MIREX2020: AUDIO MELODY EXTRACTION USING NEW MULTI-TASK CONVOLUTIONAL RECURRENT NEURAL NETWORK

An-qi Huang, Hua-ping Liu
NetEase Cloud Music, China
{huanganqi01, liuhuaping}@corp.netease.com

ABSTRACT

This abstract presents our submission to the MIREX 2020 melody extraction task, whose goal is the identification of the melody pitch sequence from polyphonic musical audio. We propose a multi-task Convolutional Recurrent Neural Network (CRNN) that allows the model to improve accuracy when performing chroma detection, octave detection and singing voicing detection tasks simultaneously. In the frequency range corresponding to the chroma and octave results, the pitch results are calculated by weighted averaging based on the Constant Q Transform (CQT) signal.

1.INTRODUCTION

Melody extraction is a task that tracks pitch contour of singing voice in polyphonic music. Extracting melody, particularly from popular music,is an essential module for melody-based music retrieval systems, such as cover song identification [1] and query by humming [2]. In this paper, we focus on multi-task CRNN to extract the melody from audio signals, and improve the resolutions by signal-processing based method.

2.MELODY EXTRATION

2.1 Model

The audio files are resampled to 16 kHz and merged into mono channel. We use a blackmanharris window and a hop size of 256 samples (16ms) for CQT spectrogram . We used 365 bins from 65 Hz to 4200 Hz and consecutive 517 frames as CRNN input.

The CNN modules configured with three 2D convolutional blocks. The convolution filters have a filter size of 3×3 and the number of filters in the convolutional blocks gradually increases as 16, 32, and up to 64.We apply batch normalization on each convolutional layer and use the LeakyReLU as an activation function. Then, feature concat is applied only to the frequency axis.

Two layer of bidirectional RNN-LSTM is connected after the convolutional module to learn the temporal information of the features extracted by CNN. Two layer of full connect and Softmax activation function is used for three task (chroma detection, octave detection and singing voicing detection) output layer. We include dropout to the full connect layer to alleviate overfitting.

We quantized the pitch labels on 1 semitone scale with low resolutions. The pitch labels cover from C2(65.406Hz) to

C6(1046.5Hz). By detecting 4 octaves and 12 chromas, we can recognize this 48 pitches. A frame determined to be not a melody is assigned a zero pitch.

Within the range of 20 CQT frequencies (4 semitone) of pitch output from CRNN, we weighted the CQT energy as a weight and averaged it to get the pitch output on 1/5 semitone scale high resolutions.

2.2 Datasets and Versions

Some or all of the following datasets are used for our different submissions:

1KP: partial of the MIR-1K dataset [3].

1KPG: generated data based on 1KP.

M05T: mirex05TrainFiles, containing 13 clips, with average length of 30 seconds [4].

GM: generated MIDI files using [5].

The task of melody extraction is related to a pitch. Therefore, data augmentation using pitch shift is effective for task of melody extraction. We augment our training set by changing the global pitch of the audio content and labels. To reducing over fi tting and improve the performance, we expanded the existing training datasets by applying pitch-shifting by $\pm$1,2 semitones.

Our different versions of submissions use different lable of training data, for training voicing detection model.

HL1: Only the singing voice (male, female) will have pitch labels, and the rest will be assigned a zero pitch.

AH1: All the main melodic voice will have pitch labels, and the rest will be assigned a zero pitch.

3.REFERENCES

[1] Joan Serra, Emilia G´omez, and Perfecto Herrera. Audio cover song identification and similarity: Background,approaches,evaluation,andbeyond.Advances in Music Information Retrieval, 274:307–332, 2010.

[2] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C Smith. Query by humming: musical informationretrievalinanaudiodatabase.InnProc.oftheACM international conference on Multimedia, pages 231–236. ACM, 1995.

[3] C.-L. Hsu and J.-S. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," IEEE Trans. Audio, Speech, and Language Processing, volume 18, issue 2, p.p 310-319, 2010

[4] G. Poliner and D. Ellis, "A Classification Approach to Melody Transcription," Proc. Int. Conf. on Music Info. Retrieval (ISMIR), London, September 2005.

[5] K. Schutte, "MIDI file tools for MATLAB," Available: https://github.com/kts/matlabmidi/blob/master/src/synth.m