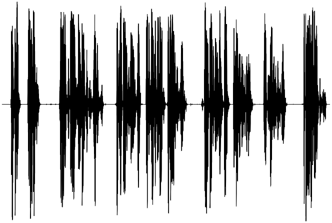
How to create a labeled dataset for Speech recognition?

Sudheer Kolachina

July 29, 2020

AUDIO



TRANSCRIPT

Chapter 1

Robin of Locksley

Robin Hood was born in the small town of Locksley in Nottighamshire. Locksley's buildings were made of wood and it was smoky and dirty, like many other towns in England ...

When transcript is approximate, how can you train an ASR model?

FORCED ALIGNMENT refers to the process by which transcriptions are aligned to audio recordings to automatically generate phone level segmentation.

- Forced alignment can be based on
    - Monophones
    - Diphones
    - Triphones
    - Words
    - Syllables
    - Phonetic classes like vowels, stop, sonorants, etc.
- Many different techniques to do forced alignment

- Kaldi-based aligners
    - Montreal Forced Aligner
    - Gentle Aligner
        - Read audio in chunks
        - Transcribe audio chunk using an ASR model
        - Estimate alignment between transcript and transcript of chunk extracted using the ASR model using Dynamic Time Warping

Useful Links

https://montreal-forced-aligner.readthedocs.io/en/latest/introduction.html
http://www.phon.ox.ac.uk/jcoleman/BAAP_ASR.pdf
https://cmusphinx.github.io/wiki/longaudioalignment/
https://github.com/pettarin/forced-alignment-tools