DEPARTMENT OF ELECTRONICS

BIOLOGICALLY INSPIRED COMPUTING

ELE00017M

TOP THREE USES OF NEURAL NETWORKS IN VOICE RESEARCH

Nathan Billis

Contents

1 Introduction

Voice research is currently a very important research area, especially with the rise of smart speakers such as the Amazon Echo [1], Google Nest [2] and Apple Homepod [3]. The importance of accurate voice recognition is not only important from an accessibility standpoint but also for people using these devices.

Speech has been an important way of communicating with each other and although speech recognition has moved a lot very quickly it's still not perfect. When working with speech there are a variety of different issues. Some of the most common issues include [4]:

- Speaker Variation: where the same words are pronounced differently by different people because of variations with gender, age, speed of speech and many other tiny changes.
- Background Noise: where the event is noisy it can add to the signal and even the speaker can add unwanted noise to the signal.
- Continuous Speech: when we speak there are occasions where there are no breaks between words making it difficult to disguise the start and ends of words.
- Other External Factors: in addition to the issues above there can be a variety of other factors such as the position of the microphone all having a knock-on effect on the quality of the recording.

This report will look at why Automatic Speech Recognition, Large Vocabulary Speech Recognition, and Noise Robust Speech Recognition are the top three applications of neural networks within voice research.

2 Neural Networks Overview

Neural networks (NN) are the most common form of biologically inspired computing in the real world. There are many types of Neural Networks each having their advantages and disadvantages. Many are treated as a black box style model with inputs on one side and outputs on the other, and they all generally require a lot of training to present good results.

Figure 1 displays some of the popular neural network types. Each network has differing strengths and because of this, they're adapted for their differing uses. The two main networks that are used in voice research are Feed Forward Network (FF) and Recurrent Neural Network (RNN), other types of networks are used however these two are the main one's we'll be focusing on. Neural Networks are made up of connected processes called neurons [6], each being activated in different ways with input nodes being activated by the environment. The complexity of the problem changes how many computational stages there needs to be in the network.

2.1 Feed Forward

When we combine many neurons we create a network and depending on the configuration changes the network type. The most basic is feed forward networks (FF). FF are arranged in layers [7], with each layer leading to the next layer and then onto the output. Unlike other forms of neural networks, there are no connections between layers. The connection between each layer has differing weights and bias changing the result of the output.

FF tend to "learn" from training and don't adapt after the initial training phase, the weights and bias which are created during the training phase remain for the duration of the existence of the network.

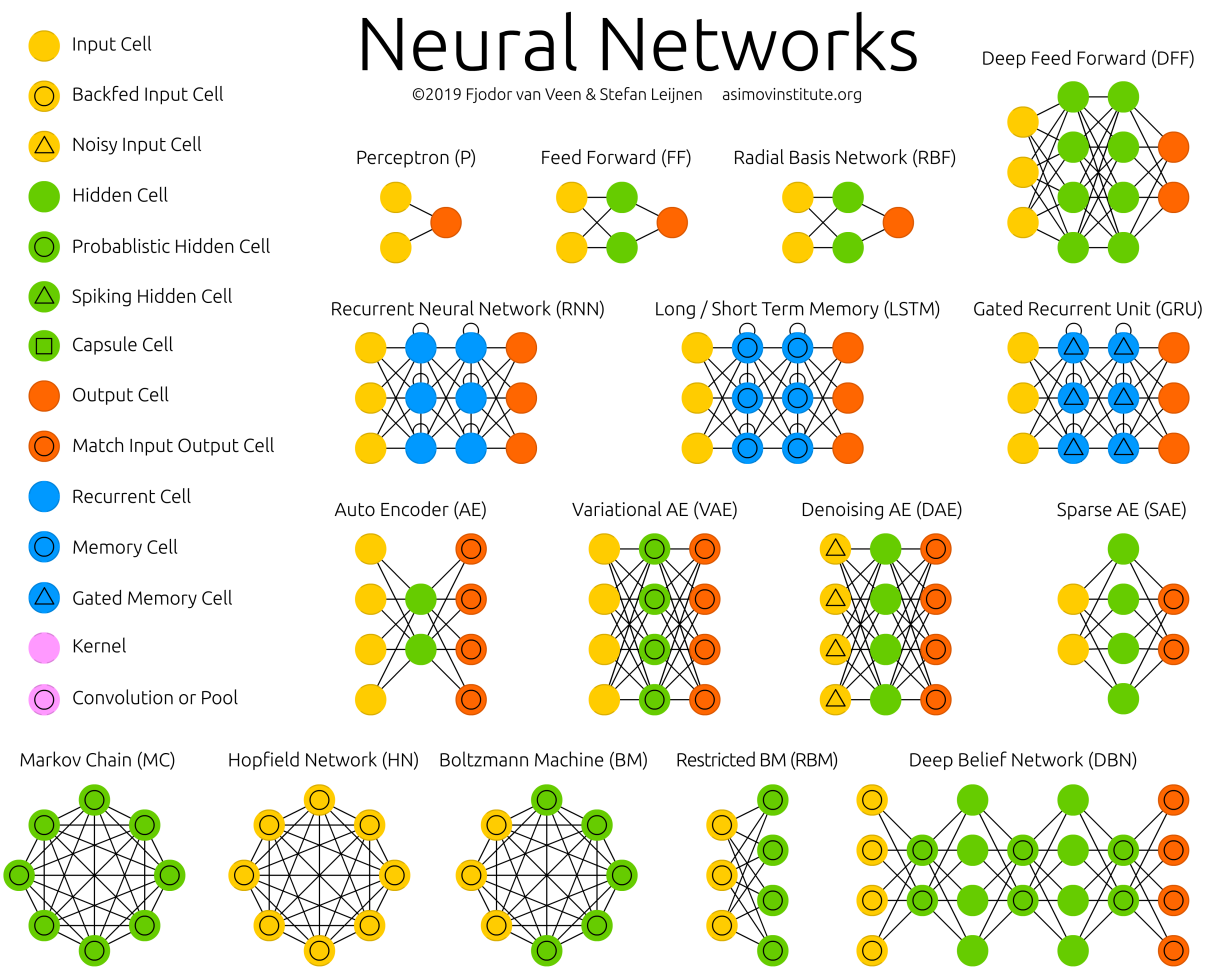

Figure 1: Main Types of Neural Networks [5]

These networks are used for supervised machine learning where we know what the outputs should be.

2.2 Recurrent Neural Networks

Recurrent neural networks (RNN) are similar to feed forward networks however unlike FF they learn from the past [8] and their decisions are influenced by that in addition to training. RNN are useful when the data needs to look at around the key item of interest. This is because RNN not only has access to their node but also those around it. This makes it useful for looking at things like the next word in a sentence because it's important to have the words that come before to use it in context.

Because the network operates in layers the order is important. When training these networks it is possible information may get lost over time, due to a vanishing or exploding gradient which is where the weights and bias are against the correct output causing data loss.

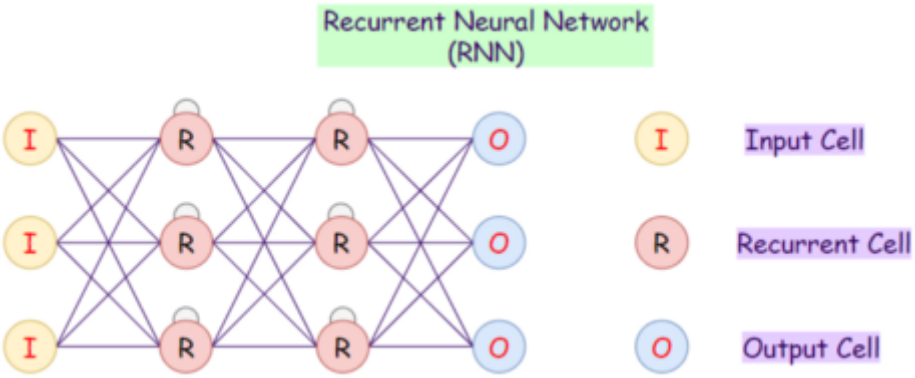

Figure 2: Recurrent Neural Network [9]

3 Automatic Speech Recognition (ASR)

One of the top uses of NN in voice research is within ASR. Before using neural networks in ASR, it relied heavily on statistical models such as hidden Markov models. Neural networks are already used in devices such as the Amazon Echo [10] and help to reduce the word error rate.

3.1 Phonemes and Voice Acoustics

Before looking at how ASR works it helps to have an understanding of Voice Acoustics. The best place to start is by first looking at phonemes, phonemes are a basic linguistic "unit" and every word is made up a series of phonemes, phonemes are not the same as syllables and if you change a phoneme you change the word. Phonemes are described from where they are produced in the mouth. The International Phonetic Alphabet is a system of phonetic notations that quantifies the speech that forms oral language.

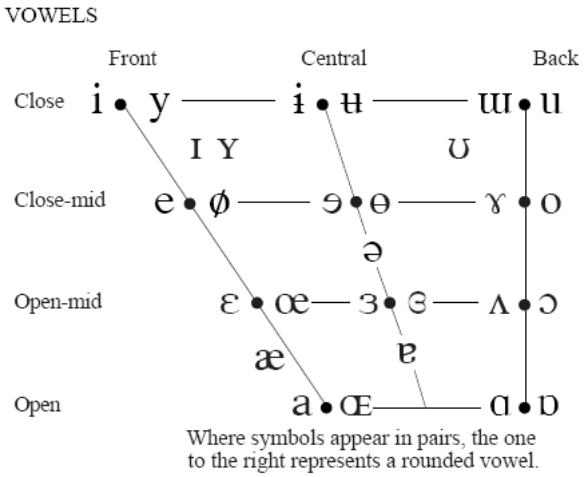

Figure 3: IPA: vowels [11]

Each phoneme has a unique spectral signature based on the format making it easier to identify. Where two vowels are next to each other that creates a diphthong which again has a slightly different spectral signature. The data from speech can be represented in several different ways depending on it's intended

use, this can be a Waveform, a Spectogram or Mel Frequency Cepstrum Coefficients.

The approximate formula to calculate mels for a frequency in Hz [4]:

$$mel(f) = 2595.^{10}log(1 + \frac{f}{700}) \tag{1}$$

When looking at feature extraction from speech Mel Frequency Cepstrum Coefficients analysis (MFCC) is a popular choice because it generally allows the same result irrespective of any time dependant problems. It has a much higher recognition success rate with used in combination with other methods of feature extraction such as using linear prediction coefficients or Power Spectral Density points [12]. Before the signal enters the neural network some pre-processing [13] is done to the signal to clean it up and get it into a format that will be easier understood by the network.

3.2 ASR and neutral networks

After gaining a basic understanding of the functions of phonemes in the speech chain it's now possible to look at Automatic Speech Recognition (ASR) and why phonemes are so important in its usage and why neural networks are used.

Feedforward Networks (FF) using Back Propagation algorithm for training is a popular network for ASR [14], RNN are also quite popular. FF neural networks need a constant input to ensure the output will be as expected because of this the all of the data from the full spectrograms is not used as an input and instead the MFCC, as discussed earlier, is used. Because the MFCC is used the dataset is a lot smaller meaning that the calculations and training are a lot quicker.

Depending on the number of words we want to recognise, the size of the network and number of output nodes will change. To estimate the number of neurons needed in the network Oja rule of thumb [15] is used:

$$H = \frac{T}{5(N + M)} \tag{2}$$

Where H is the number of hidden layer neurons, N is the input later size, M the output layer size, and T is the size of the training set. After we've created the network it can be trained using supervised training and then used to classify words. From this equation we see that with more words in the dictionary the bigger the network needs to become.

The shift from using statistical models or hybrid methods has allowed ASR to become more accurate and have a higher confidence level. Neural networks are very powerful at classifying speech and work well when paired with good pre-processing such as using Mel Frequency Cepstrum Coefficients over spectrograms.

4 Large Vocabulary Speech Recognition (LVSR)

Another use of neural networks in voice research is for large vocabulary speech recognition, when looking at ASR we discussed how having a large data-set meant that the neural network would be reasonably large, because of this different approaches have to be taken when looking at large vocabulary speech recognition.

Using hybrid networks with Neural Networks and Hidden Markov Models (HMM) it is possible to use data-sets much larger than traditional networks or models. Hybrid networks work by first training the

HMM in order to create training data to train the neural network with, once this data is created and the network has been trained and then fine tuned by back propagation, the outputs are then used for the HMM states. Using this combination approach it allows much bigger data-sets to be used and reduces the word error rate [16].

In addition to using hybrid models for LVSR, studies have been done looking at context-dependent networks for LVSR. These networks are also pre trained and they use a mixture of Deep Neural Networks (DNN) and HMM these hybrid context-dependent networks are simplified to CD-DNN-HMM. It works in a similar way to other hybrid networks but with the context dependent nature it improves sentence accuracy by over 5% [17]. A more cutting edge approach to LVSR is by using Recurrent Neural Networks that perform sentence prediction at character level. This works by replacing the HMM with a RNN, which learns the alignment between input and character sequence using an attention mechanism which is built into the network. The result of this work is to create simpler large vocabulary speech recognition system compared to HMM-DNN models [18]. This use of neural networks is really useful after getting the phonemes from words and helps to identify words either in context or out of context.

5 Noise Robust Speech Recognition

The final top use of neural networks in voice research is for use in noise robust speech recognition. When looking at noisy signals there are two main methods to look at the problem. The first is to attempt to clean up the signal, the second is to train the model to deal with the noisy signal. This area of research is large, not only looking at signals but also at data loss and reconstruction.

The majority of methods used are shared between conventional Gaussian model recogniser [19]:

- Using feature enhancement to remove issues before training.
- Training with multi-condition data.
- Adding a noise model to the network.

In addition to these another training method called drop out training is used which is more robust to unseen variables. Drop out training is used to help reduce the problem of deep neural networks form over fitting, this is where a large NN is trained to a smaller data set and try's to fit itself to the small number of data points. Drop out training works by randomly removing a certain percentage of neurons in the hidden later in each presentation of the sample during training sessions [20], this in turn increases the networks resilience for new or unknown inputs making it ideal for noisy signals.

Adding a noise model to the network or noise aware training is as simple as adapting the NN by in the training phase adding noise to the data set. Deep Neural network training also use feature enhancing algorithm prior to training to clean up the signal, this is similar to the pre-processing step in normal ASR but instead of producing a tradition MFCC the signal is processed again using the noise reduction algorithm [21].

6 Conclusion

The applications of neural networks are vast, in this report we focused on only those within voice research. Each of these uses are very important in increasing the quality and confidence when using smart voice controlled devices, and are all important in their own right.

Looking further afield not just at ASR, LVSR and Noise Robust Speech recognition there is also usage of neural networks in [22]:

- Low Resource speech recognition
- Multilingual speech recognition
- Speaker Adaptation
- Speech Separation

The uses of neural networks that are investigated in this report are important at the moment because there has been some recent research looking at identifying COVID-19 from coughing, which works in a similar way to ASR but instead of looking to identify phonemes in the neural network it identifies if the cough and checks if is similar to that of someone with coronavirus [23]. Neural networks improve the reliability and accuracy of automatic speech recognition, dealing with noisy signals well and identifying large vocabulary data sets, and they are leading the way to improving the world that we live in.

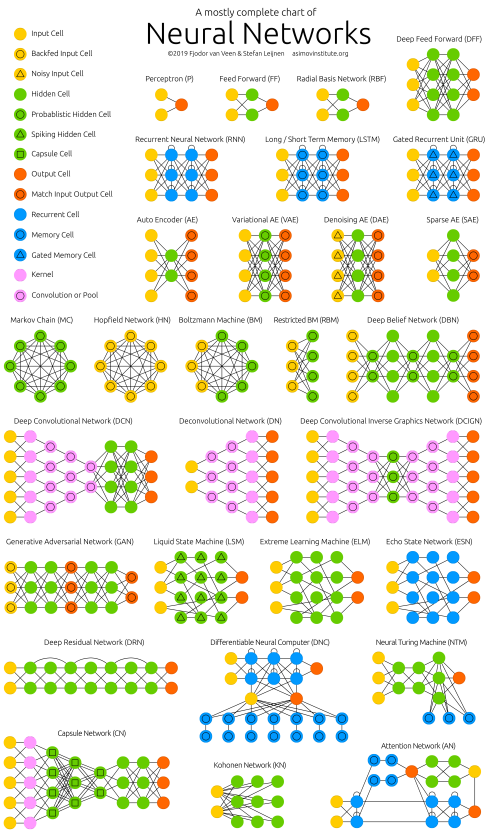

Figure 4: Mostly complete diagram of Neural Networks [24]

References

[1] *Introducing the All-New Echo Family—Reimagined, Inside and Out — Amazon.com, Inc. - Press Room*. URL: https://press.aboutamazon.com/news-releases/news-release-details/introducing-all-new-echo-family-reimagined-inside-and-out.

[2] *Google Nest smart speakers and displays – Google Store*. URL: https://store.google.com/gb/magazine/compare_nest_speakers_displays.

[3] *HomePod - Apple (UK)*. URL: https://www.apple.com/uk/homepod/.

[4] Wouter Gevaert, Georgi Tsenov, and Valeri Mladenov. "Neural networks used for speech recognition". In: *Journal of Automatic Control* 20.1 (2010), pp. 1–7. ISSN: 1450-9903. DOI: 10.2298/jac1001001g.

[5] Stefan Leijnen and Fjodor Van Veen. "The Neural Network Zoo". In: *Proceedings 2020, Vol. 47, Page 9* 47.1 (May 2020), p. 9. DOI: 10.3390/proceedings2020047009. URL: www.mdpi.com/journal/proceedings.

[6] Jürgen Schmidhuber. *Deep Learning in neural networks: An overview*. Jan. 2015. DOI: 10.1016/j.neunet.2014.09.003.

[7] *Neural Networks - Architecture*. URL: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Architecture/feedforward.html.

[8] *Recurrent Neural Networks. Remembering what's important — by Mahendran Venkatachalam — Towards Data Science*. URL: https://towardsdatascience.com/recurrent-neural-networks-d4642c9bc7ce.

[9] *Main Types of Neural Networks and its Applications– Tutorial - Towards AI*. URL: https://medium.com/towards-artificial-intelligence/main-types-of-neural-networks-and-its-applications-tutorial-734480d7ec8e.

[10] Prakhar Swarup, Roland Maas, Sri Garimella, et al. *Improving ASR confidence scores for Alexa using acoustic and hypothesis embeddings*. Tech. rep. URL: https://nlp.stanford.edu/projects/glove/..

[11] International Phonetic Association. *IPA: vowels — International Phonetic Association*. URL: https://www.internationalphoneticassociation.org/content/ipa-vowels.

[12] V. Moonasar and G. K. Venayagamoorthy. "A committee of neural networks for automatic speaker recognition (ASR) systems". In: *Proceedings of the International Joint Conference on Neural Networks*. Vol. 4. 2001, pp. 2936–2940. DOI: 10.1109/ijcnn.2001.938844.

[13] Fu-Hua Liu, Richard M. Stern, Xuedong Huang, et al. "Efficient cepstral normalization for robust speech recognition". In: *Proceedings of the workshop on Human Language Technology - HLT '93*. Morristown, NJ, USA: Association for Computational Linguistics (ACL), 1993, p. 69. DOI: 10.3115/1075671.1075688. URL: http://portal.acm.org/citation.cfm?doid=1075671.1075688.

[14] Lekshmi K.R and Sherly Elizabeth. "Automatic Speech Recognition using different Neural Network Architectures – A Survey". In: *International Journal of Computer Science and Information Technologies* 7(6) (Nov. 2016), pp. 2422–2427.

[15] Erkki Oja. "Simplified neuron model as a principal component analyzer". In: *Journal of Mathematical Biology* 15.3 (1982), pp. 267–273. ISSN: 14321416. DOI: 10.1007/BF00275687. URL: https://link.springer.com/article/10.1007/BF00275687.

[16] Navdeep Jaitly, Patrick Nguyen, Andrew Senior, et al. *Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition*. Tech. rep.

[17] George E. Dahl, Dong Yu, Li Deng, et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition". In: *IEEE Transactions on Audio, Speech and Language Processing* 20.1 (2012), pp. 30–42. ISSN: 15587916. DOI: 10.1109/TASL.2011.2134090.

[18] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, et al. "End-to-end attention-based large vocabulary speech recognition". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 2016-May. Institute of Electrical and Electronics Engineers Inc., May 2016, pp. 4945–4949. ISBN: 9781479999880. DOI: 10.1109/ICASSP.2016.7472618.

[19] Michael L. Seltzer, Dong Yu, and Yongqiang Wang. "An investigation of deep neural networks for noise robust speech recognition". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Oct. 2013, pp. 7398–7402. ISBN: 9781479903566. DOI: 10.1109/ICASSP.2013.6639100.

[20] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: (July 2012). URL: http://arxiv.org/abs/1207.0580.

[21] Dong Yu, Li Deng, Jasha Droppo, et al. *A MINIMUM-MEAN-SQUARE-ERROR NOISE REDUCTION ALGORITHM ON MEL-FREQUENCY CEPSTRA FOR ROBUST SPEECH RECOGNITION*. Tech. rep.

[22] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, et al. "Speech Recognition Using Deep Neural Networks: A Systematic Review". In: *IEEE Access* 7 (2019), pp. 19143–19165. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2896880.

[23] Jordi Laguarta, Ferran Hueto, and Brian Subirana. "COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings". In: *IEEE Open Journal of Engineering in Medicine and Biology* (Sept. 2020), pp. 1–1. DOI: 10.1109/ojemb.2020.3026928.

[24] *The Neural Network Zoo - The Asimov Institute*. URL: https://www.asimovinstitute.org/neural-network-zoo/.