Interspeech 2021

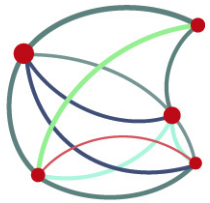
Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data Filtering/Verbatimization

http://doi.org/10.21437/Interspeech.2021-1905

Gonçal V. Garcés Díaz-Munío, Joan Albert Silvestre-Cerdà, Javier Jorge, Adrià Giménez, Javier Iranzo-Sánchez, Pau Baquero-Arnal, Nahuel Roselló, Alejandro Pérez-González-de-Martos, Jorge Civera, Albert Sanchis and Alfons Juan

{gogardia,jsilvestre,jajorca,adgipas,jairsan,pabaar,narobe1, alpegon2,jorcisai,josanna2,ajuanci}@vrain.upv.es

www.mllp.upv.es



MLLP | Machine Learning and Language Processing

VRAIN

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

1 Motivation

- Quality ASR requires **thousands of hours of speech data**.
- Frequently, **existing transcriptions** are **not 100% verbatim**.
  - Most large public corpora are non-verbatim.
- We need techniques to **make the most of speech data**.
  - Filter out inaccurately transcribed parts.
  - What about **auto-improving** inaccurate transcriptions?

2 The Europarl-ASR corpus

- A new large speech corpus for:
  - Training & benchmarking (streaming) ASR.
  - Benchmarking speech data filtering and verbatimization.
- **1300 hours** of EN transcribed speech data.
  - 3 full sets of timed transcriptions:
    official non-verbatim; auto noise-filtered; auto verbatimized.
- Dev/test: 18 hours of **manually revised** transcriptions.
  - 2 transcription sets: official non-verbatim; revised verbatim.
  - 2 independent dev/test partitions for 2 realistic ASR tasks.

https://www.mllp.upv.es/europarl-asr

MLLP

3 Data gathering and selection

- European Parliament recordings, transcriptions and translations.
  - Focus on English (v1.0).
  - 1340 raw hours of EN transcribed speech (non-verbatim!).
  - Auto-segmentation at the speech level.

- **Data selection**:
  - Filter out noisiest data (wrong times, unrelated/empty texts...).
  - At the speech level, based on character error rate (CER) of official transcription vs automatic transcription.
  - Selected data: 1263 hours (94%) (non-verbatim!).

4 Tasks, data partition, manual revision

- 2 realistic streaming ASR tasks:
  - Speaker-dependent (*MEP*): prior knowledge about MEPs.
  - Speaker-independent (*Guest*): no prior knowledge of guests.

Set →	MEP-dev	MEP-test	Guest-dev	Guest-test	train
Speakers	21	21	6	6	1034
Length	4.6h	4.7h	4.3h	3.9h	1230h

- **Manual revision** of dev/test: verbatim & non-verbatim texts.
- Text data for language modelling (in-domain):

Data source	Tokens
EP training set speeches	10M
EP speeches without audio	6M
EP translations into English	42M
Europarl-v10 EN no overlap	11M
DCEP English	104M
Total	173M

5 Speech data filtering and verbatimization

- **Goal**: filter/verbatimize training data to improve ASR results.
- Speech data **filtering**:
  - Force-align audio and transcription. Accept/reject at word level based on phoneme duration and alignment score.
  - 33% of the speech data is filtered out.
- Speech data **verbatimization**:
  - For each speech, automatic transcription with an LM based on the speech's non-verbatim transcription.
  - No speech data is lost.

Data set	Segments	Duration (h)
raw	1.13M	1007
filtered	1.24M	672
verbatimized	1.03M	1054

6 Europarl-ASR baselines

- Strong baseline ASR results, offline and streaming (lat. 0.65 s)

Training set	Word error rate			
	MEP-test (Speaker-dependent)		Guest-test (Speaker-independent)	
	Offline	Streaming	Offline	Streaming
raw	8.6	8.8	7.6	7.8
filtered	**7.8**	**7.9**	7.4	7.5
verbatimized	8.2	8.3	**7.0**	**7.3**

7 Conclusions

- Download the Europarl-ASR corpus:
  https://www.mllp.upv.es/europarl-asr
  - Goals: Assessing streaming ASR, filtering & verbatimization.
  - 1300h EN speech data (non-verbatim, filtered, verbatimized).
  - 18h dev/test, manually revised (verbatim, non-verbatim).

- Strong ASR baseline WERs, offline and streaming (lat. 0.65 s):
  - MEP task (speaker-dependent): offline 7.8, streaming 7.9.
  - Guest task (speaker-independent): offline 7.0, streaming 7.3.

- Speech data filtering & verbatimization improved WER by 9%.

Annex
Comparison of Europarl corpora

	Europarl-ST	VoxPopuli	Europarl-ASR
Release date	2020	2021	2021
Focus	Speech Translation	Speech-to-speech, unsupervised	ASR & filtering
EN transcribed speech (h)	186	543	1263
dev/test sets	Non-verbatim	Non-verbatim	Verbatim & non-verbatim

- Other speech or text corpora
  - Europarl (2001): Parallel text corpus. English 54M tokens.
  - GigaSpeech (2021): ASR speech corpus. English 10K hours.