Contents

1 Introduction

Deep learning as a broad framework of methods has taken speech and natural language system-building by a storm. The deluge of work that has happened in the last 25 or so years has shown remarkable promise, and growth. Therefore we don't need to underscore the need to present here the progress that has been made since the late 80's. One of the primary goals of our paper, therefore, is to highlight some of the most significant approaches and the findings therein. While our approach mostly will be chronological, we will also focus on how deep neural networks (DNN) have fundamentally revolutionized our approach to speech processing systems in general, and Automatic Speech Recognition (ASR) and Text-to-Speech Synthesis (TTS) in particular. In addition, we want to focus on a very specific aspect within the use of DNNs in speech processing, namely the integration of linguistic knowledge in achieving some of the remarkable successes in the core tasks of speech processing. At the outset, we would like to outline that the goals of this paper are not to introduce the concepts of machine learning, but to specifically treat a class of learning algorithms that variously appear in the literature under the cover term deep learning. Essentially, all deep learning systems and architectures are a specific form of artificial neural networks which have been in existence since the earliest formulation by <insert reference>. While the most basic functions of the artificial neural network or perceptron remain the same and lot of advancement has been in the way the basic ingredient, in this case, the perceptron has been used to create architectures that are remarkable improvements over the initial attempts to use these machines for both classification and regression tasks.

In a series of seminal papers, **??** outline the use of multilayered neural networks in ASR and speaker recognition, respectively.

2 The architecture of Deep Neural Netowrks

Typically, DNNs refer to feedforward multi-layered artificial neural networks (ANN) with more than one layer of hidden units with a logistic function to traverse between the hidden layers and the output. Here we rely on **?** to outline the general architecture of DNNs. We will illustrate the functioning of the algorithms and the processes with an acoustic modeling task as discussed in **?**. Information from each hidden unit, j, is used along with a logistic function in order to map the total input from the previous layer, x_j, to a scalar state, y_j which is then sent to the following layer.

Here, as in 1 below, b_j refers to the bias associated with the unit j

$$y_j = logistic(x_j) = \frac{1}{1+e^{-x_j}}, x_j = b_j + \sum_i y_i w_{ij}, \quad (1)$$

$$p_j = \frac{exp(x_j)}{\sum_k exp(x_k)} \quad (2)$$

$$C = -\sum_j d_j log p_j \quad (3)$$

$$\Delta w_{ij}(t) = \alpha \Delta w_{ij}(t-1) - \epsilon \frac{\delta C}{\delta w_{ij}(t)} \quad (4)$$