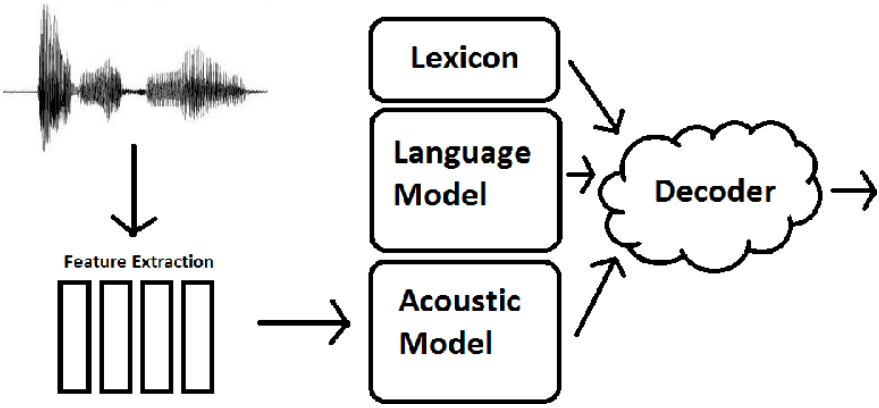
Automatic speech recognition with Vosk

Vosk is speech recognition software which translates speech into text, which is also called speech-to-text (or STT in short).

The software is made by Alpha Cephei whose *primary mission is scientific*. They release Vosk in 3 versions: Vosk open source, Vosk enterprise and Vosk mobile.

As automatic speech recognition (ASR) is one of the sub domains of the computational linguistics field, there is a whole range of software available to work with speech recognition, such as CMU Sphinx or Kaldi. Vosk is particulary made to be used in a "plug and play" manner. Other software might provide many more options and ways to tweak how speech is recognised, but usually they are also much more difficult in their use.

Encoding & Decoding



Automatic speech regonition is based on processes of **encoding** and **decoding**.

To do that, speech is **encoded** in multiple ways, which is a process that is also called **knowledge representation**. The output of this encoding process are **models**, such as: **acoustic models**, **language models**, **lexicons**, and **phonetic dictionaries**.

The speech recognition software uses these models to decode speech.

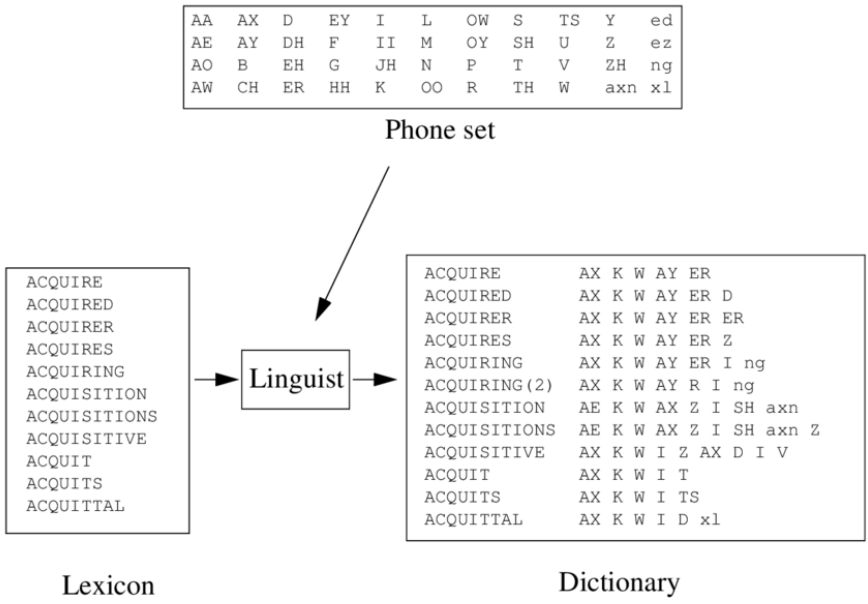
What is an **acoustic model**? (source, and nice clear step-by-step description)

> An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word.

What is a **language model**? (source)

> A Statistical Language Model is a file used by a Speech Recognition Engine to recognize speech. It contains a large list of words and their probability of occurrence.

Example of a **phonetic dictionary**: https://raw.githubusercontent.com/cmusphinx/cmudict/master/cmudict.dict (source)



How do the Vosk models decode speech?

From https://alphacephei.com/vosk/lm:

> The knowledge representation in speech recognition is an open question.
>
> Traditionally Vosk models compile the following data sources to build recognition graph:
>
> - Acoustic model - model of sounds of the language
> - Language model - model of word sequences
> - Phonetic dictionary - model of the decomposition on words to sounds

Using Vosk

Install Vosk

- you can install it using pip: `$ pip3 install vosk`
- use it as a python library: `import vosk`

Install Vosk in a virtual environment (recommended)

- navigate to the folder you work in: `$ cd ~/my/current/folder/`
- make a virtual environment: `$ python3 -m venv myvenv`
- activate your virtualenvironment: `$ source myvenv/bin/activate`
- now you should see that your prompt changed into something that looks like this: `$(myvenv)`
- install vosk: `$ pip3 install vosk`
- now vosk is installed in the folder: `myvenv/lib/python3.7/site-packages/vosk/`
- you can only use vosk now with your virtual environment activated. So when you run a python script that uses vosk, you need to run `$ python3 myscript.py` while you are in this virtual environment.

Download a model

Before you can try to run Vosk, you need to do download a model:

- choose a model: https://alphacephei.com/vosk/models
  - `NOTE:` check the filesize of the model, maybe you don't want to choose a big one if you use the soupboat

- download it to the soupboat (for example for the small EN-US model):
  - `$ cd the/folder/that/you/are/working/in/`
  - `$ wget https://alphacephei.com/vosk/models/vosk-model-small-en-us-0.15.zip`
  - `$ unzip vosk-model-small-en-us-0.15.zip`
- rename the folder into: `model`
  - `NOTE:` this is the folder that Vosk will look for; it is important that your notebook/script and "model" folder are in the same place

Connect your mic

`REMEMBER:` you need a microphone to use Vosk, and the Soupboat does not have one (yet?)!

So in order to work with Vosk, you need to work on a computer that has either a **build-in microphone**, or is connected to an **external microphone**.

Start from examples

There are multiple example scripts provided by the makers of Vosk, you can find them in their GitHub repository: https://github.com/alphacep/vosk-api/tree/master/python/example

Start from a customized example

See `speech-to-text.py` for a script that we prepared for you, that takes **speech as input** and **plain text as output**.

Run it with: `$ python3 speech-to-text.py`

In [ ]: