Brenden Schmitz and Farisa Adi

Look who’s talking

The way both you and Siri speak might be more similar than you realise

Look who's Talking

Brenden Schmitz and Farisa Adi

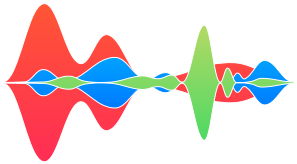
A man looks out his window at a cloudy sky. He pulls out his iPhone and asks Apple's virtual assistant Siri "Hey Siri, will I need an umbrella today?". Promptly, Siri voices a reply with "Yes, there is a forecast of rain in the afternoon." A simple response, but an incredible premise: An artificial intelligence is able to both understand a statement and respond not only with a relevant answer, but with its own distinct voice. So how does Siri do it? The answer comes in two parts: **Speech Synthesis**, the artificial production of human speech, and **Deep Learning**, the machine learning process that allows Artificial Intelligence (AI) to learn and respond to stimuli.

Speech Synthesis

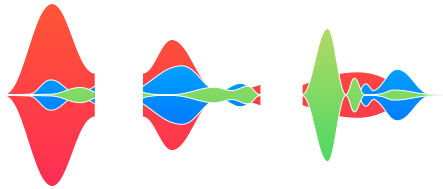
First comes Speech Synthesis. In order for an AI to speak in its own voice, it must use the voice of an actual person as a scaffold. Typically this involves a variety of speech recordings (between 10-20 hours worth) from a professional voice actor, including segments like responding to a joke or reading from a book. It is impossible to record every utterance of natural speech, so these recordings are instead segmented into individual **phones** (a distinct speech sound independent of contextual use) and reconstructed into new words.

This process might sound straightforward, but it is quite the opposite. After calculating a response, an AI must first understand the meaning behind it and determine its **prosody** (the rhythm and tune of speech that contribute to its meaning, such as asking a question or speaking sarcastically). Subsequently, it must find a sequence of phones that can be joined together without audible glitches that both matches the correct response to the query and sounds like an actual humans voice. This is where Deep Learning comes in.

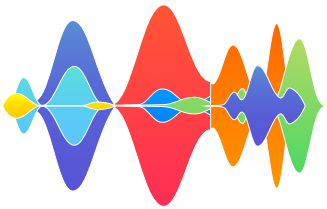
How Speech Synthesis Works



1. Recording. Speech samples are recorded and saved to a database.



2. Segmentation. Each sample is split into its phonetic components (phones).

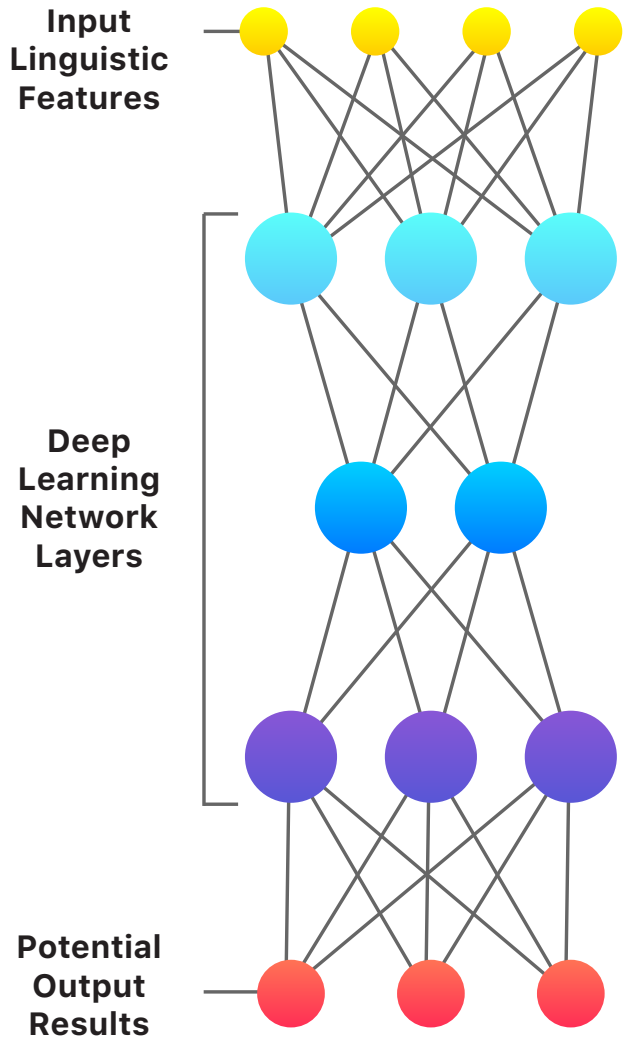


3. Reassembly. New words are created by stitching phonetic components together.

Deep Learning

Beginning in iOS 10, Apple has based the voice of Siri on deep learning to provide a more natural and smoother voice. Put simply, deep learning allows for an AI to learn using algorithms that loosely replicate the human brain and its ability to transfer data. This ability is called a neural network.

A neural network consists of thousands of **nodes** (an algorithm that computes a result from input data values) that are densely interconnected. When a network is activated, a node will receive data from each of its connections and calculates an output that is compared against a threshold value. If the output is above the threshold, it passes the data onto the next layer of nodes. This process repeats until a final output is achieved and compared to a desired result. Depending on how close the output is to the desired result, some nodes will be prioritized by the AI to produce a more accurate result.



Applying this to speech synthesis, Siri can train itself to more clearly understand the qualities of speech (such as the contextual meanings of pitch and duration of sounds) by learning to predict which nodes to prioritize depending on its input query. What this results in is faster and more coherent responses, as well as a more natural cadence that matches human speech. Siri can handle more advanced queries successfully such as making digital payments safely, translating articles of different languages, and personal tailoring the iOS experience to your individual needs. This innovation stands as another step in creating the perfect virtual assistant.

References:

Acero, A. (2017). ***Stanford Seminar - Deep Learning in Speech Recognition***. Retrieved 15 October, 2018, from https://www.youtube.com/watch?v=RBgfLvAOrss

Bengio, Y., Hinton, G., & Lecun, Y. (2015). Deep Learning. ***Nature, 521***, 436–444. doi: 10.1038/nature14539

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. ***ScienceDirect***. Retrieved 15 October, 2018, from https://arxiv.org/pdf/1404.7828.pdf

Shah, D. (2018). ***AI, Machine Learning, & Deep Learning Explained in 5 Minutes***. Medium. Retrieved 15 October, 2018, from https://becominghuman.ai/ai-machine-learning-deep-learning-explained-in-5-minutes-b88b6ee65846

Siri Team. (2017). ***Deep Learning for Siri's Voice: On-device Deep Mixture Density Networks for Hybrid Unit Selection Synthesis***. Machine Learning Journal. Retrieved 15 October, 2018, from https://machinelearning.apple.com/2017/08/06/siri-voices.html

How Siri Speaks

Text to Speech

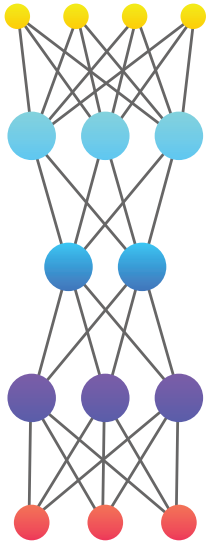
1. Text Output. After recieiving a information request, Siri generates an answer in text form.

Yes, there is a forecast of rain in the afternoon.

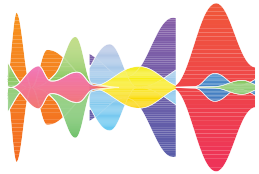
2. Conversion. Through linguistic analysis Siri converts the text into phonetic text.

jɛs, ðeər ɪz ə ˈfɔːkɑːst ɒv reɪn ɪn ði ˈɑːftəˈnuːn.

3. Calculation. The prosody of the sentence is determined through deep learning algorithms, which selects the correct rhythm and intonation by processing the statement though its node layers, which are pre-trained to think of a viable solution.



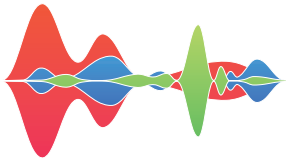
4. Unit Selection. Siri then chooses recorded phones from its data-base that matches the prosody determined by the deep learning algorithm and stiches them together.



5. Speaking. The process complete, Siri speaks the answer.

Audio Recording

1. Recording. A professional voice actor records speech of all different natures, such as news, jokes, poetry, and much more.



2. Segmentation. The recorded speech is divided up into its individual phones and saved to a database.