INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

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

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

Machine Learning and Language Processing (MLLP) research group
Institut Valencià d'Investigació en Intel·ligència Artificial (VRAIN)
Universitat Politècnica de València
Camí de Vera s/n, 46022, València, Spain
{gogardia, jsilvestre, jajorca, adgipas, jairsan, pabaar, narobe1}@vrain.upv.es,
{alpegon2, jorcisai, josanna2, ajuanci}@vrain.upv.es

Abstract

We introduce Europarl-ASR, a large speech and text corpus of parliamentary debates including 1 300 hours of transcribed speeches and 70 million tokens of text in English extracted from European Parliament sessions. The training set is labelled with the Parliament's non-fully-verbatim official transcripts, time-aligned. As verbatimness is critical for acoustic model training, we also provide automatically noise-filtered and automatically verbatimized transcripts of all speeches based on speech data filtering and verbatimization techniques. Additionally, 18 hours of transcribed speeches were manually verbatimized to build reliable speaker-dependent and speaker-independent development/test sets for streaming ASR benchmarking. The availability of manual non-verbatim and verbatim transcripts for dev/test speeches makes this corpus useful for the assessment of automatic filtering and verbatimization techniques. This paper describes the corpus and its creation, and provides off-line and streaming ASR baselines for both the speaker-dependent and speaker-independent tasks using the three training transcription sets. The corpus is publicly released under an open licence.

Index Terms: automatic speech recognition, speech corpus, speech data filtering, speech data verbatimization.

1. Introduction

Significant advances in automatic speech recognition (ASR) have recently taken place thanks to an increasing availability of speech and text resources, both for supervised [1, 2, 3, 4] and unsupervised [5, 6] learning. Nevertheless, there is still a lack of publicly available, realistic ASR tasks addressing real-life problems [7], especially for streaming ASR.

Moreover, optimal acoustic model training requires verbatim transcripts for speech data. However, available transcripts are usually not truly verbatim, since human transcriptions are typically rephrased and linguistically corrected. Thus, most existing web-sourced ASR corpora can only offer non-verbatim transcripts [1, 2, 4, 6] for ASR training and evaluation. In this scenario, the use of automatic speech data filtering and speech data verbatimization techniques can be key in enhancing ASR quality by making the data less noisy and closer to verbatim. Nevertheless, progress on filtering and verbatimization has been hindered by a lack of appropriate benchmarks.

In this context, some recent speech corpora have exploited the vast multilingual speech and text data available from the European Parliament (EP). In 2019, we released Europarl-ST [4] as a multilingual corpus for speech translation training and benchmarking (which has already found some success [8, 9, 10, 6]), based on EP speech data with transcriptions and translations from 2008–2012 (including 186 h of English labelled speech data). This was followed by the recent release in March 2021 of the VoxPopuli multilingual speech corpus [6] for unsupervised/semi-supervised learning and speech-to-speech interpretation, which uses more EP speech and text data, labelled and unlabelled, for more languages, from 1996–2020 (including 552 h of English labelled speech data). Each corpus offers a predefined speaker-independent training/development/test partition based only on non-verbatim transcripts. But there is more EP data to be exploited in more ways by a corpus focusing on monolingual supervised ASR.

In this article, we introduce Europarl-ASR, a large corpus of parliamentary debates for (streaming) ASR benchmarking and speech data filtering/verbatimization extracted from EP sessions from 1996–2020. This corpus was developed independently by the authors from EP website-sourced data, beginning in late 2019 concurrently with the creation of Europarl-ST.

In its initial release, the corpus is focused on English (EN) monolingual annotated data. Its main highlights, which no other EP-based corpus offers, are:

- 1 300 hours of EN transcribed speech data.
- 18 hours of EN speech data with both revised verbatim and official non-verbatim transcriptions, split in 2 independent dev/test partitions for 2 realistic ASR tasks (with vs. without previous knowledge of the speaker).
- 3 full sets of timed transcriptions for the training data: official non-verbatim, automatically noise-filtered, and automatically verbatimized.
- 70 million tokens of EN text data.

The availability of manual verbatim transcriptions ensures reliable ASR benchmarking, and, together with the original non-verbatim transcriptions, enables the assessment of filtering and verbatimization techniques. Baseline ASR performances are provided on both proposed tasks using the three training transcription sets, considering not only off-line ASR but also streaming ASR under strict low-latency requirements, as this corpus is a genuine task for streaming ASR benchmarking.

The Europarl-ASR corpus is released under an open licence at `https://www.mllp.upv.es/europarl-asr`.

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

2. Data gathering and selection

The Europarl-ASR corpus is sourced from EP debates, which consist of single-speaker speeches produced in turns, in any of the 24 official EU languages. Also, oral interpretation in the other 23 languages is carried out in real time by EP interpreters.

On May 2020, we crawled all publicly available data from the so-called *verbatim reports of proceedings* in the EP site [11]. This included the speeches from all parliamentary sessions held from July 1999 to May 2020. Within this period, however, there are three kinds of data resources spanning different subperiods:

- *Manual speech transcriptions:* July 1999 to May 2020.
- *Manual speech translations:* July 1999 to Nov. 2012.
- *Audio and video files,* including original and dubbed audio tracks: Sep. 2008 to May 2020.

In brief, the data gathering process resulted in roughly 2.9M speeches from 1.2K parliamentary sessions, 4.9K raw hours of accumulated original audio speech data from 2.5K different speakers, and 108K extra raw hours of interpreted/dubbed audio tracks[1]. Considering the original and dubbed audio tracks, each language has in total about 4.9K raw hours.

As discussed in the introduction, the Europarl-ASR corpus is (in its initial release) limited to English, which is the most resourced language with 98K speeches, 38K of which come with audio data, accounting for 1.3K raw hours of original English audio. In this regard, as transcribed audio is provided at the session's agenda item level, we first proceeded with its segmentation at speech level. To this end, we took advantage of the approximate manual start and end time stamps provided by the EP for each speech, which were heuristically adjusted when needed (e.g., to avoid overlapping between consecutive speeches). Although the time boundaries obtained by adjusting EP start-end time stamps may not be perfect for all speeches, we found that they are fairly clean in most cases and, when not, they are generally realistic, unclear boundaries between consecutive speeches that may need more in-depth subsequent processing.

After setting the start and end times for all speeches, a data selection procedure was applied at the speech level on the basis of character error rate (CER) scores. We automatically transcribed all English speeches with audio data (38 086 files, 1 340 raw hours) using an in-house low-profile English ASR system, based on a feed-forward DNN-HMM acoustic model trained on 5.6K speech hours [4]. Then, CER scores were computed for these automatic transcriptions against the official transcriptions, and all speeches with CER scores greater than 50% were discarded. This threshold was empirically defined by means of manual inspection of randomly selected samples, aiming to filter out noisy data (incorrect timestamps, unrelated/empty transcriptions, etc.). As a result, we retained roughly 33 002 speeches and 1 263 raw hours, that is, 94% of the total initial audio. In addition, we force-aligned all these speeches with their corresponding transcriptions, using the aforementioned acoustic model, in order to obtain audio-to-word alignments.

Regarding text data, we gathered the manual transcriptions of the 98K English speeches produced since July 1999, along with the English translations of 180K other speeches produced until November 2012, reaching 62.7M tokens in total.

3. Data partition and tasks

The EP is composed of 705 members (MEP) elected in the 27 Member States of the European Union every 5 years. Most EP speeches (∼80%) are delivered by MEPs, though many (∼20%) are given by guest speakers. For ASR, it is worth noting that MEPs are known beforehand, at the beginning of each parliamentary term, and even earlier in the case of reelected MEPs. Thus, a first *MEP* use case or task consists in building *speaker-dependent* ASR systems of (potentially) enhanced performance by exploiting prior knowledge (data resources) for MEPs already *seen* in previous (re)training stages. Contrastingly, guest speakers may intervene a few times or just once, so limited to no information might be available beforehand. As is usual in ASR nowadays, this leads to a second *Guest* task in which *speaker-independent* ASR systems are required to deal with speech data from guest speakers *not seen* in system (re)training.

Selected English text and audio data were used to source each task with its own development and test sets, plus a common training set comprising the rest of the data. For the MEP speaker-dependent task, we randomly selected 21 MEPs with comparatively large speech data available for them to be adequately represented in the train, dev and test sets. In accordance with the gender distribution in the whole data crawl (∼65% male, ∼35% female), 13 male and 8 female speakers were selected[2]. Then, a number of speeches from selected speakers were randomly drawn so as to build equally distributed dev and test sets of about 3.5 speech hours each, with roughly 10 minutes of audio in each set for every selected speaker.

For the Guest speaker-independent task, 12 guests with limited speech data were randomly selected, thus reserving guests with more data to maximize the amount of guest data in the training set. To preserve gender balance, 6 female and 6 male speakers were selected and evenly distributed over the dev and test sets (i.e., 3 female and 3 male speakers in each set). As in the MEP task, speeches were randomly drawn for the selected speakers, resulting in dev and test sets of 3 hours each, with about 30 min of speech in each set per selected speaker.

All speeches not included in any of the dev and test sets, except those from the 12 guest speakers selected for the Guest task, were allocated to the training set. In total, this set comprises 32 335 speeches from 1 034 speakers (681 male, 353 female), accounting for 1 230 raw hours and 9.7M text tokens. Also, previously computed audio-to-word alignments were used to build 1.1M acoustic model training segments, lasting 3.2 seconds on average and 1 007 hours overall (including silences), with pure speech duration (excluding non-speech events and silences) standing at 920 hours. It is important to remark that, due to the non-verbatimness of the official transcripts (Sec. 4), these force-aligned segments may include alignment errors (which are not handled). Table 1 provides overall statistics for the train, dev and test sets.

Table 1: *Overall statistics of the dev, test and train sets.*

Set	Speakers	Speeches	Raw	Speech	Tokens
MEP-dev	21	159	4.6h	3.4h	37.9K
MEP-test	21	145	4.7h	3.5h	38.8K
Guest-dev	6	52	4.3h	3.0h	31.8K
Guest-test	6	56	3.9h	3.0h	32.0K
train	1034	32335	1230h	920h	9.7M

Apart from the speech data above, we also prepared five sets of in-domain English text data for language modelling:

1. *Training set speeches:* Sep. 2008 to May 2020.

[1] We use K for "thousand" and M for "million" as abbreviations.

[2] The gender of the 1 046 speakers was automatically inferred from their names using the https://genderize.io/ REST API.

2. *Speeches without audio:* July 1999 to Aug. 2008.
3. *Translations into English:* July 1999 to Nov. 2012.
4. *Europarl-v10 [12] English text:* Apr. 1996 to July 1999.
5. *DCEP corpus [13] English text.*

Table 2 shows overall statistics for these sets. The Europarl-v10 text was selected from English transcriptions and translations with no overlap in dates with the first three sets, so that data deduplication is not needed. The DCEP text is from published EP-published official documents, excluding "verbatim" reports of proceedings from parliamentary debates [13]. It was included to increase the total amount of in-domain tokens so that competitive language models could be built.

Table 2: *Overall statistics of in-domain English text data.*

Data source	Speeches	Tokens	Sum
Training set speeches	32.3K	9.7M	9.7M
Speeches without audio	20.5K	6.4M	16.1M
Translations into English	180.5K	42.3M	58.4M
Europarl-v10 English subset	N/A	11.0M	69.4M
DCEP corpus English text	-	103.5M	172.9M

4. Manual revision of transcriptions

The EP's official transcripts are not fully verbatim, despite being close to the actual uttered speech. On one hand, they suffer from non-transcribed speech as they are subject to summarization and rephrasing, as well as to deliberate omissions of spontaneous speech lapses (false starts, repetitions...). On the other hand, they contain unuttered text, again due to summarization and rephrasing, but also to linguistic corrections and to the insertion of unuttered standard phrases and formalisms, such as "*Mister President*" at the beginning of speeches.

To provide fully verbatim transcriptions for reliable ASR benchmarking, all development and test sets were manually post-edited using the TLP Player [14] at the MLLP Transcription and Translation Platform [15]. Specific post-editing guidelines[3] were applied to obtain truly verbatim transcripts for each speech revised. Thus, we provide two transcription references for each development and test speech: the original, raw reference (*raw*), and a revised, verbatim reference (*verb*). Considering the manually revised transcripts as the true reference, the original raw transcripts are at a WER of 11.0% overall.

5. Filtering and verbatimization

Having fully verbatim transcriptions for development and test purposes is ideal in order to reliably assess and compare ASR systems. Similarly, it would be best to have fully verbatim transcriptions for training too, but the sheer size of the train set makes it extremely costly to manually revise. This is in fact a common situation for ASR system builders nowadays: in general, it is expected that (minor) discrepancies between available raw transcriptions and their unavailable verbatim counterparts will be largely compensated by exploiting more training data. Although we agree with this view to a certain extent, we have recently observed that, when available transcriptions are at a significant (WER) distance of the true (verbatim) transcriptions, preprocessing of training data by refined *"noise" filtering* certainly pays off [16, 17]. This being the case with EP data, we decided to apply this refined filtering to the training set, and also a novel, more advanced kind of transcription "reconstruction" preprocessing which we refer to as *verbatimization*.

The goal of filtering is to detect and discard acoustic segments unrelated to the available approximate transcription. As described in [16], a pre-existing acoustic model is first used to force-align a given audio and its approximate transcription, and then to decide whether to accept or reject each audio-to-word unit on the basis of phoneme duration and alignment score statistics; finally, training segments are built by joining consecutive accepted words. This procedure was applied to the whole training set using the already computed forced-alignments (Sec. 2). As a result, 33% of the audio data was filtered out, leaving 1.2M training segments of 2 seconds on average (672 hours overall).

In contrast to filtering, which is limited to discarding unreliable audio segments, verbatimization is allowed to freely "reconstruct" verbatim transcriptions from their raw, approximate counterparts. Broadly speaking, the proposed verbatimization technique consists in replacing approximate transcriptions by the automatic transcriptions produced by an ASR system using an LM heavily biased towards a given transcription, though still free enough to deliver completely different output. In particular, for the train set, KenLM [18] was used to train in-domain (ID) trigram LMs from each approximate transcription. These ID LMs were linearly interpolated with a pre-existing out-of-domain (OOD) general-purpose LM trained with 17.9G tokens. Interpolation weights were optimized to minimize perplexity on the (verbatim) MEP-dev set, resulting in ID and OOD weights of 93% and 7%, respectively. Then, each approximate transcription in the training set was verbatimized by using its corresponding interpolated LM in conjunction with a pre-existing acoustic model, resulting in 1M training segments of 3.7 seconds on average, lasting 1 054 hours overall.

Table 3 shows basic statistics for the three training transcription sets: *raw*, *filtered (filt)* and *verbatimized (verb)*.

Table 3: *Statistics of the three training transcription sets.*

Set	Segments	Duration (h)	Avg. len. (s)
raw	1.13M	1007	3.2
filt	1.24M	672	2.0
verb	1.03M	1054	3.7

From the statistics in Table 3, we can see that the filtering process resulted in an aggressive reduction of the effective training data while, contrarily, verbatimization not only did not reduce the duration of the raw training data but, apparently, it was even capable of uncovering missing transcription parts and thus enlarge its input. This behaviour suggests the proposed technique could be useful for purposes other than just "repairing" training data, though we defer this study for brevity. At this point, it is worth noting that the verbatimization procedure was also assessed on the MEP-dev set, for which it produced automatic verbatim transcriptions at a 6.5% WER distance from those manually generated. This amounts to a 35% relative improvement when compared to the 10.1% WER distance from the raw MEP-dev transcriptions to their revised counterparts.

6. Baseline experiments and results

To provide baseline figures for offline and streaming ASR in the MEP and Guest tasks, a common experimental setting was used to build a different hybrid ASR system for each of the train-

[3] https://www.mllp.upv.es/europarl-asr

ing transcription sets provided: raw, filt and verb. As in [17], acoustic modelling was done by first training context-dependent feed-forward DNN-HMMs with three left-to-right tied states using the transLectures-UPV toolkit [19]. State tying was based on a phonetic decision tree approach [20] which, in our case, produced 15K, 11K, and 15K tied states for, respectively, the raw, filt and verb data. Then, feed-forward models were used to bootstrap and train a BLSTM-HMM AM as in [21], though combining the MEP and Guest development sets for validation.

Language modelling was conducted on the basis of the in-domain English text data collected for that purpose (Sec. 3), from which we trained common n-gram and Transformer LMs for the three ASR systems, while keeping the MEP and Guest development sets apart for validation and task adaptation. On one hand, KenLM [18] was used to train a 4-gram LM with vocabulary size limited to 250K words and OOV ratio 0.4% on the dev sets. On the other hand, an in-house version of FairSeq [22] was used to train a Transformer LM (TLM) with the 4-gram LM vocabulary and the setup described in [17]. As in [21], variance regularization was applied to speed up computation of TLM scores during inference. Table 4 shows perplexities computed over development data for each LM and their linear interpolation. While the TLM clearly outperforms the n-gram model by almost halving its perplexity, the interpolation of both provides even better results in all cases.

Table 4: *LM perplexities computed over development data.*

	MEP	*Guest*	Both
4-gram	127.9	132.8	130.2
TLM	67.7	77.1	72.0
4-gram+TLM	63.6	69.9	66.5
TLM weight	82%	78%	80%

A real-time one-pass decoding based on a *history-conditioned search* strategy was used to build the desired hybrid ASR systems from the acoustic and language models described above [23]. Also, for the systems to effectively work under streaming conditions, BLSTM AMs were queried with a sliding, overlapping context window of 500 ms [24], while TLM history was limited to a maximum 40 words [21].

Table 5 shows WER figures for the *raw*, *filt* and *verb* offline ASR systems using either the 4-gram LM alone or interpolated with the TLM. For the computation of these figures, system hyperparameters were optimized for minimum WER under the constraint of decoding to be completed with a real time factor (RTF) close to one (from raw audio and no prior segmentation).

Table 5: *WERs for the three systems under an offline ASR setup.*

		raw		*filt*		*verb*	
		dev	test	dev	test	dev	test
MEP	4-gram	10.2	9.9	9.4	**9.2**	9.5	9.3
	+TLM	8.8	8.6	7.9	**7.8**	8.1	8.2
Guest	4-gram	11.2	8.7	10.7	8.8	10.2	**8.2**
	+TLM	9.5	7.6	9.0	7.4	8.7	**7.0**

From the results in Table 5, we see that, as anticipated by LM perplexities, LM interpolation consistently outperforms the 4-gram LM alone by up to 1.4 absolute WER points. Note also that the filtering and verbatimization techniques led to consistent WER improvements of up to 9% relative in both tasks w.r.t. the system trained with original, raw transcriptions. Moreover, when comparing filtering and verbatimization, we can observe that filtering resulted in a better WER on the MEP task, while verbatimization provided a better WER on the Guest task.

Table 6 shows WERs and latencies for the three systems when running under strict low-latency streaming conditions. Latencies are computed as the delay between the time instant when an acoustic frame is provided to the decoder, and the time instant when that frame is fully processed by the decoder and its corresponding output has been delivered. Search parameters were tuned accordingly to ensure low and stable latencies. With slight degradations of 1–4% WER, these systems were capable of providing live transcription outputs with a mean latency of 0.65 seconds and a very small standard deviation of 50 ms.

Table 6: *WER and latencies in seconds ($\mu \pm \sigma$) for the three systems running under a streaming ASR setup.*

	MEP		*Guest*	
	dev / test	Latency	dev / test	Latency
raw	9.0 / 8.8	0.66 ± 0.05	9.8 / 7.8	0.66 ± 0.05
filt	8.1 / **7.9**	0.65 ± 0.05	9.3 / 7.5	0.65 ± 0.05
verb	8.4 / 8.3	0.65 ± 0.05	9.1 / **7.3**	0.65 ± 0.05

7. Conclusions

A new corpus of parliamentary debates, Europarl-ASR, has been introduced for streaming ASR benchmarking and, pioneeringly, for assessing filtering and verbatimization techniques. Its first version has been publicly released under an open licence, featuring roughly 1 300 hours of transcribed English speeches plus 18 hours of manually verbatimized evaluation data. Strong baseline ASR results have been reported on the two test sets (*MEP* / *Guest*): 7.8 / 7.0 WER for the offline setup, and 7.9 / 7.3 WER when considering a realistic use of ASR at the EP, that is, under tight streaming conditions, with an empirical mean latency of 0.65 seconds. Also, the application of filtering and verbatimization techniques over the official non-verbatim training transcripts has been shown to result in systematic and consistent WER gains of 9% relative in both tasks.

As for future releases of this corpus, first, we plan to extend the English-language data with up to 3.5K hours of speeches interpreted into English, applying verbatimization over manual (1.4K hours) and automatically generated (2.1K hours) English translations. Second, we intend to define and incorporate assessment resources and metrics (based on harmonic F-measures) to gauge filtering techniques. Finally, we will add support for other EU languages such as German or Spanish.

8. Acknowledgements

This work has received funding from the EU's H2020 research and innovation programme under grant agreements 761758 (X5gon) and 952215 (TAILOR); the Government of Spain's research project Multisub (RTI2018-094879-B-I00, MCIU/AEI/FEDER,EU) and FPU scholarships FPU14/03981 and FPU18/04135; the Generalitat Valenciana's research project Classroom Activity Recognition (PROMETEO/2019/111) and predoctoral research scholarship ACIF/2017/055; and the Universitat Politècnica de València's PAID-01-17 R&D support programme.

9. References

[1] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," in *Proc. LREC 2014*, Reykjavik, Iceland, 2014, pp. 3935–3939.

[2] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proc. 2019 Conf. of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, 2019, pp. 2012–2017.

[3] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Proc. Interspeech 2020*, 2020, pp. 2757–2761.

[4] J. Iranzo-Sánchez *et al.*, "Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates," in *Proc. 45th Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, 2020, pp. 8229–8233.

[5] J. Kahn *et al.*, "Libri-Light: A Benchmark for ASR with Limited or No Supervision," in *Proc. ICASSP 2020*, 2020, pp. 7669–7673.

[6] C. Wang *et al.*, "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," *arXiv:2101.00390*, 2021.

[7] T. Likhomanenko *et al.*, "Rethinking Evaluation in ASR: Are Our Models Robust Enough?" *arXiv:2010.11745*, 2020.

[8] E. Ansari *et al.*, "FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN," in *Proc. 17th Intl. Conf. on Spoken Language Translation (IWSLT 2020)*, Seattle, WA, USA, 2020, pp. 1–34.

[9] H. Nguyen, Y. Estève, and L. Besacier, "An Empirical Study of End-to-end Simultaneous Speech Translation Decoding Strategies," *arXiv:2103.03233*, 2021, (accepted for publication at ICASSP 2021).

[10] M. Zanon Boito, W. Havard, M. Garnerin, É. Le Ferrand, and L. Besacier, ""MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible"," in *Proc. 12th Language Resources and Evaluation Conf. (LREC 2020)*, Marseille, France, 2020, pp. 6486–6493.

[11] European Parliament, "Verbatim reports of proceedings of the European Parliament Plenary," European Parliament. https://www.europarl.europa.eu/plenary/en/debates-video.html, (accessed 27 May 2020).

[12] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Proc. MT Summit 2005*, Phuket, Thailand, 2005, pp. 79–86.

[13] N. Hajlaoui, D. Kolovratnik, J. Väyrynen, R. Steinberger, and D. Varga, "DCEP -Digital Corpus of the European Parliament," in *Proc. 9th Intl. Conf. on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014, pp. 3164–3171.

[14] J. A. Silvestre-Cerdà, A. Pérez, M. Jiménez, C. Turró, A. Juan, and J. Civera, "A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories," in *Proc. IEEE Intl. Conf. on Systems, Man, and Cybernetics SMC 2013*, Manchester, UK, 2013, pp. 3994–3999.

[15] The MLLP Team, "MLLP Platform for Transcription and Translation," MLLP. https://ttp.mllp.upv.es, (accessed 26 March 2021).

[16] J. Jorge *et al.*, "MLLP-UPV and RWTH Aachen Spanish ASR Systems for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge," in *Proc. IberSPEECH 2018: 10th Jornadas en Tecnologías del Habla and 6th Iberian SLTech Workshop*, Barcelona, Spain, 2018, pp. 257–261.

[17] J. Jorge, A. Giménez *et al.*, "MLLP-VRAIN Spanish ASR Systems for the Albayzin-RTVE 2020 Speech-To-Text Challenge," in *Proc. IberSPEECH 2021*, 2021, pp. 118–122.

[18] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proc. 6th Workshop on Statistical Machine Translation*, Edinburgh, Scotland, 2011, pp. 187–197.

[19] M. A. del Agua *et al.*, "The transLectures-UPV toolkit," in *Proc. VIII Jornadas en Tecnología del Habla and IV Iberian SLTech Workshop (IberSPEECH 2014)*, Las Palmas de Gran Canaria, Spain, 2014, pp. 269–278.

[20] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based State Tying for High Accuracy Acoustic Modelling," in *Proc. Workshop on Human Language Technology*, Plainsboro, NJ, USA, 1994, pp. 307–312.

[21] P. Baquero-Arnal *et al.*, "Improved Hybrid Streaming ASR with Transformer Language Models," in *Proc. InterSpeech 2020*, 2020, pp. 2127–2131.

[22] M. Ott *et al.*, "fairseq: A Fast, Extensible Toolkit for Sequence Modeling," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 48–53.

[23] J. Jorge, A. Giménez, J. Iranzo-Sánchez, J. Civera, A. Sanchis, and A. Juan, "Real-Time One-Pass Decoder for Speech Recognition Using LSTM Language Models," in *Proc. Interspeech 2019*, Graz, Austria, 2019, pp. 3820–3824.

[24] J. Jorge *et al.*, "LSTM-Based One-Pass Decoder for Low-Latency Streaming," in *Proc. ICASSP 2020*, 2020, pp. 7814–7818.