Speech to text and text to speech

Siri works on primarily 2 technologies – Speech Recognition and Natural Language Processing. Speech Recognition is the task of converting a human speech into its corresponding textual form. For instance, when you trigger Siri by saying “Hey Siri”, in the back-end, a powerful speech recognition system by Apple kicks off and converts your audio into its corresponding textual form – “Hey Siri.” This is an extremely challenging task simply because we humans have a highly diverse set of tones as well as accents. The accents vary not only across countries, but also across states/cities within a country. Some people speak fast, some speak slowly. Characteristics of male and female voices are also very different.

The engineers at Apple train Machine Learning models on large, transcribed datasets in order to create efficient speech recognition models for Siri. These models are trained with highly diverse datasets that comprise of the voice samples of a large group of people. This way, Siri is able to cater to various accents.

In the recent years, deep learning has proven to produce phenomenal results in speech recognition. The word error rate of speech recognition engines has drastically gone down to less than 10%. This has been possible due to the availability of not only large datasets, but also powerful hardware using speech recognition algorithms that can be trained on the datasets.

Once Siri has understood what you are saying, the converted text is sent to Apple servers for further processing. Apple servers then run Natural Language Processing (NLP) algorithms on this text to understand the intent of what the user is trying to say. For instance, the NLP engines are able to differentiate that when a user is saying “set an alarm for 7AM tomorrow,” the user is asking about setting an alarm and not about making a call.

Overall, Siri is based on large-scale Machine Learning systems that employ 2 main aspects of data science – Speech Recognition and Natural Language Processing.

Amazing isn’t it? In our technology-heavy world, we tend to take things around us for granted. However, when we try to uncover them, we realize that there is a lot of technological magic happening behind the scene.

Speech synthesis

Speech synthesis—the artificial production of human speech—is widely used for various applications from assistive technology to gaming and entertainment. Ml models in apple server, translate text to sythesiszed speech