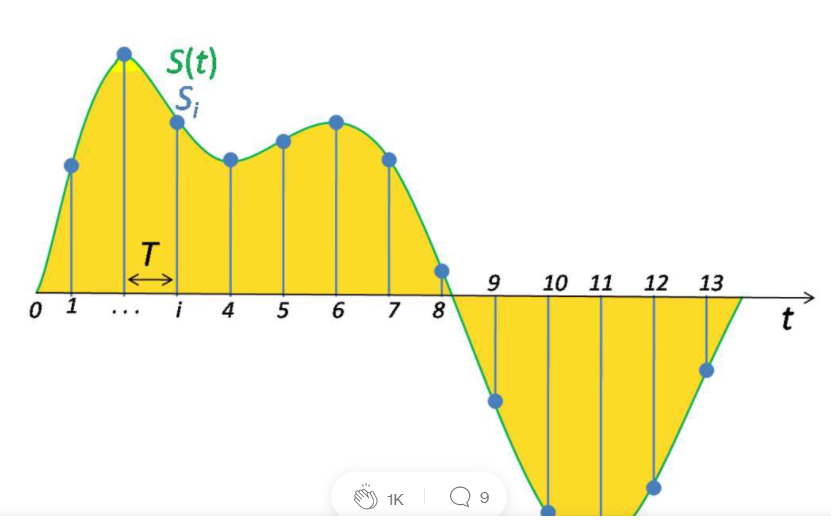
Audio Deep Learning

<https://towardsdatascience.com/audio-deep-learning-made-simple-part-1-state-of-the-art-techniques-da1d3dff2504>

A website gathering explanations and techniques about Deep Learning in audio processing.

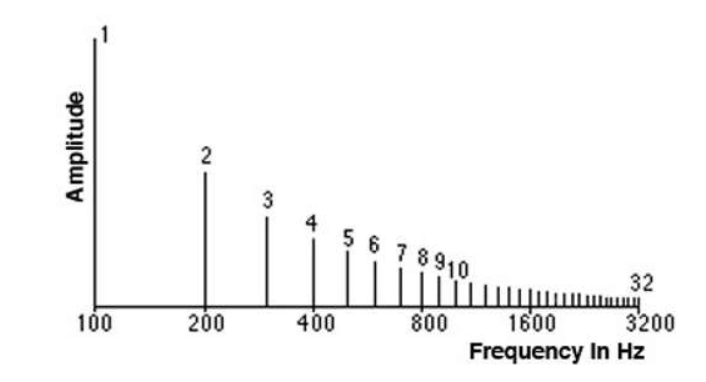
First, how to we represent the sound digitally? We represent it by turning it into a series of numbers that we can input in our model:



*Time discretization of a sound signal*

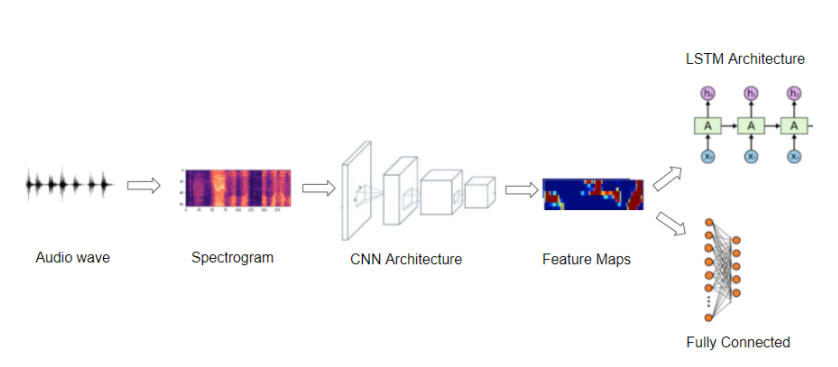
Recently, Deep Learning has been used a lot to process sounds as well. Some techniques have been developed. **Depending on the type of application we search for sound, we can extract different types of data that are relevant for our application, such as “phonemes”.**

To analyze the sound and do calculations on it, we first need to create a spectrum:



From the audio we have discretized. Spectrums are another way to see a specific signal, and are done using Fourier transforms.

**Most deep learning applications use the following steps:**



Raw audio waves are transformed into a spectrogram. This spectrogram is put through a CNN architecture, and is transformed into a feature map. Then, we apply the modifications we want on the feature map.

**What we are capable to do nowadays is: Music Generation & Music Transcription.** We are able to generate synthetic music that matches a particular genre, instrument, or a composer’s style.

(**Guide on how to process data using Python):**

[**https://towardsdatascience.com/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505**](https://towardsdatascience.com/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505)

**Creating a song converter (converting a speech into a singing speech) necessitates several steps:**

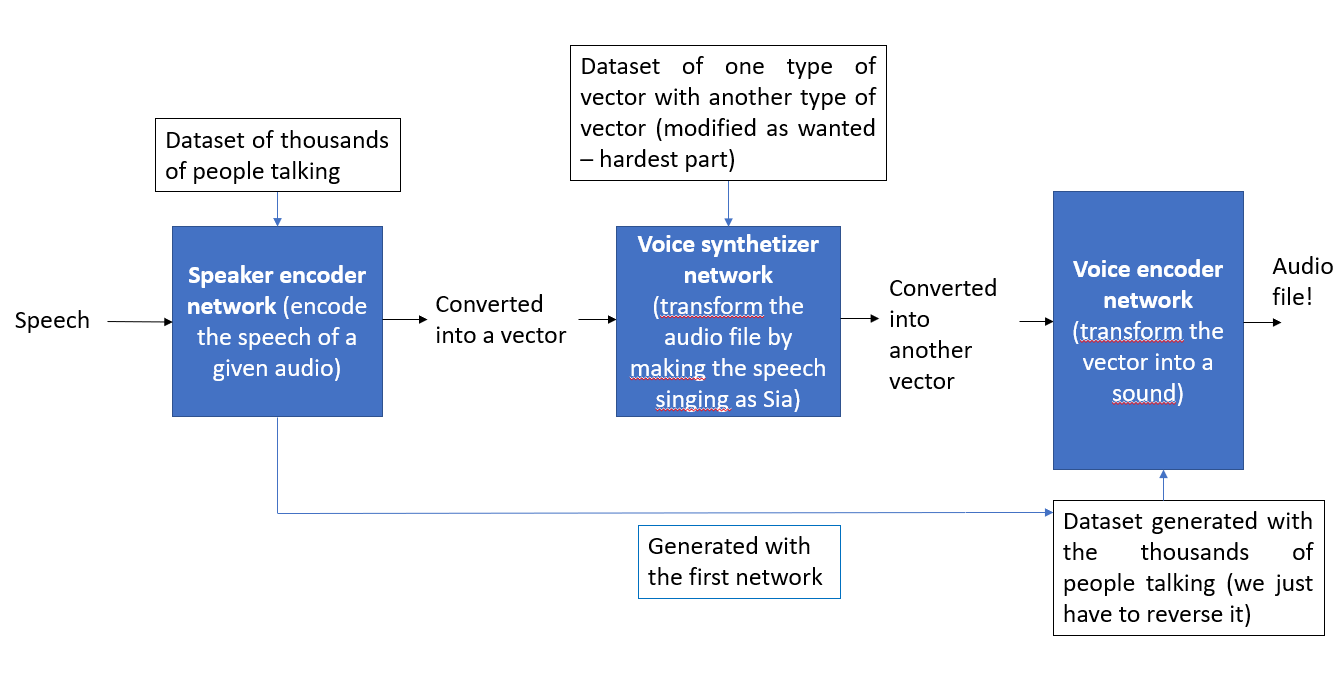
Basically, what we are going to do is called “Voice Cloning”. We could add instruments or at least a melody when the speech is given back!

<https://medium.com/the-research-nest/voice-cloning-using-deep-learning-166f1b8d8595> About voice cloning techniques.

<https://towardsdatascience.com/you-can-now-speak-using-someone-elses-voice-with-deep-learning-8be24368fa2b> Another article about our original idea

<https://github.com/CorentinJ/Real-Time-Voice-Cloning.git> Github of a voice cloning algorithm.

We could adopt a structure of multi networks that would communicate to each other and performs specific tasks.



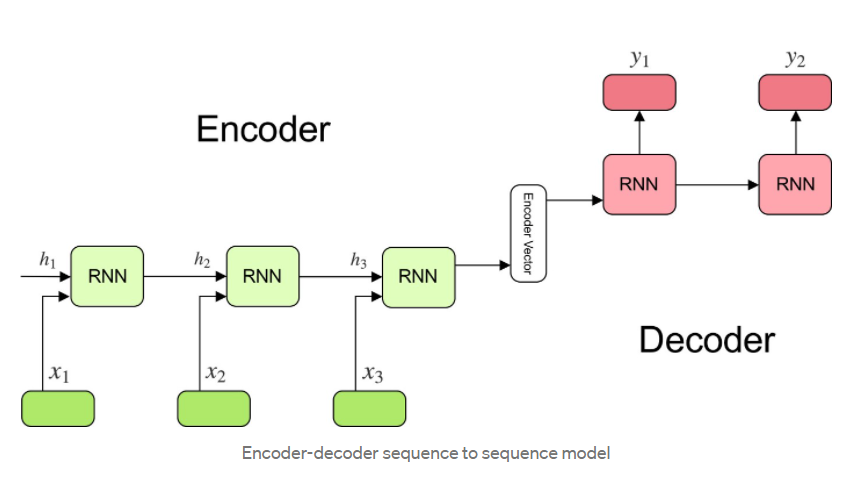
(Inspired by SV2TTS)

**Speaker encoder network:**

* Usually, involves the use of RNN: Recurrent Neural Networks. It has some issues, so models like LSTM or GRU are sometimes also used. **Usually, the voice encoder converts**

**Voice synthetized network:**

* Seq2Seq (created by Google), seems to be a fine choice. The Seq2Seq model is the following one:



**Voice encoder network:**

* RNN or (LSTM/GRU) network.