedding dimension is $3 \times 320 \times 64$. We split a music track into several segments, each segment has 64 frames. We use Adam algorithm to optimize the proposed framework. The learning rate is set to $1e-4$. All training components are trained on a machine with two NVIDIA RTX 3090 GPUs. The batch size is set to 64.

The HKDSME consists of two trainable components: harmonic supervision, heterogeneous knowledge distillation. We will introduce the details of the two modules one by one: The harmonic supervision module consists of convolution layer that outputs four channels. The four channels are then fed into a softmax layer to perform four-class prediction. The heterogeneous knowledge distillation consists of a three-layer neural network: a convolution layer with kernel size of 1×1 , a SeLU layer and a fully connected layer to transform the feature map to 320×64 .

1.1 Details About Datasets

Since we use several public datasets for train and evaluate the proposed framework HKDSME, we would like to introduce the details about the dataset we used. We use the full MIR-1K dataset and 35 popular music tracks in MedleyDB to as the labeled data to train the proposed model, the total duration of the labeled training data is 4 hours and 33 minutes. Since our framework is under a semi-supervised setting, we also use 700 popular music tracks selected from FMA dataset as unlabeled data to train the model, the duration is about 8.5 hours. For testing, we use four public datasets: ADC2004, MIREX 05, MedleyDB and iKala, the total duration is 3 hours and 2 minutes. Please note that though we use MedleyDB to train and test our proposed model, the tracks in training and testing do not have an overlap.

1.2 Details About Metrics

Following the convention in the literature, we use the following metrics for performance evaluation: overall accuracy (OA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), voicing recall (VR) and voicing false alarm (VFA).

OA is defined as the proportion of all frames correctly estimated by the algorithm, where for non-melody frames this means the algorithm labeled them as non-melody, and for melody frames the algorithm both labeled them as melody frames and provided a correct F0 estimate for the melody. RPA measures the proportion of melody frames in the ground truth for which \hat{f} is considered correct (i.e., within half a semitone of the ground truth f^*). RCA gives a measure of pitch accuracy that ignores octave errors, a common error made by melody extraction systems. VR measures the proportion of frames labeled as melody frames in the ground truth that are estimated as melody frames by the algorithm. VFA

measures the proportion of frames labeled as non-melody in the ground truth that are mistakenly estimated as melody frames by the algorithm. These metrics are computed by the mir_eval library with the default setting e.g., a pitch estimate is considered correct if it is within 50 cents of the ground truth one. Among the metrics, OA is often considered more important.

1.3 Codes Open Source

In order to follow the anonymous policy of ACM MM, the codes link will be public upon the acceptance.