A HYBRID NEURAL NETWORK BASED ON THE DUPLEX MODEL OF PITCH PERCEPTION FOR SINGING MELODY EXTRACTION

Hsin Chou, Ming-Tso Chen, and Tai-Shih Chi

Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 300, Taiwan

ABSTRACT

In this paper, we build up a hybrid neural network (NN) for singing melody extraction from polyphonic music by imitating human pitch perception. For human hearing, there are two pitch perception models, the spectral model and the temporal model, in accordance with whether harmonics are resolved or not. Here, we first use NNs to implement individual models and evaluate their performance in the task of singing melody extraction. Then, we combine the NNs to constitute the composite NN to simulate the duplex model, which complements the pitch perception from unresolved harmonics of the spectral model using the temporal model. Simulation results show the proposed composite NN outperforms other conventional methods in singing melody extraction.

Index Terms— pitch perception, duplex model, melody extraction, deep neural network, CNN

1. INTRODUCTION

Melody extraction is one important topic in the research field of music information retrieval [1]. Many high level musical tasks use the main melody as prior information, such as in music genre classification [2], query-by-humming [3], cover song identification [4] and voice separation [5]. Many algorithms have been proposed for melody extraction and most of them consider characteristics of melody for their designs [6]. In contrast, we think melody stems from human pitch perception. Therefore, we tackle this problem from the viewpoint of perception.

For human pitch perception, two major models have been proposed, the spectral model and the temporal model [7]. For the spectral model, pitch is assumed derived from resolved spectral components on the auditory spectrogram [8][9][10]. Most of these studies suggest the human brain has templates to sift the harmonic structure of the sound to determine the pitch. However, the spectral model cannot explain some auditory sensation such as the residue pitch which is associated with unresolved harmonics [11]. On the other hand, studies on the temporal model typically emphasize on higher-number harmonics which cannot be resolved by the spectral model [12][13]. Autocorrelation is often used in these studies for pitch detection. It is treated as one possible way to approximately detect the within-channel periodicity of the sound [14]. These two models emphasize on different parts of harmonics and each of them cannot explain all pitch-related psychoacoustic experiment results alone. Therefore, it has been proposed that the human brain may have more than one mechanism for detecting pitch [15].

Based on the pitch perception theory of human hearing, we design a NN to extract singing melody. First, we build a convolutional neural network (CNN) to simulate the spectral model because it just behaves like using templates (kernels) to sift harmonic components. Then, we build a deep neural network (DNN) to simulate the temporal model using the time domain autocorrelation function as input of the DNN. Finally, we combine the NNs for the spectral and the temporal models to form a hybrid NN.

The rest of this paper is organized as follows. In Section 2, we describe the proposed hybrid NN and related signal pre-processing and post-processing. In Section 3, we demonstrate details of each individual NN, which simulates either the spectral model or the temporal model, and discuss corresponding experiment results. The overall results from the hybrid NN are demonstrated in Section 4 and the conclusion is given in Section 5.

2. PROPOSED HYBRID NEURAL NETWORK

2.1. Architecture and pitch labels

The architecture of the proposed hybrid NN for melody extraction is shown in Fig. 1. It consists of two NNs which mimic the spectral model and the temporal model to detect pitch from resolved (lower-number) and unresolved (higher-number) harmonics, respectively. Both NNs can work independently for melody extraction.

We treated melody extraction as a classification problem. We followed the study in [16] to quatize the plausible pitch range from D2 (f0=73 Hz) to F#5 (f0=740 Hz) in a step of 50 cent (1/2 semitone) to give us a total of 82 states. In addition, an "unvoice" state was also incorporated for encoding/decoding the melody contour. Therefore, the fully-connected output layer of the hybrid NN had a total of 83

This research is supported by the Ministry of Science and Technology, Taiwan under Grant No MOST 105-2221-E-009-152-MY2.

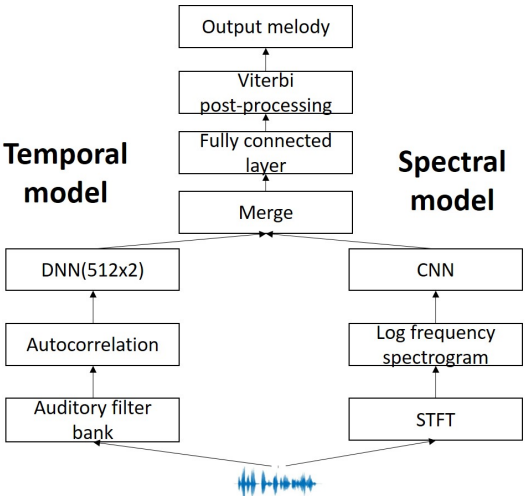

Fig. 1. Architecture of the proposed hybrid NN.

output units. To produce the probability for each state, the cross-entropy was used as the objective function in training the hybrid NN.

2.2. Viterbi post-processing

Each value of the 83 output units of the hybrid NN indicates the state probability. Traditionally, the state with the maximum value is selected as the final output of pitch/melody. However, the proposed hybrid NN sometimes produces temporal discontinuity along the pitch/melody contour just like any pitch tracking algorithm. Therefore, similar to these approaches in [17][18], the Viterbi tracking algorithm is applied on the outputs of the NN for temporal smoothing to correct some transient errors. The transition matrix used by the Viterbi tracking algorithm was generated from the training data.

2.3. Dataset and evaluation metrics

For evaluations, we used the iKala [19] and the MIR-1K [20] datasets. The MIR-1K dataset consists of 1000 song clips extracted from 110 Chinese karaoke pop songs sung by 8 female and 11 male nonprofessional singers. A clean music accompaniment track and a mixture track are available for each clip, which is sampled at 16 kHz and roughly with a duration of 4 to 13 seconds. We used 740 clips for training, and the remaining 260 clips for test. The iKala dataset contains 252 Chinese karaoke song clips sampled at 44.1 kHz. We used the first 200 clips for training and the remaining 52 clips for test. All the used clips were generated by mixing the singing voice with the background music of equal energy. For scalability, we resampled all signals to 16 kHz.

The standard evaluation metrics in melody extraction are voicing recall rate (VR), voicing false alarm rate (VFA), raw

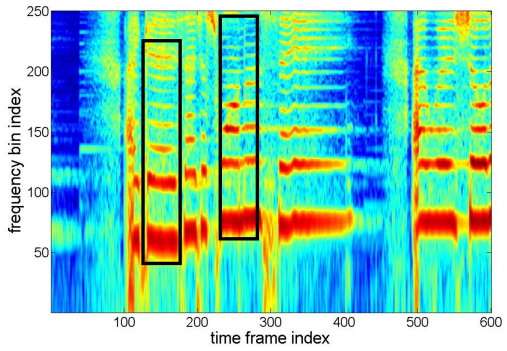

Fig. 2. The log-frequency spectrogram of a sample clip. Clearly, the harmonic pattern is almost invariant on the log-frequency axis.

pitch accuracy (RPA), raw chroma accuracy (RCA) and overall accuracy (OA). We computed these rates using the Python library of the mir_eval tool [21].

3. EXPERIMENT RESULTS OF INDIVIDUAL NN

In this section, we demonstrate and discuss results from each of the NNs which simulate the spectral model and the temporal model of pitch perception. To optimize both NNs, we used Adam optimizer with the learning rate automatically tuned. Both NNs were trained with 15 epochs. Since the architectures of both NNs are not very deep, 15 epochs are enough to make the NNs converge.

3.1. Results from spectral-model inspired CNN

We use a CNN to simulate the spectral model of pitch perception. Not like the study [22], which used the CNN on the linear-frequency spectrogram (LinFS) to detect pitch, we use the CNN on the log-frequency spectrogram (LogFS) due to the fact that the harmonic pattern is invariant on the LogFS with changing pitch. Fig. 2 shows the LogFS of a sample clip with the invariant harmonic pattern clearly. As in the human auditory system, the harmonics are logarithm arranged and resolved up to about the 10th harmonic due to the critical bandwidth of the cochlea [7]. It is postulated that a template of harmonics may exist in human auditory system to match the harmonic pattern of the sound for determining pitch [10]. Accordingly, we use the CNN, whose kernels behave like templates and the convolution operation along the log-frequency axis behaves like sifting the harmonics of the input sound, to simulate the spectral model of human pitch perception.

As the input of the CNN, we first generated the LinFS of the sound using the short-time Fourier transform (STFT) with a window size of 48 ms and 50% overlap between windows, then we rearranged the LinFS into the LogFS with the resolution of 48 bins per octave. The CNN had one convolutional

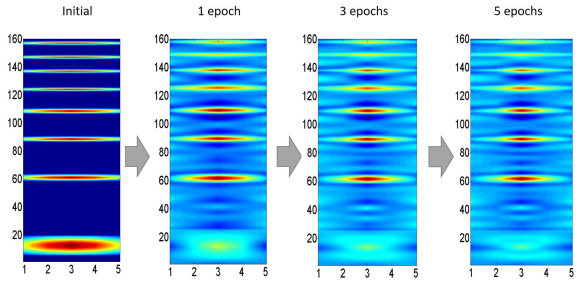

Fig. 3. The designed initial kernel for the CNN and its evolving shapes during training.

layer and two fully connected layers with 512 nodes. Only one kernel with the size of 160×5 (3 $1/3$ octaves $\times 120$ ms) was used in the CNN to account for the invariant harmonic pattern.

We designed the initial kernel on the log-frequency axis by imitating the excitation pattern on the auditory spectrum to the first 8 resolvable harmonics [10]. In addition, we applied a Gaussian window on the 5 frames across the time axis. The left panel of Fig. 3 shows the designed initial kernel of the CNN. The other panels show the kernel after different numbers of epochs during training. We can observe that the kernel wasn't changing a lot during training. It still looked like the initial harmonic pattern.

This 1-kernel CNN subsystem was tested using the iKala dataset and the OA scores were 79.07% and 78.55% using the designed and randomly initialized kernels, respectively. The performance difference was very small because input signals were LogFS such that the randomly initialized kernel eventually evolved into a shape very similar to the designed kernel. On the other hand, The CNN with the designed kernel converged at a much faster speed than the randomly-initialized CNN. We also tested the system with 5, 10 and 15 kernels. Comparing with the 1-kernel system, the improvement in the OA scores was less than 1% but at the cost of huge computational load. In addition, many of those kernels evolved into similar patterns at the end, which implied they provided redundant information. These results kind of support the hypothesis that there may be one harmonic template in human auditory system for pitch detection.

3.2. Results from temporal-model inspired DNN

Temporal models for pitch perception often use autocorrelation to estimate pitch. Therefore, we used temporal autocorrelation as input to train the DNN. First, all signals were passed through a bank of gammatone filters, which are often used in modeling cochlear filters. We used 40 gammatone filters whose center frequencies are equally distributed on the ERB scale between 80 Hz and 6 kHz. Then, we computed normalized autocorrelation out of the duration of 48 ms in each

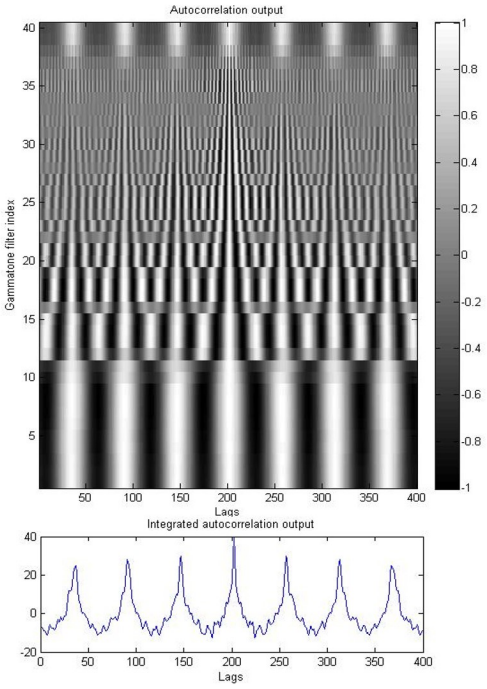

Fig. 4. Autocorrelation outputs of gammatone filters for a sample signal (top panel) and the integrated autocorrelation output (bottom panel).

channel. The top panel of Fig. 4 displays the autocorrelation results of a sample section of a signal in all subbands. Some studies further performed dimension reduction on the multi-channel autocorrelation output for pitch detection [17][23]. Here we simply integrated all information across channels by collapsing the autocorrelation results over channels. The bottom panel of Fig. 4 shows the integrated autocorrelation result of the top panel. Clearly, the periodicity of the sample signal is prominently shown in the bottom panel. The integrated result with dimension of 1136 was directly put into the DNN as the training data. We didn't perform any further procedures, such as dimension reduction, but hoped the DNN can learn the optimal procedures itself. The DNN consisted of two hidden layers of 512 units with the ReLU activation function and the output layer of 83 units corresponding to pitch labels.

The OA score of the DNN was 75.27% for iKala dataset. Although the temporal-model inspired DNN performed worse than the spectral-model inspired CNN, we can still deduce the integrated autocorrelation function can provide valid information for pitch/melody detection.

4. EXPERIMENT RESULTS OF THE HYBRID NN

In the previous section, we demonstrate both the spectral-model and temporal-model inspired NNs work in melody extraction. However, in pitch perception theory, neither spec-

Table 1. Melody extraction evaluations in terms of VR, VFA, RPA, RCA and OA using iKala dataset. All scores are displayed in %. Results from the spectral and the temporal models inspired NNs are also listed in the bottom two rows.

	VR	VFA	RPA	RCA	OA
Proposed	86.14	14.04	79.98	81.54	81.28
HPSS+Prop.	83.42	13.92	74.43	75.97	78.28
MCDNN [16]	85.85	15.05	77.88	79.60	80.22
Melodia [25]	82.02	26.71	75.99	78.36	72.80
Spec. Model	85.44	15.51	76.40	78.22	79.07
Temp. Model	83.17	27.43	76.61	78.47	75.27

Table 2. Melody extraction evaluations in terms of VR, VFA, RPA, RCA and OA using MIR-1k dataset. All scores are displayed in %. Results from the spectral and the temporal models inspired NNs are also listed in the bottom two rows.

	VR	VFA	RPA	RCA	OA
Proposed	82.73	16.14	72.23	75.38	75.64
HPSS+Prop.	75.35	12.37	64.29	67.72	71.12
MCDNN [16]	78.36	14.25	65.21	68.30	71.22
Melodia [25]	85.10	30.80	72.95	75.74	69.61
Spec. Model	83.63	21.31	68.81	72.16	71.70
Temp. Model	81.57	26.76	67.87	71.71	69.44

tral nor temporal model can explain all psychoacoustic experiment results. Some researchers believe both mechanisms coexist for pitch perception and proposed the duplex (or unity) pitch perception model [15][24]. Accordingly, we integrate these two NNs into a hybrid NN, as shown in Fig. 1, to combine complementary information for melody extraction. The hybrid NN was formed by cascading the output layers of the CNN and the DNN followed by two 512-node fully connected layers for higher level integration. For the purpose of scalability, batch normalization was done in the first hidden layer and 20% dropout was adopted to prevent data over-fitting.

For performance comparison, we implemented another recently proposed DNN-based method [16]. In addition, we also compared the proposed hybrid NN with the well-known non-DNN method, Melodia [25]. For NN-based methods, including the proposed method and the method in [16], both the iKala and the MIR-1k dataset were used for training. The test results for each dataset are listed in Table 1 and Table 2 separately. It was suggested that using the HPSS (harmonic/percussive source separation) method for per-processing help the melody extraction performance [26]. Therefore, following the settings of that study, we performed HPSS on the STFT of the input signal with a window size of 300 ms. The HPSS decomposed the signal into h1 and p1 under the condition of high frequency resolution. As suggested in that study, we used p1 as training data for our proposed hybrid NN. The test results of the proposed NN with the HPSS pre-processing are listed in Table 1 and Table 2 as from the "HPSS+Prop." method.

Test results shown in Table 1 using the iKala dataset demonstrate the proposed hybrid NN outperforms the individual CNN (with the OA score of 79.07%) and DNN (with the OA score of 75.27%) in melody extraction. Results in both tables consistently confirm that the proposed method performs better than the DNN-based method MCDNN [16] and the non-DNN method Melodia [25], and all DNN-based methods outperform the non-DNN method. Surprisingly, we found using HPSS as a preprocessing method does not provide any benefits to our system.

5. CONCLUSION

Inspired by the duplex (or unity) model of pitch perception, we built up a hybrid neural network, including a 1-kernel CNN and a DNN, for melody extraction. Considering characteristics of resolved harmonics in the spectral domain, we designed the initial pattern of the kernel of the CNN. Considering characteristics of unresolved harmonics in the temporal domain, we integrated the autocorrelation across gammatone filters as the input to the DNN. Experiment results show that the temporal-model inspired DNN does provide complementary information to the spectral-model inspired CNN when extracting singing melody. In addition, the proposed hybrid NN produces higher OA scores than the compared DNN-based method and non-DNN method using both the iKala and the MIR-1k dataset. This study shows the spectral and temporal information should be jointly utilized for better detection of pitch/melody.

6. REFERENCES

[1] Justin Salamon, Emilia Gómez, Daniel PW Ellis, and Gaël Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[2] Justin Salamon, Bruno Rocha, and Emilia Gómez, "Musical genre classification using melody features extracted from polyphonic music signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 81–84.

[3] Chung-Che Wang and Jyh-Shing Roger Jang, "Improving query-by-singing/humming by combining melody and lyric information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 798–806, 2015.

[4] Peter Foster, Simon Dixon, and Anssi Klapuri, "Identifying cover songs using information-theoretic measures of similarity," *IEEE/ACM Transactions on Audio,*

Speech and Language Processing (TASLP), vol. 23, no. 6, pp. 993–1005, 2015.

[5] Yukara Ikemiya, Kazuyoshi Yoshii, and Katsutoshi Itoyama, "Singing voice analysis and editing based on mutually dependent f0 estimation and source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 574–578.

[6] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[7] William A Yost, "Pitch perception," *Attention, Perception, & Psychophysics*, vol. 71, no. 8, pp. 1701–1715, 2009.

[8] Julius L Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *The Journal of the Acoustical Society of America*, vol. 54, no. 6, pp. 1496–1516, 1973.

[9] Frederic L Wightman, "The pattern-transformation model of pitch," *The Journal of the Acoustical Society of America*, vol. 54, no. 2, pp. 407–416, 1973.

[10] Shihab Shamma and David Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *The Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2631–2644, 2000.

[11] Roelof J Ritsma, "Existence region of the tonal residue. i," *The Journal of the Acoustical Society of America*, vol. 34, no. 9A, pp. 1224–1229, 1962.

[12] Ray Meddis and Michael J Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification," *The Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2866–2882, 1991.

[13] JCR Licklider, "periodicity pitch and place pitch," *The Journal of the Acoustical Society of America*, vol. 26, no. 5, pp. 945–945, 1954.

[14] Malcolm Slaney and Richard F Lyon, "On the importance of time-a temporal representation of sound," *Visual representations of speech signals*, vol. 95116, 1993.

[15] Ray Meddis and Lowel OMard, "A unitary model of pitch perception," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–1820, 1997.

[16] Sangeun Kum, Changheun Oh, and Juhan Nam, "Melody extraction on vocal segments using multi-column deep neural networks.," in *ISMIR*, 2016, pp. 819–825.

[17] Kun Han and DeLiang Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 2158–2168, 2014.

[18] Daniel PW Ellis and Graham E Poliner, "Classification-based melody transcription," *Machine Learning*, vol. 65, no. 2, pp. 439–456, 2006.

[19] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 718–722.

[20] Chao-Ling Hsu and Jyh-Shing Roger Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[21] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[22] Hong Su, Hui Zhang, Xueliang Zhang, and Guanglai Gao, "Convolutional neural network for robust pitch determination," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 579–583.

[23] Mingyang Wu, DeLiang Wang, and Guy J Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[24] Joseph Carl Robnett Licklider, "A duplex theory of pitch perception," *The Journal of the Acoustical Society of America*, vol. 23, no. 1, pp. 147–147, 1951.

[25] Rachel M Bittner, Justin Salamon, Slim Essid, and Juan Pablo Bello, "Melody extraction by contour classification.," in *ISMIR*, 2015, pp. 500–506.

[26] François Rigaud and Mathieu Radenen, "Singing voice melody transcription using deep neural networks.," in *ISMIR*, 2016, pp. 737–743.