LHCP-ASR: An English Speech Corpus of High-Energy Particle Physics Talks for Narrow-Domain ASR Benchmarking

Jaume Santamaría-Jordà[1] **Pablo Segovia-Martínez**[1] **Gonçal V. Garcés Díaz-Munío**[1] **Joan Albert Silvestre-Cerdà**[1]
Adrià Giménez[2] **Rubén Gaspar Aparicio**[3] **René Fernandez Sánchez**[3] **Jorge Civera**[1]
Albert Sanchis[1] **Alfons Juan**[1]

[1]MLLP research group, VRAIN, Universitat Politècnica de València, Spain
[2]Departament d'Informàtica, Escola Tècnica Superior d'Enginyeria, Universitat de València, Spain
[3]European Organization for Nuclear Research (CERN), Genève

www.mllp.upv.es

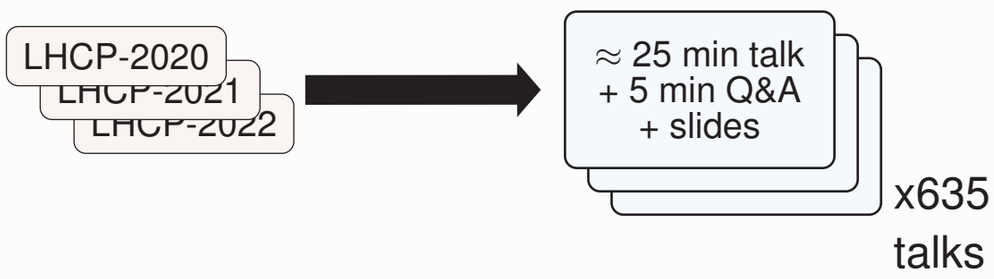
Introduction

- ASR is evaluated under highly controlled conditions.
- It is difficult to assess real progress.
- We introduce LHCP-ASR, an English speech corpus for narrow-domain and domain adaptation.
- Sourced from LHCP 2020-22 conferences.
- Featuring technical vocabulary and multi-speakers.

Data gathering

Speech data

- Source: Large Hadron Collider Physics conference
- Online and recorded due to the COVID-19 pandemic



- For a total of 235 hours
- Inherently complex audio:
  - Specialized content on high-energy physics
  - Diverse speakers: chairs, mods, audience...
  - Compressed audio and variable acoustics
  - Most speakers are non-native.

Text Data

- CERN Document Server (CDS)
  - Over 550K records
  - Scientific papers, reports, PhD thesis...
  - Very technical vocabulary
- CERN News repository
  - English news published since 1993
  - Easier vocabulary to facilitate communication

	Vocab. size	Tokens
CERN-news	40.9K	2.3M
CDS-abstracts	170.4K	20.1M
CDS-documents	2.7M	1.5G
Total	3.0M	1.5G

Dataset contents

LHCP-2020

- Developing ASR system for internal use at CERN
- Need of in-domain evaluation dataset:
  - **Manually transcribed** by CERN volunteers
  - Cross-reviewed, using guidelines
  - Speaker-independent partitions defined

Pseudo-labelling

- Iterative process to refine in-domain ASR system
- Used to obtain automatic verbatim transcriptions

LHCP-2022

- Second evaluation task for ensure performance
- Manual post-edition from pseudo-labels
- Speaker-independent partitions defined

LHCP-ASR speech data

Set		Talks	Hours	Vocab. size	Tokens
LHCP-2020	Dev	14	5.6	4.0K	53.0K
	Test	15	6.2	4.0K	56.4K
LHCP-2022	Dev	11	4.8	3.3K	42.8K
	Test	32	13.4	5.8K	116.9K
LHCP-train		560	204.8	18.1K	1.8M
Total		632	234.9	19.7K	2.1M

ASR reference results

- Initial results with our best-performing system
- Also with different Whisper sizes for comparison

ASR System	*LHCP-2020*		*LHCP-2022*	
	Dev	Test	Dev	Test
Whisper-turbo	13.2	15.9	17.3	17.7
Whisper-medium	13.8	15.4	16.7	16.7
Our in-domain system	**12.4**	**13.6**	**14.2**	**15.0**

Conclusions

- LHCP-ASR is a challenging and realiable benchmark
- This corpus offers a valuable resource for research
- Our in-domain ASR system outperforms general-purpose models like Whisper

Acknowledgements

The authors would like to thank the European Organisation for Nuclear Research (CERN) for their support during the PO OV9177345.

The research leading to these results has received funding from EU4Health Programme 2021–2027 as part of Europe's Beating Cancer Plan under Grant Agreements nos. 101056995 and 101129375; and from the Government of Spain's grant PID2021–122443OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU", and grant PDC2022–133049-I00 funded by MICIU/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR". The authors gratefully acknowledge the financial support of Generalitat Valenciana under project IDIFEDER/2021/059.