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.

Spectral model
- We use CNN to simulate the spectral model of pitch perception.
- We train 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.

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.

Result
- 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.