KU LEUVEN

Resources for Flemish ASR: what is needed ?

ELG workshop on resources for Luxemburgish and Flemish

Hugo Van hamme
KU Leuven, PSI, Dept ESAT
Leuven, 8 July 2021

Progress in ASR technology

Enabling factors

- Algorithms: GMM to DNN
- Computational resources
  - x10000 since 1993
- **Data in training**

2

DL revolution: data

Example from a research paper Google (English)

- Arun Narayanan, et al., "Recognizing Long-Form Speech Using Streaming End-to-End Models", 2019
- *The training sets include data from four domains: anonymized and hand-transcribed utterances representative of …*

Application Domain	Total (hours)	Mean (sec.)	Median (sec.)
Search	56k	6.2	4.8
Farfield	38k	3.9	3.5
Telephony	4k	4.4	3.0
YouTube	190k	5.9	4.5
Total	288k		

- Multi-domain corpus
- Compare to CGN:
  - 0.25kh Flemish and 0.5kh Dutch
- Need MUCH more data

3

More data for Flemish ASR - cause

Bottom-up demand

- Flemish/Dutch is lagging behind compared to English
  - E.g. captions in Google Meet
- Wait for FAGMA ?
  - cfr. vaccination …
  - Relevant for digital economy, services, inclusion, …
- Interest from Flemish industry – cfr. meetings organised by EWI
  - Ambition to make a 10kh annotated corpus @ 1 mio EURO

4

Where to get data for a large CGN ?

Scale up from CGN

	Scaling	Legal	Coverage
Call centers	☺	☹	☺
Parliament, city council	☺	☺	☺
Talkshows	☺	😐	😐
Soaps and movies	☺	😐	☹
Audio books	☺	😐	☹
Lecture recordings	😐	☹	😐
Crowd sourcing	😐	☺	☺
Face-to-face meetings	☹	☺	☺

- Coverage: dialects, (non-natives), age

5

Legal constraints

Life is more complicated today

6

GDPR: restrictions

- Informed consent not trivial:
  - Call centres …
  - Talk shows, lectures, ...
- Pseudonyms not trivial for voice data
- Limitations on use
- Limitations in time
- Geographical limitations
- Recall contributions

7

Author rights

- Creative content (soap, movie, play, …)
- Actors, commedians, moderators

8

Should we join forces with The Netherlands ?

How different are we really ?

How different are we ?

Use of Factorized Hierarchical Variational Auto Encoder

- Unsupervised learning of speaker and speech representations
- Trained on CGN
- Single-layer classifier
- Tested on segments of 20s

- Seem to be very different acoustically !

α_b	α_c	**Dialect acc**	**Gender acc**
0	0	0.980	0.993
10	10	0.946	0.997
10	0	0.872	0.980
10	10	0.946	0.997
10	100	0.999	0.994
0	10	0.910	0.955
10	10	0.946	0.997
100	10	0.843	0.993

Master thesis by Robrecht Meersman, 2019

Should we team up with The Netherlands ?

Test with CGN corpus – E2E technology

Train	Test VL – WER (%)	Test NL – WER (%)
VL	22.3	41.4
NL	34.0	27.8
VL + NL	20.4	26.5

Conclusion

- It does help … a little

Thanks to Steven Vander Eeckt, 2020

More, cheaper, faster, …

Do we need annotation ?

12

Doing away with labels
Principles

- Unsupervised training
  - Mask out present speech and predict it from past and future
  - No labels required
  - Other options
- Self-supervised training
  - Label your training set with your ASR system
  - Then train ASR system again

Speech representation learning: amount of data
Using unlabeled Flemish data

10h	30h	50h	100h	150h	250h	350h	500h	700h
31.87	20.85	16.76	15.55	15.74	14.76	14.34	14.73	13.52

(a) Unlabelled data for pre-training a base wav2vec 2.0 model (no finetuning), large ASR DNN.

Thanks to Jakob Poncelet, 2021

Baselines:
- MFCC 15.10% WER (standard Kaldi – hybrid DNN/HMM)
- Using 4.5k Dutch parliament data instead of Flemish: 16.32% WER

Conclusion:
- More data beyond 700h likely to help
- Language needs to match
- Analysis is limited to small wav2vec model
- **Ambition should still be thousands of hours**

Speech representation learning: finetuning

Using **labeled** Flemish data

0h	1h	10h	20h	30h	50h	90h	150h	250h
27.75	13.84	12.08	11.32	11.19	10.71	10.61	10.53	10.50

(b) Labelled data for finetuning XLSR-53, small ASR DNN.

Thanks to Jakob Poncelet, 2021

Trained unsupervisedly on 56k hours of 53 languages, incl. 1.6kh Dutch

Baselines:
- MFCC 15.10% WER (standard Kaldi)
- Using small unsupervised model on Flemish data:
  - No fine-tuning 13.52% WER
  - With fine-tuning: 11.76% WER

Conclusion:
- More unsupervised data helps
- Fine-tuning effective, saturates at 150h
- Domain/language/dialect/age match/transfer in pre-training and fine-tuning ?

15

Conclusion - the plan

Two tracks

16

Ambition based on science

Data needs for Flemish

- Thousands of hours
- Domain coverage
  - Meetings (parliament, city council, lectures, …)
  - Read speech and interviews (news broadcast, audiobooks, …)
  - Spontaneous (soaps, …)
  - Call centres and voice bots
- Dialect/age coverage

17

Track 1: Unsupervised and self-supervised approaches

More, cheaper, scalable

- Research project
  - weakly supervised learning
  - self-supervised learning
  - unsupervised learning
- GDPR and author rights:
  - usage for research is less restricted
  - data not published
- Serve industry and government through derivatives
  - ASR models for open source toolkits

18

Track 2: Public data

- More future-proof
- Data published by owner
  - Can specify legal constraints
  - If further clearance needed: direct link between user and provider
  - Universities help in formatting and (automatic/manual) annotation
- AI community will work on it

19

Thanks you for your attention

Thanks to Robrecht Meersman, Steven Vander Eeckt & Jakob Poncelet

Questions ?

20