Abstract

In this work we analyze a neural network structure capable of achieving a degree of invariance to speaker vocal tracts for speech recognition applications. It will be shown that invariance to a speaker's pitch can be built into the classification stage of the speech recognition process using convolutional neural networks, whereas in the past attempts have been made to achieve invariance on the feature set used in the classification stage. We conduct experiments for the segment-level phoneme classification task using convolutional neural networks and compare them to neural network structures previously used in speech recognition, primarily the time-delayed neural network and the standard multilayer perceptron. The results show that convolutional neural networks can in many cases achieve superior performance than the classical structures.