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, National Chiao Tung University, Hsinchu, Taiwan

Introduction

- Based on the pitch perception theory of human hearing, we design a NN to extract singing melody.
- For human pitch perception, two major models have been proposed, the spectral model and the temporal model .

Architecture

- 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.
- The Viterbi tracking algorithm is applied on the outputs of the NN for temporal smoothing to correct some transient errors.

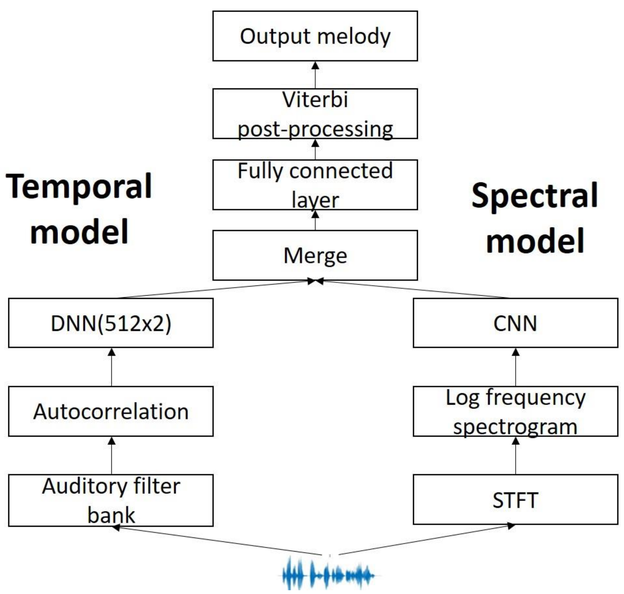

Fig. 1. Architecture of the proposed hybrid NN.

Spectral model

- We use CNN to simulate the spectral model of pitch perception.
- We training CNN on log-frequency spectrogram (LogFS) due to the fact that the harmonic pattern is invariant on the LogFS with changing pitch.
- 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

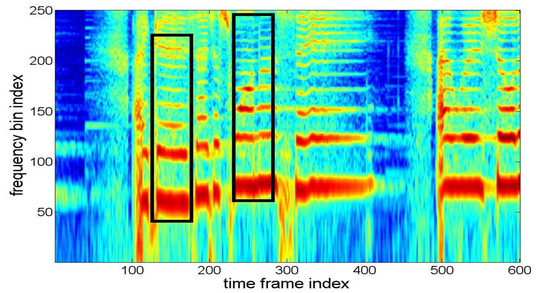

Fig. 2. The log-frequency spectrogram of a sample clip. Clearly, the harmonic pattern is almost invariant on the log-frequency axis.

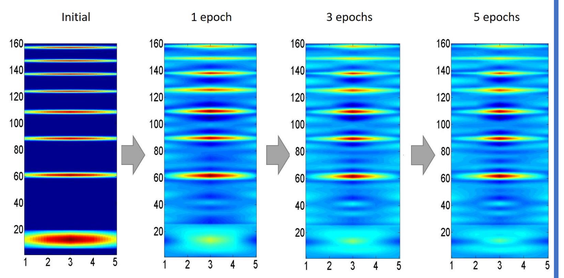

Fig. 3. The designed initial kernel for the CNN and its evolving shapes during training.

Temporal model

- We used temporal autocorrelation as input to train the DNN
- Signals were passed through a bank of gammatone filters, which are often used in modeling cochlear filters.

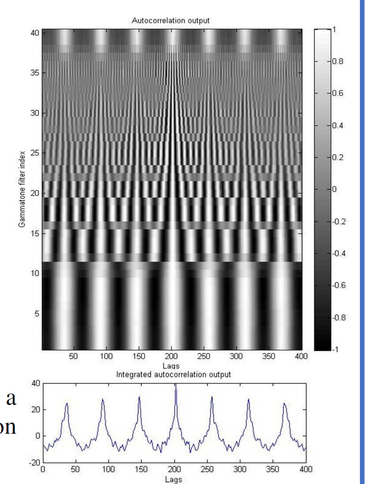

Fig. 4. Autocorrelation outputs of gammatone filters for a sample signal (top panel) and the integrated autocorrelation output (bottom panel).

Result

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

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.
- Experiment results show that the temporal-model inspired DNN does provide complementary information to the spectral-model inspired CNN when extracting singing melody.
- 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.