Elaine Lee & Madeleine Jackson

Hey Siri?

Are you really using the voice system Siri,
whilst understanding how it truly works?

Hey Siri?

Elaine Lee & Madeleine Jackson

Siri is a built-in, voice system platform; a personal assistant planted in all Apple devices for the users. Available in the IOS app it was originally developed on the 4th of October 2011, simultaneously changing the IT perspective of the world along with the introduction of iPhone. Siri's main objective as a personal assistant include: a fully detailed yet comfortable interaction between the user and the device through the verbal communication with Siri. The casual conversation allows the user to ask questions, favours (organisation of dates, calling, texting and even facetiming) or simply for fun, all hands-free.

Speech Synthesis

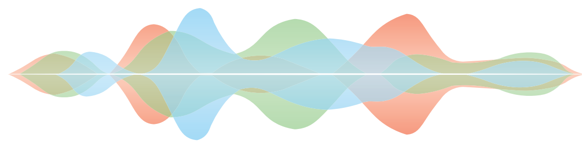
In order for AI to speak in its own voice, it must use the voice of an actual person as a scaffold. This is where **speech synthesis** comes in play. Typically, Apple has provided various recordings (roughly 10-20 hours' worth) from a professional actor, including witty jokes, and reading books.

It is impossible to record every utterance of natural speech, hence these recordings are instead segmented into individual **phones** (a distinct speech sound independent of contextual use) and reconstructed into new words through unit selection. Once the recordings are collected, unit selection begins on slicing the recorded speech into its elementary components, such as half-phones, and then recombining them according to the input text to create entirely new speech.

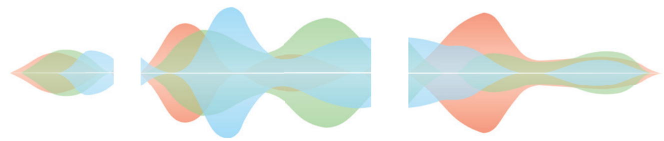
Deep Learning

Deep Learning, also referred to as deep machine learning is a subset of machine learning, an artificial intelligence developed by humans. Mostly known as an A.I in the contemporary society, deep learning consists of neural networks with the ability of inputting data's increasingly abstract representation. Thus, allowing the machine to mimic human behaviours in a very precise and detailed manner. Heavily

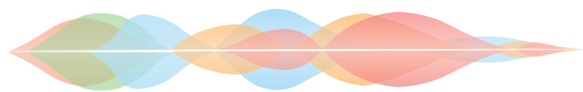
How Speech Synthesis Works



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



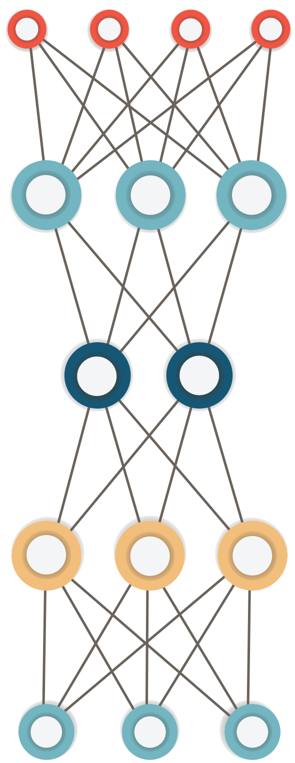
2. Segmentation. Each samples is split into its phonetic components (phones).



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

inspired by the structure of a human brain, the recorded set of data and multi-layered structure of algorithms within the network connects the computer system and the machine, thus producing many successful results, to great extents from self-driving cars, voice system to academic research.

To each of its incoming connections, a **node** will assign a number known as "**weight**." When a neural net is trained, all its weights and thresholds are initially set at random value. Training data is first passed to the layer (**input layer**) and flows through the next layers of nodes, multiplying and adding together in complex ways, before it inevitably arrives, completely transformed, at the **output layer**. During training, the weights and thresholds are constantly adjusted (i.e. passing the weight on the next layer of nodes if that weight is above the threshold value) until the training data with the same labels consistently produces remarkably similar outputs.



Overall, through the use of speech synthesis; allowing the AI to produce a human voices and deep learning; artificial intelligence developed by humans which are integrated into the platform, has successfully invented the voice system 'Siri'. This built-in technology/platform; a personal assistant in all Apple devices for the users, was designed with the contemporary technology which dramatically improved from the time period 2010. From then till now, it is still continuing to support the users of Apple products through the smart yet detailed interactions, all hands-free.

References:

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. ScienceDirect. Retrieved 28 September, 2020, from https://arxiv.org/pdf/1404.7828.pdf

Siri Team. (2017). Deep Learning for Siri's Voice: On-device Deep Mixture Density Networks for Hybrid Unit Selection Synthesis. Machine Learning Journal. Retrieved 28 September, 2020, from https://machinelearning.apple.com/2017/08/06/siri-voices.html

Magnimind. (2019, August 5). Deep learning structure guide for beginners. Medium. Retrieved from https://becominghuman.ai/deep-learning-structure-guide-for-beginners-9681cab342b6

Oppermann, A. (2020, August 11). What is deep learning and how does it work? Medium. Retrieved from https://towardsdatascience.com/what-is-deep-learning-and-how-does-it-work-2ce44bb692ac

Aäron van den Oord. (2016). WaveNet: A Generative Model for Raw Audio. Cornell University. Retrieved 4 October, 2020, from https://arxiv.org/pdf/1609.03499.pdf

How Siri Speaks

Text to Speech

1. Text Output. After receiving a information request, Siri generates a reply in a text form.

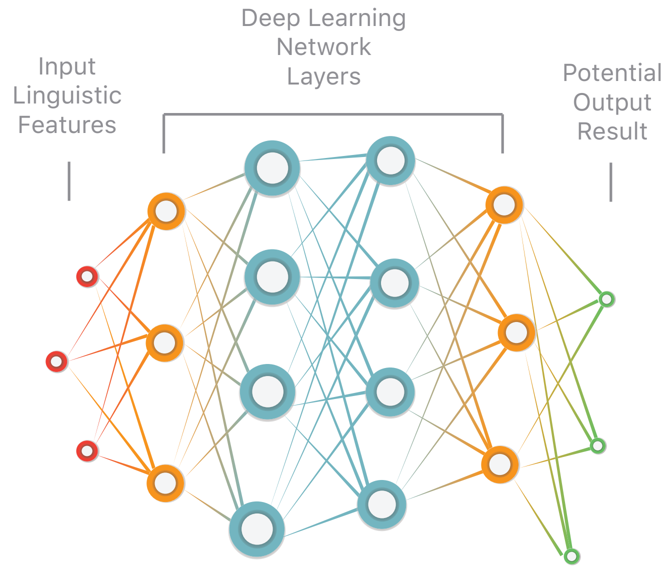
The time is
8:30 pm

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

ðə taɪm ɪz
8:30 pm

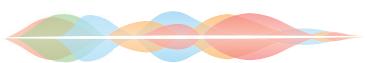
Deep Learning

3. Calculation. The prosody of the sentence is then calculated by **deep learning** algorithms, which choose the appropriate rhythm and intonation by decoding the sentence through its node layers, which are pre-trained to think of a practical solution.

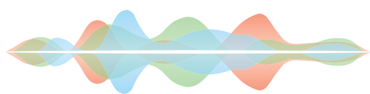


Speech Synthesis

4. Unit Selection. Siri then selects recorded phones from its database through slicing the recorded speech that fits the prosody defined by the deep learning algorithm and then stitching them together to develop a new speech.



4.1. Recording. Various recordings (approximately 10-20 hours) by a professional actor of all natures, including but not limited to: witty jokes, news, poetry, and more.



4.2. Segmentation. Each samples is split into its phonetic components (phones).

5. Speak. After the speech has been accomplished, Siri speaks out the practical solution.