MELODY EXTRACTION USING MULTI-COLUMN DEEP NEURAL NETWORKS (MIREX 2016)

Sangeun Kum, Changheun Oh, Juhan Nam
Graduate School of Culture Technology
Korea Advanced Institute of Science and Technology
{keums, thecow, juhannam}@kaist.ac.kr

ABSTRACT

In this paper we describe our system for the audio melody extraction task of the Music Information Retrieval Evaluation eXchange (MIREX) 2016. We present a classification-based approach for melody extraction on vocal segments using multi-column deep neural networks. In the proposed model, each of neural networks is trained to predict a pitch label of singing voice from spectrogram, but their outputs have different pitch resolutions. The final melody contour is inferred by combining the outputs of the networks and post-processing it with a hidden Markov model. Our system also includes a singing voice active detector to select singing voice frames using an additional deep neural network. It is trained with spectrogram and the output of deep neural networks for melody extraction.

1. INTRODUCTION

Extracting melody, particularly from singing voice, is important to implement systems which use a melody to search songs, such as cover song identification [6] and query by humming [2]. In this paper, we focus on algorithms to extract the singing melody from audio signals. Singing melody extraction is a task that tracks pitch contour of singing voice in polyphonic music. While the majority of melody extraction algorithms are based on computing a saliency function of pitch candidates or separating the melody source from the mixture, data-driven approaches based on classification have been rarely explored [3]. we present a classification-based approach using a deep learning algorithm. The system is comprised of five main block, a preprocessing, a data augmentation, a Multi-Column Deep Neural Networks (MCDNN) for melody extraction, Hidden Markov Model (HMM), and Singing Voice Detector (SVD).

2. METHOD

2.1 Preprocessing

The audio files are resampled to 8 kHz and merged into mono channel. We use a 1024 point Hann window and a hop size of 80 samples for spectrogram, and finally compress the magnitude by a log scale. The only 256 bins from 0 Hz to 2000 Hz are used for training, because the human singing voices are mainly presented in the frequency bands and a level of singing voice is greater than a level of background music. We use the RWC pop music database as our main training set [5] and 60 vocal tracks of the MedleyDB dataset as an additional training set [1].

2.2 Data Augmentation

Data augmentation is an important technique to help reducing overfitting and to improve the performance. The task of melody extraction is related to a pitch. Therefore, we expect that the pitch shifting will be effective for our task. Specifically, we expanded the existing training datasets by applying pitch-shifting by $\pm 1, 2$ semitones. The result showed a significant improvement of predicting the pitch accuracy. we augment our training set by changing the global pitch of the audio content. A phase-vocoder method approach is used for more natural transposition [4].

2.3 Multi-Column Deep Neural Networks

Diagram of process units are shown in Figure 1. In first step for training multi-column deep neural network, we take multi-frame spectrogram as input of each column of deep neural network unit that make DNNs capture contextual information and predict a pitch label. The DNNs have same configuration with three layers and ReLUs. Otherwise, Each DNNs has different pitch resolutions which are one semitone, half semitone, and quarter semitone. Given the outputs of the columns, we compute the combined posterior as follows:

$$y^{N}_{MCDNN} = 10^{\sum_{i=1}^{N} \log\left(y^{i}_{DNN}\right)} \tag{1}$$

where y^{i}_{DNN} corresponds to the prediction from i^{th} column DNN, and N corresponds to the number of total columns. We use multiplication in a maximum-likelihood sense, assuming that the column DNNs are independent.

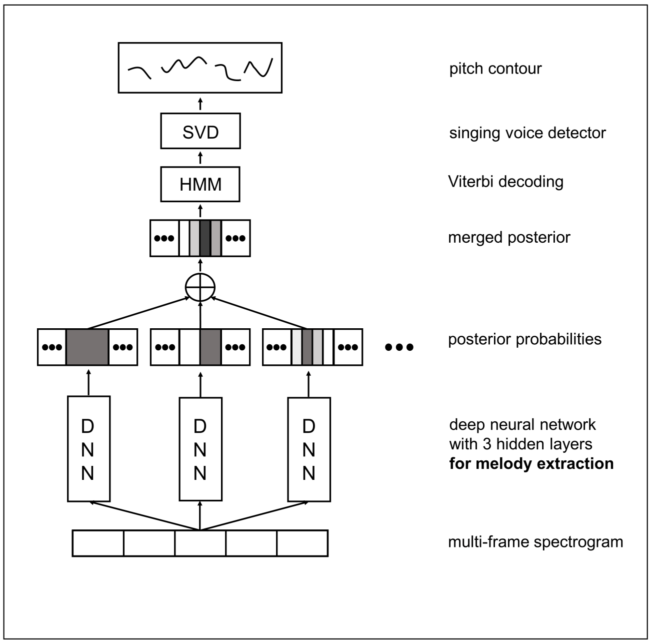

Figure 1. Block diagram of our proposed multi-column deep neural networks for singing melody extraction

2.4 Temporal Smoothing by HMM

The Viterbi decoding based on a HMM is conducted to capture long-term temporal information that appear on the pitch contours of singing voices. We implemented the HMM, following the procedure in [3]. The prior probabilities and transition matrix can be estimated from ground-truth of the training set. The prediction of whole tracks is used as posterior probabilities.

2.5 Singing Voice Detection

we apply our proposed singing voice detector using the deep neural network. The DNN for voice detector is trained to predicts the singing voice frame from a spectrogram and an output of the SCDNN(res=1) [1] we train a single-column DNN for voice detection using RWC and the MedleyDB datasets. The DNN takes a single frame spectrogram as input data and is configured with three hidden layers and ReLUs for the nonlinear function in common. The output layers predict singing voice segment.

3. REFERENCES

[1] Bittner, Rachel M and Salamon, Justin and Tierney, Mike and Mauch, Matthias and Cannam, Chris and Bello, Juan Pablo: “Multi-column deep neural networks for image classification,” *Proceedings of the International Symposium on Music Information Retrieval*, pp.155–160, 2014.

[2] Dannenberg, Roger B and Birmingham, William P and Tzanetakis, George and Meek, Colin and Hu, Ning and Pardo, Bryan: “The MUSART testbed for query-by-humming evaluation,” *Computer Music Journal*, Vol.28, No.2, pp.34–48, 2004.

[3] Ellis, Daniel PW and Poliner, Graham E: “Classification-based melody transcription,” *Machine Learning*, Vol.65, No.2, pp.439–456, 2006.

[4] Laroche and Jean: *Applications of Digital Signal Processing to Audio and Acoustics*, Springer US, Boston, MA, 2002.

[5] Masataka Goto and Hiroki Hashiguchi and Takuichi Nishimura and Ryuichi Okar: “RWC Music Database: Popular, Classical, and Jazz Music Databases,” *Proceedings of the International Symposium on Music Information Retrieval*, pp.287–288, 2002.

[6] Serra, Joan, Gómez, Emilia and Herrera, Perfecto: *Advances in Music Information Retrieval*, Springer, Utrecht, 2010.

[1] “res=1” indicates pitch resolution in semitone unit. “res=2”, and “res=4” means progressively higher resolutions by a factor of 2.