Orange
Innovation

Data preparation for Automatic Speech Recognition
with Kaldi

Elodie GAUTHIER (Orange Innovation)

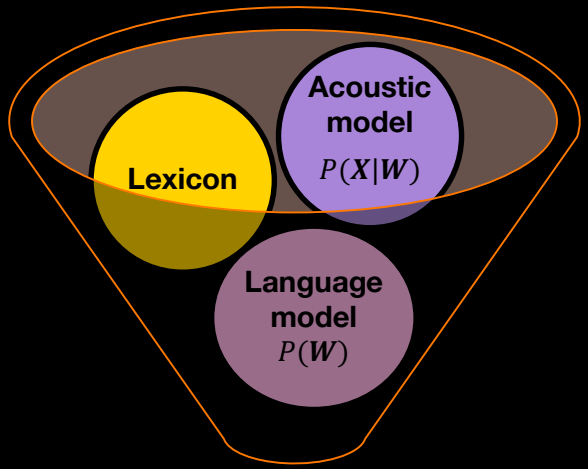
orange™

ASR focus:

1. Conventional ASR pipeline
2. Training phase
3. Evaluation phase
4. Datasets size & collection

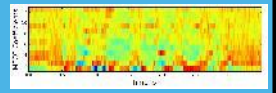
ASR focus 1/4: Conventional ASR pipeline

Conventional ASR pipeline



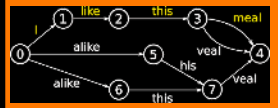
Input speech

Features Extraction



Decoder

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \; p(\mathbf{X}|\mathbf{W}) \; P(\mathbf{W})$$



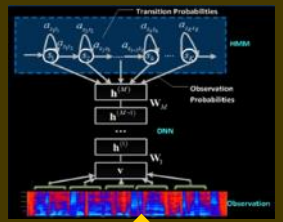
Output words

« I like this meal »

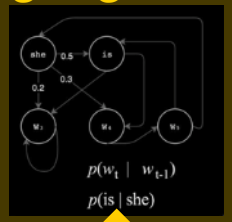
ASR: data for modeling

MODELS

Acoustic model



Language model



RESOURCES

Speech dataset +

- hello there
- the weather is cloudy today
- Barack Obama served as the 44th President of the United State

Pronunciation lexicon

#	SIL
<UNK>	SPN
<aaah>	SPN
<mmm>	SPN
eight	EY T
ate	EY T
data	D EY T AH
data	D AE T AH

Textual dataset

what is necessary to the completeness of the story at this stage is not to recapitulate but to take up some of the loose ends of threads woven in and follow them through until the clear and comprehensive picture of events can be seen
the way of the inventor is hard
he can sometimes raise capital to help him in working out his crude conceptions but even then it is frequently done at a distressful cost of personal surrender when the result is achieved the invention makes its appeal on the score of economy of material or of effort and then labor often awaits with crushing and tyrannical spirit to smash the apparatus or forbid its very use possibly our national optimism as revealed in invention the seeking a higher good needs some check

ASR focus 2/4: Training phase

Train an ASR: which data?

Train an ASR system consists in making the system learn the orthographic transcription of a speech stream.

Training materials : 3 elements

1

Speech dataset

Set of audio files along with the corresponding orthographic transcription

- hello there
- the weather is cloudy today
- Barack Obama served as the 44th President of the United State

2

Pronunciation lexicon

File containing a word followed by its phonetic transcription (machine-readable phonetic alphabets exist)

#	SIL
<UNK>	SPN
<aaah>	SPN
<mmm>	SPN
eight	EY T
ate	EY T
data	D EY T AH
data	D AE T AH

3

Textual dataset

Set of contemporary texts containing common orthographic words sequence

what is necessary to the completeness of the story at this stage is not to recapitulate but to take up some of the loose ends of threads woven in and follow them through until the clear and comprehensive picture of events can be seen
the way of the inventor is hard
he can sometimes raise capital to help him in working out his crude conceptions but even then it is frequently done at a distressful cost of personal surrender
when the result is achieved the invention makes its appeal on the score of economy of material or of effort and then labor often awaits with crushing and tyrannical spirit to smash the apparatus or forbid its very use
possibly our national optimism as revealed in invention the seeking a higher good needs some check
possibly the leaders would travel too fast and too far on the road to perfection if conservatism did not also play its salutary part in insisting that the procession move forward as a whole
on the contrary the conditions for its acceptance had been ripening fast yet the very vogue of the electric arc light made harder the arrival of the incandescent
a number of parent arc lighting companies were in existence and a great many local companies had been called into being under franchises for commercial business and to execute regular city contracts for street lighting
thus in a curious manner the modern art of electric lighting was in a very true sense divided against itself with intense rivalries and jealousies which were none the less real because they were but temporary and occurred in a field where ultimate union of forces was inevitable
hence twenty years after the first edison stations were established the methods they involved could be fairly credited with no less than sixty seven per cent
it will be readily understood that under these conditions the modern lighting company supplies to its customers both incandescent and arc lighting frequently from the same dynamo electric machinery as a source of current and that the old feud as between the rival systems has died out

ASR focus 3/4: Evaluation phase

Evaluate an ASR: which data?

Evaluate an ASR system consists in testing the ability of the system to correctly transcribe the speech stream.

Evaluation materials : 4 elements

1

Speech dataset

- do Melbourne trains run all night
- it could not be caused by uranium alone
- light travels at 300,000 kilometers per second

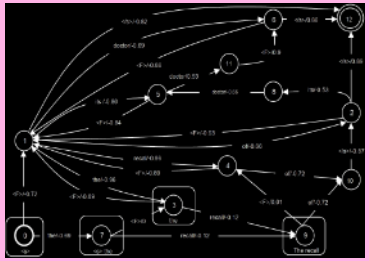
2

Pronunciation lexicon

#	SIL
<UNK>	SPN
<aaah>	SPN
<mmm>	SPN
eight	EY T
ate	EY T
data	D EY T AH
data	D AE T AH

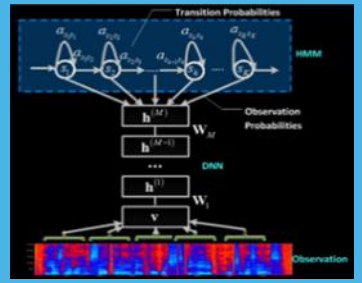
3

Language model



4

Acoustic model



Evaluate an ASR: key metrics

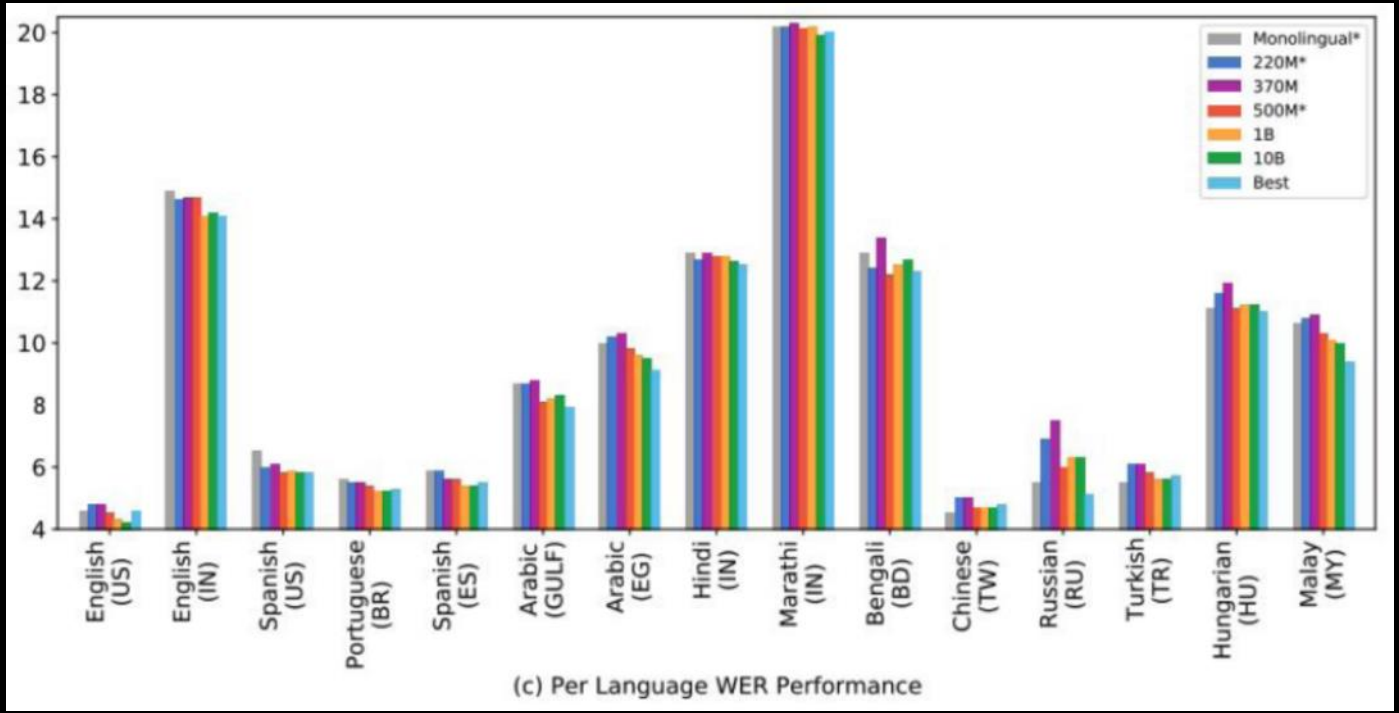
Objective tests:

- **Word Error Rate (WER) – achieve 20% of WER is a good start**

Human transcription	le	chat	boit	du	lait
Automatic transcription	le	chien	aboie	***	***
Error type	/	substitution	substitution	omission	omission
Error? (yes = 1 ; no = 0)	0	1	1	1	1

$$WER = \frac{\text{error count in the automatic transcription}}{\text{word count in the human transcription}} = \frac{4}{5} = 0{,}8 = 80\ \%$$

Google state-of-the-art ASR performance 2021



ASR: which tools?

Deep neural networks (DNNs) show best results nowadays.

For some kind of architectures (and a reduced training time), the use of graphic cards (GPUs) might be mandatory.

Kaldi is a widely used toolkit to train an ASR system.

Kaldi proposes recipes that guide the user from data preparation to model evaluation[1].

Kaldi runs on Linux OS (installation scripts for Windows OS exist but not recommanded).

1. see http://kaldi-asr.org/doc/

ASR focus 4/4: Dataset size & collection

Low-resourced languages: definition and issues

- Most of today's NLP research focuses on only 20 languages
- Over the 7,000+ world spoken languages, this is leaving the vast majority of languages un(der)studied.

Low-resource setting means:

Data scarcity

Unstable spelling

Limited electronic documents

Limited presence on the Web

None or few computerization

Dataset size recommandations

when working with low-resourced languages

Speech dataset	Textual dataset	Pronunciation lexicon
20 to 50 hours of speech, gender and age balanced, with various spoken style, accents, language level	Utterances composed of **10 millions of words at least,** containing a large contemporary vocabulary, wide range of topics (society, health, economy, politics, sports, etc.)	**15 000 entries at least,** covering all the sounds that exists in the language

How to collect?

Option 1: Buy the data

From a catalog

- ✓ Quick to get
- ✓ Quick launch
- ✗ Cost
- ✗ Data checking time

On demand

- ✓ Custom size
- ✓ Data collection process control
- ✓ Speech or text data already exists
- ✗ Contractualization time
- ✗ Speaker or transcriber training time
- ✗ Speech recordings or manual transcription planning

Option 2: Collect the data

By your own

- ✓ Semi-auto collection
- ✓ Data collection process control
- ✗ Data collection time
- ✗ Transcriber training time
- ✗ Manual transcription of recordings time

By contracting with a partner

- ✓ Partner skills & knowledge
- ✓ Partner connections
- ✗ Data collection time
- ✗ Transcriber training time
- ✗ Manual transcription of recordings time

Checking, cleaning and pre-processing

- Audio data
  - Listen to a sample of audio files with an audio player
    - Split audio files if long pauses are found
    - Put appart files if to much noise is found
    - Segment audio into speaker if multispeakers in recordings
  - Sample audio files to 16kHz, 16bit
  - Convert to WAVE format

- Textual data
  - Check encoding of text files (UTF-8 preferred with Unicode character set)
  - Convert words to lowercase to avoid ambiguities
  - Keep diacritics (accent, like « é » in French) on letters if any
    - e.g: in French, « email » and « émail » have different meaning

Thank you for your attention.

We are ready to prepare data for Kaldi !

Get the Notebook here to play with me:

https://github.com/gauthelo/contribuling2022-kaldi-workshop