Word Tonality-conditioned Mandarin Singing Voice Synthesis Via a Long Short-term Memory Recurrent Neural Network

Yi-Jhe Lee[1], Chan-Chuan Chu[1], Fu-Rong Yang[1], Shan-Hung Wu[2], Yi-Wen Liu[1*]

[1]*Electrical Engineering, National Tsing Hua University, Hsinchu City, Taiwan*
[2]*Computer Science, National Tsing Hua University, Hsinchu City, Taiwan*

With the rapid advancements in artificial neural networks over recent years, it becomes possible for a machine to create musical contents that sound natural to human ears. In this research, we aim to teach a machine to sing by "sight-reading". Ten hours of Mandarin singing voice data have been recorded and symbolic representation of the songs is transcribed manually; by marking the musical phrases, the note boundaries, the word boundaries, and the phoneme onset and offset time, the machine can learn to find a mapping from the sequential symbolic representations of the songs (words + music) to a sequential parametric (called vocoder) representation of the singing voice. To demonstrate the idea, a long short-term memory (LSTM) network was built to learn this sequence-to-sequence mapping. The network has two stacked layers of LSTM cells, and can be configured to pass information recurrently in both the forward and backward directions in time. Since past research suggested that Mandarin word tonality tends to affect voice production during singing, we trained the system with or without revealing the tonality of words so results could be compared. Pilot results based on about 80 minutes of training data recorded from one female singer suggest that the machine can "sing" Mandarin-like utterances while faithfully reproducing the voice timbre of the singer. However, the pitch contour and the rhythm control are harder to train partly because unnaturalness in these musical aspects are less tolerable to human listeners.

Keywords: artificial intelligence, language, music, singing voice, long short-term memory

Email:ywliu@ee.nthu.edu.tw