Deep learning architectures in speech processing: Modeling dynamic speech gestures

March 27, 2017

1 Introduction

Deep learning as a broad framework of methods has taken speech and natural language processing 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. Our approach here is both to present and condense these findings, while we 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. The articulatory-acoustic portion of the speech chain is fundamentally non-linear and non-unique and our specific focus in the use of deep learning architectures is to highlight the success in a specific sub-task within broad speech processing and that is the articulatory-acoustic domain mapping. In that sense, the objective of this paper is to cover the most significant advances made in speech processing though deep learning models and architectures, but also focus on how deep learning architectures have helped unearth crucial generalizations between speech articulatory gestures and their acoustic manifestations. 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 for a significant amount of time, but regained currency in the mid-80s with publications emerging from the parallel distributed processing group at San Diego. The paper is organized as follows: In the following section, we motivate the need to look closely into the various deep learning initiatives and outline the importance of the major impacts of these learning mechanisms. In Section 3, we briefly, and

very generally, discuss the various architectures that help build deep learning models. In section 4, we concentrate on a specific use-case and outline the successes in this use-case, namely, the articulatory-acoustic mapping problem. In section 5, we offer some perspective and perhaps, hazard some speculation as to the future of these learning models.

2 Why Deep Learning?

We will begin this section with Manning [2015] who in his paper points to a few crucial changes that have come as a consequence of the expansion of deep learning systems in natural language processing. One of Manning's 2015 primary concerns and engagement in this paper has been to point out how natural language processing takes center stage as far as being one of the most important challenges for machine learning scientists and deep learning enthusiasts, alike. Manning's 2015 entreatment is a clarion call for linguists, NLP engineers and data scientists to shift focus away from beating benchmark dataset tasks and challenges, and to concentrate on "problems, approaches, and architectures". The import of Manning's 2015 appeal to re-engage with the cognitive and design goals of the study of human languages has been felt and responded to by work that has sought to understand deep learning architectures in the context of language cognition, and essentially has managed to re-center focus on NLP tasks as not just an engineering challenge, but also to re-imagine the goals of NLP research broadly within the cognitive and language sciences. In this paper, and especially section4, we discuss in detail how these goals have been achieved, and in which direction NLP research is moving armed with deep learning tools. We will treat the use of deep learning methods on solving acoustic variation problems that come about as a consequence of an essentially non-linear problem that of articulatory-acoustic mapping and articulatory-acoustic inversion. The non-linearities that arise in the articulatory-acoustic mapping are a direct consequence of the way speech articulators overlap with each other in a non-discrete fashion to produce contrastive sequences of segments.

Schmidhuber [2015] is a fairly comprehensive introduction to available architectures within deep learning. Notable concepts that find mention in this survey article include an essential concept in automatic learning mechanisms; *fundamental credit assignment problem* [Minsky, 1961]. Within the context of neural networks *learning* and *credit assignment* are synonomous. In that, input sequences of neurons produce activations which manage to activate other neurons and spread activations based on weighted connections between previously activated neurons and hidden layers. Ascertaining these activation weights, therefore, refers to the fundamental credit assignment problem. The aggregate activations of neurons between many stages depends on being able to assign credit so as to achieve the desired outcome. Just like in any other machine learning problem, these activations, and the associated weights can be utilized for solving both classification and regression problems, the advantage of deep neural networks in comparison with other learning methods is the non-linear na-

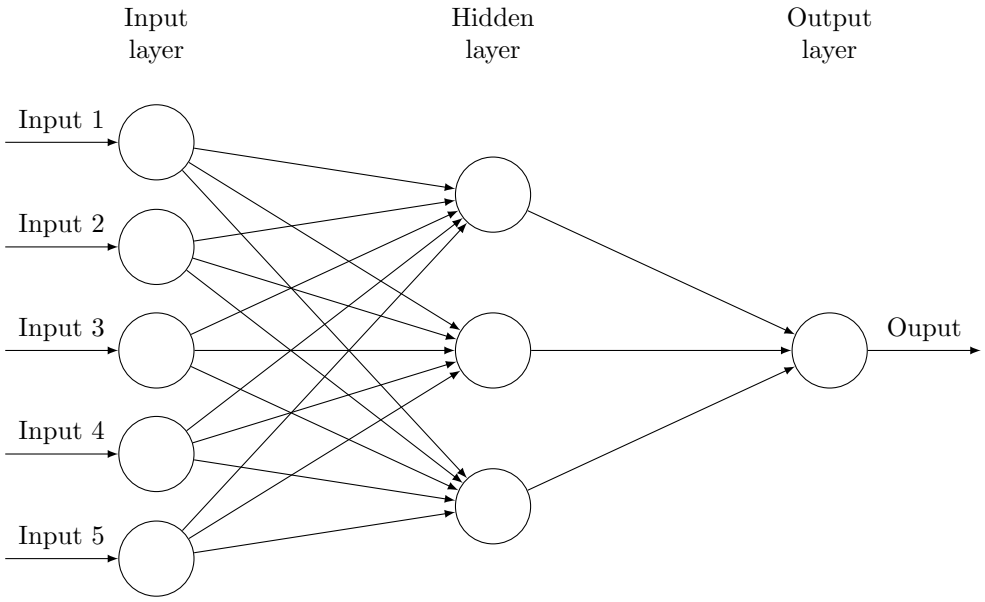


Figure 1: Basic design of a single hidden layer artificial neural network

ture of the computations that achieve the aggregate activation of the network.

The success of deep learning methods crucially hinges on several sets of advantages that these learning architectures provide when compared to other machine learning methods and algorithms. In most induction learning processes features need to be handcrafted and curated. In that respect, deep learning methods do not require feature selection as a process. Much of the learning involved in speech processing tasks requires understanding layers of representations which may indeed become more abstract as we get into greater neuronal depths. Deep learning methods are essentially task-neutral and since, they do not need handcrafted features the training, testing, and validation cycles tend to be agnostic to the nature of the task, and therefore highly replicable across domains and datasets. What is applicable to a task such as digit recognition can be easily applied to speech recognition with limited changes to the architectures. Though the architectures themselves may offer lesser or greater accuracy on the task at hand. In the most basic of formulations, sensory inputs are associated with nodes in a hidden layer through a vector of connection weights with each activation in turn associated with a bias term. These hidden units have corresponding connection weights that associate them with the output. For instance, we can characterize a simple artificial neural network as below:

The input, in this case any kind of sensory input, and for our case acoustic input (but also see articulatory parameters as input in section 4) could be

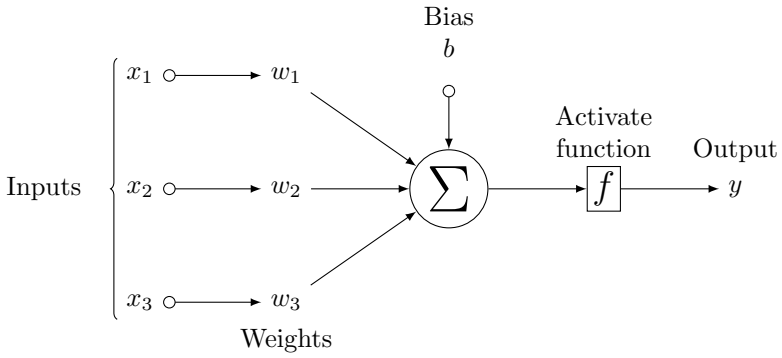


Figure 2: Connection weights and activation functions

associated with connection weights that are summed and computed using an activation function as in figure 2.

A consolidated

3 The architecture of Deep Neural Netowrks

While the basic structure and functions of the artificial feed-forward neural network or single-layer perceptron remain the same, a lot of advancement has been made in recent years by adding more layers to the architecture. In this way, the perceptron has been used as a building block 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, Bengio et al. [1989a,b] outline the use of multilayered neural networks in ASR and speaker recognition, respectively.

Typically, DNNs refer to feed-forward 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 Hinton et al. [2012] 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 Hinton et al. [2012]. 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

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

$$\Delta w_{ij}(t) = \alpha \Delta w_{ij}(t-1) - \epsilon \frac{\delta C}{\delta w_{ij}(t)} \tag{2}$$

3.1 Sequence Learning with Recurrent Neural Networks

Feed-forward neural networks (deep or otherwise) are effective for modelling non-linear relationships between fixed-length inputs and outputs. This however limits their application in learning problems which involve sequential data, such as to be found in the linguistic domain. Speech as well as text are inherently sequential in form, and to effectively model these, the basic feed-forward architecture needed to be extended.

Recurrent Neural Networks (RNN) Rumelhart et al. [1986] provided these necessary extensions to the basic feed-forward architecture in the form of internal feedback connections that allowed the network to propagate learned weights from past processing steps to the current one. Instead of a single fixed-length input vector x, RNNs take in a sequence of input vectors $\mathrm{x}^1,\ldots,\mathrm{x}^t$, which correspond (though not necessarily) to a sequence in time with timesteps t. At Intuitively, this internal state may be understood as providing the network with a memory mechanism. This feature of RNNs makes them suitable for learning/processing tasks on data that is temporally dynamic and sequential, such as speech. The use of RNNs has driven much of the recent advances in the state-of-the-art of artificial speech processing and recognition.

3.2 Deep Belief Networks

Restricted Boltzmann Machines (RBMs) are a class of stochastic neural networks typically applied to unsupervised learning problems. The architecture consists of a layer of visible units, which correspond to the input data vector, and a layer of hidden units. The 'restriction' here, in reference to regular Boltzmann Machines, is that there are no visible-visible or hidden-hidden connections. The result of this is that the hidden units are mutually independent and can be sampled parallelly in an unbiased way. This allows the RBM to efficiently find

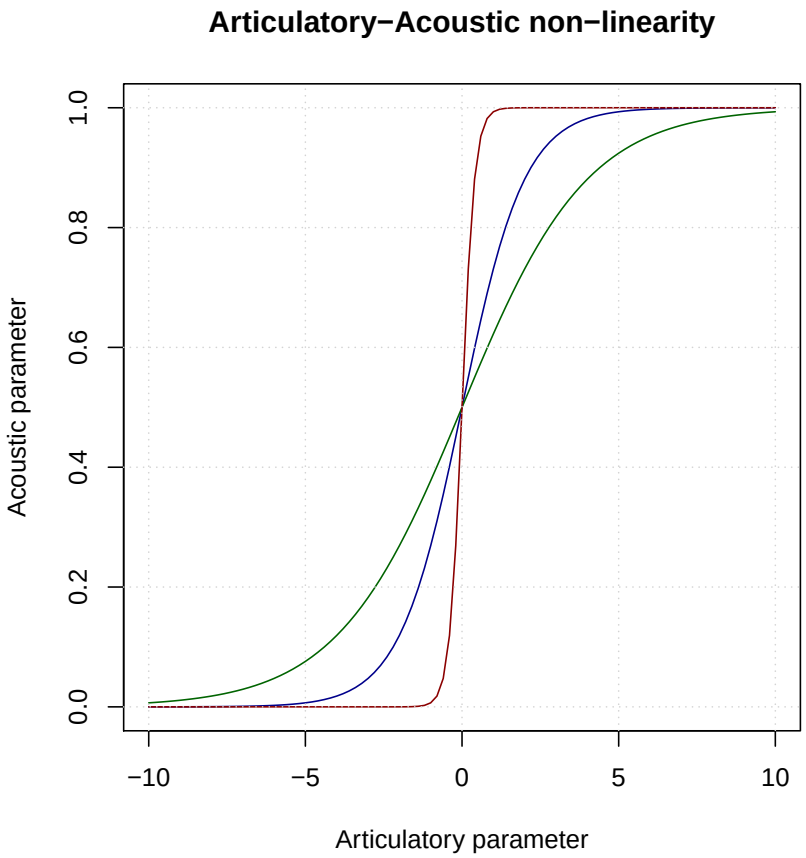
4 Modelling dynamic processes in speech processing

In this section we will cover a special use case of applying deep learning methods to an essentially non-linear set of problems, namely, modeling the dynamic nature of the articulatory-acoustic mapping (AAM). Since Stevens [1968], it has been known that the relationship between the articulatory and acoustic parameters is non-linear and has been explicated within the broad reach of *Quantal theory* [Stevens, 1968]

4.1 Acoustic-to-Articulatory inversion

Acoustic variation can be considered a direct consequence of changes in the articulatory parameter. While a given articulatory state will always have a

unitary and unique acoustic realization, a given acoustic signal could be the outcome of more than one articulatory states. This is exemplified in 4.1 below, where large changes in the articulatory parameter may not consequently produce different acoustic variations, however, every articulatory state has a unique acoustic consequence. The mapping between the articulatory and acoustic parameters, therefore, is highly non-linear in nature. A slight variation of the articulatory state may give rise to a whole different acoustic signal. This gives rise to a class of problems understood as the acoustic to articulatory inversion problem.



4.2 Acoustic-to-Articulatory mapping

Acoustic-to-Articulatory mapping refers to a special class of the speech inversion problem where simultaneous articulatory and acoustic data is used to learn a mapping between the parameters. This learning, due to the essential non-linear

relationship between the parameters, therefore, is also learning of various non-linearities. Recovering articulatory features from acoustics, therefore, remains a non-trivial problem. Toutios and Margaritis [2003] Badino et al. [2016], Canevari et al. [2013], Cernak et al. [2016]

References

Christopher D. Manning. Computational linguistics and deep learning. *Comput. Linguist.*, 41(4):701–707, December 2015. ISSN 0891-2017.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015. ISSN 0893-6080.

M. Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1): 8–30, Jan 1961. ISSN 0096-8390. doi: 10.1109/JRPROC.1961.287775.

Y. Bengio, R. Cardin, R. De Mori, and E. Merlo. Programmable execution of multi-layered networks for automatic speech recognition. *Commun. ACM*, 32 (2):195–199, February 1989a. ISSN 0001-0782.

Yoshua Bengio, Renato De Mori, and Regis Cardin. Speaker independent speech recognition with neural networks and speech knowledge. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, NIPS'89, pages 218–225, Cambridge, MA, USA, 1989b. MIT Press.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*, 2012.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, oct 1986. doi: 10.1038/323533a0. URL `https://doi.org/10.1038%2F323533a0`.

K.N. Stevens. *The Quantal Nature of Speech: Evidence from Articulatory-acoustic Data*. 1968.

Asterios Toutios and Konstantinos Margaritis. A rough guide to the acoustic-to-articulatory inversion of speech. In *6th Hellenic European Conference of Computer Mathematics and its Applications, HERCMA-2003*, 2003.

Leonardo Badino, Claudia Canevari, Luciano Fadiga, and Giorgio Metta. Integrating articulatory data in deep neural network-based acoustic modeling. *Comput. Speech Lang.*, 36(C):173–195, March 2016. ISSN 0885-2308.

Claudia Canevari, Leonardo Badino, Luciano Fadiga, and Giorgio Metta. Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping. In *INTERSPEECH*, 2013.

Milos Cernak, Afsaneh Asaei, and Hervé Bourlard. On structured sparsity of phonological posteriors for linguistic parsing. *Speech Communication*, 84:36 – 45, 2016. ISSN 0167-6393.