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

We describe our system for audio melody extraction task of the Music Information Retrieval Evaluation eXchange (MIREX) 2016. The system is based on multi-column deep neural networks. 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 used for melody extraction.

1. INTRODUCTION

Extracting melody, particularly from polyphonic music, is an essential module for melody-based music retrieval systems, such as cover song identification [2] and query by humming [5]. 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 [1]. We present a classification-based approach using a deep learning algorithm [6]. The system is built with five components, which include preprocessing, data augmentation, multi-column deep neural networks, hidden Markov model (HMM), and singing voice detector.

2. SYSTEM DESCRIPTION

2.1 Datasets

We used the RWC pop music database as our main training set [3] and 60 vocal tracks of the MedleyDB dataset as an additional training set [4]. They contain pitch labels for the singing voice melody.

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.

2.3 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.

2.4 Multi-Column Deep Neural Networks

The system for classifying melody is shown in Figure 1. The DNNs takes multiple frames of spectrogram as input and have three hidden layers with ReLUs in common. However, each DNNs has a different pitch resolution, specifically, one semitone, half semitone, and quarter semitone, respectively. The outputs of the columns are combined as follows:

$$y^N_{MCDNN} = \prod_{i=1}^{N} (y^i_{DNN} + \epsilon) \tag{1}$$

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. We add a small value, ϵ to prevent numerical underflow.

2.5 Temporal Smoothing by HMM

The Viterbi decoding based on a HMM is conducted to capture long-term temporal information that appears on the pitch contours of singing voices. We implemented the HMM, following the procedure in [1]. The prior probabilities and transition matrix are estimated from ground-truth

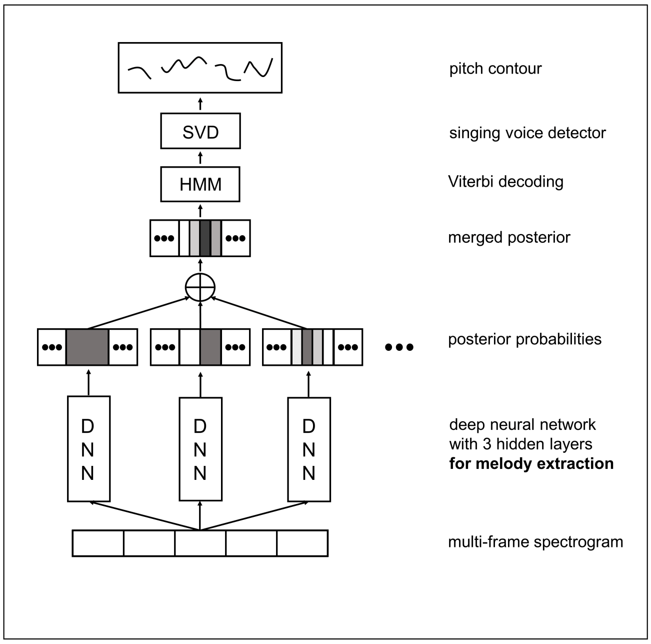

Figure 1. Block diagram of our proposed multi-column deep neural networks for singing melody extraction

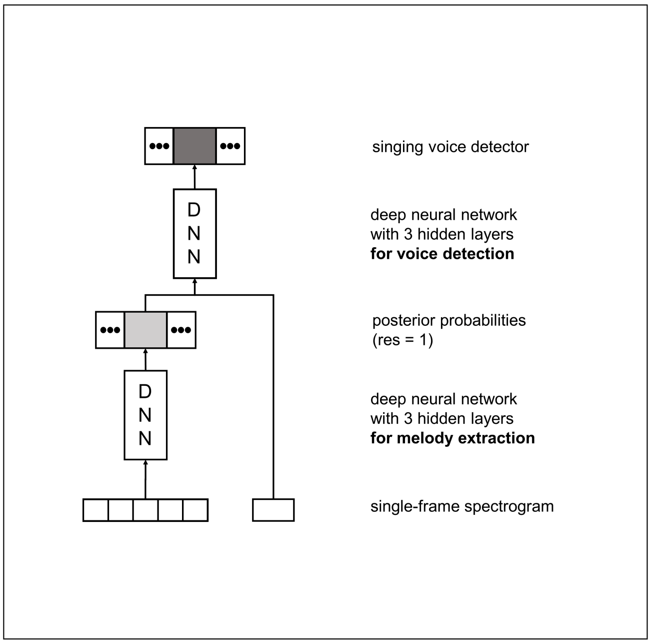

Figure 2. Block diagram of our proposed deep neural networks for singing voice detection

of the training set. The prediction of a whole track is used as posterior probabilities.

2.6 Singing Voice Detection

An additional DNN is separately trained to predict the presence of singing voice. Our architecture of the singing voice detector is illustrated in Figure 2. We combined a single frame of spectrogram and the output of the melody extraction DNN with semi-tone resolution as input. The DNN is configured with three hidden layers and ReLUs for the nonlinearity.

3. REFERENCES

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

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

[3] 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.

[4] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam and Juan Pablo Bello: "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," *Proceedings of the International Symposium on Music Information Retrieval*, pp.155–160, 2014.

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

[6] Sangeun Kum, Changheun Oh, and Juhan Nam: "Melody Extraction on Vocal Segments Using Multi-Column Deep Neural Networks" *Proceedings of the International Symposium on Music Information Retrieval*, 2016.