MLLP-UPV and RWTH Aachen Spanish ASR Systems for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge

Javier Jorge, Adrià Martínez-Villaronga, Pavel Golik, Adrià Giménez, Joan Albert Silvestre-Cerdà, Patrick Doetsch, Vicent Andreu Císcar, Hermann Ney, Alfons Juan and Albert Sanchis

DATASET

- 17 different TV shows, 2015-2018.
- 569 hours, from which:
  - 460h are provided with subtitles ($\neq$ verbatim transcripts).
  - 109h have been human-revised transcribed.
- Database is provided in 5 partitions, either speaker and TV show dependent:
  - *train*: 460h of speech data with non-verbatim subtitles from 16 TV shows.
  - *subs-C24H*: 3M sentences from the *24H Channel* subtitles during 2017.
  - *dev1*: 53h of speech data from 5 TV shows.
  - *dev2*: 15h of speech data from 2 TV shows.
  - *test*: 40h of speech data from 8 TV shows.
- We divided *dev1* set into two subsets: *dev1-train* (43h) and *dev1-dev* (15h).

CLOSED-CONDITION

Speech Data filtering

- Preexisting hybrid CD-DNN-HMM ASR system to force-align audio and text.
- Heuristic post-filtering based on phoneme lengths and alignment scores.
- Discarded those files in which more than 2/3 of the words were filtered.
- Joining words into segments delimited by large-enough silences and deleted words.
- Hours of raw data, after alignment and then filtered:

	Raw	Aligned Raw	Aligned Speech	Filtered Speech
train	463	438	252	187
dev1-train	43	31	24	18
dev1-dev	15	12	9	7
dev2	15	12	9	6
Overall	535	493	294	218

Text Data filtering

- Sentence re-segmentation, lowercase, abbreviation expansion, numbers transliteration.

	Sentences	Running words	Vocabulary
train	340K	4.3M	80K
subs-C24H	3.1M	57M	160K
RNN-train	1.8M	35M	176K
dev1-dev	9.9K	160K	13K
dev2	7.7K	150K	12K

- Final vocabulary without singletons: ∼132K words.
- OOV ratios of *dev1-dev* and *dev2* sets were 0.36% and 0.53%, respectively.

System summary

- Hybrid BLSTM-HMM acoustic model + *n-gram*/RNN interpolation.
- Toolkits: *transLectures-UPV Toolkit* (TLK) + TensorFlow + SRILM + CUED-RNNLM.

Acoustic Models

- 48-dim features (16 MFCCs + deriv.).
- 8.9K tied triphone states (using CART).
- Bi-directional LSTM 4 layers, 512 units per direction.

Language Models

- Separate 4-gram LMs trained for *train* and *subs-C24H* sets.
- Linear interpolation of both LMs → *4-gram general LM*.
  - Interpolation weights optimized on *dev1-dev*.
- Show-specific 4-gram LMs:
  - TV Show IDs for *dev1*, *dev2* and *test* files are known beforehand.
  - Created train/dev sets for each TV Show.
  - Each show-specific LM linearly interpolated with general LMs → *4-gram adapt LMs*.
- RNN LMs: trained on *train* + *subs-C24H*, 1 layer of 1024 LSTM units.

Results

- Comparison of acoustic models combinations.

	dev1-dev WER	*dev2* WER	*dev2* ΔWER
FFDNN	29.7	27.1	-
BLSTM	26.5	23.8	12.2

- Comparison of language model combinations.

	dev1-dev PPL	*dev1-dev* WER	*dev2* PPL	*dev2* WER	*dev2* ΔWER
4-gram general	107	26.5	148	23.8	-
RNN	92	26.2	111	23.0	3.4
RNN + 4-gram general	78	25.3	102	22.4	5.9
RNN + 4-gram adapt	69	24.8	99	22.4	5.9

- Comparison of audio segmentation/VAD techniques.

	dev1-dev % drop.	*dev1-dev* WER	*dev2* % drop.	*dev2* WER	*dev2* ΔWER
MLLP-UPV (1)	10.9	24.8	5.9	22.4	-
LIUM (2)	7.1	23.7	3.9	20.8	7.1
CMUseg (3)	0	23.2	0	20.9	6.7
Pre-Recognition (4)	0	22.9	0	20.6	8.0
+ MLLP-UPV (5)	3.2	**22.3**	3.3	**20.0**	10.7

- RTF analysis an its effect on WER.

	RTF	WER
Submitted system (5)	1.5	20.0
+ inc. prune	0.8	20.3

OPEN-CONDITION

System summary

- Hybrid BLSTM-HMM acoustic model + *n-gram* language model.
- Toolkit: RASR + RETURNN from RWTH Aachen University.
- Training data:
  - 3800 hours from several sources.
  - Vocabulary size: 325k words.
  - One or more pronunciation variants per word.
  - Variety of domains and acoustic conditions.
  - No overlap with IberSpeech-RTVE challenge data.

Acoustic Model

- 80-dim MFCC features.
- 5k tied triphone states (using CART).
- Bi-directional LSTM 4 layers, 512 units per direction.
- Optimization meta-parameters:
  - 30% activations dropped per Dropout.
  - Adam with Nesterov momentum.
  - Epoch based Newbob.
- Multi-GPU Training:
  - We split input utterances into overlapping chunks of roughly 10 seconds.
  - L2 normalization of the gradients for each chunk.
  - Distribute up to 50 chunks over multiple GPUs.
  - Re-combine approx. every 500 chunks.

Language model

- 5-gram with KN smoothing.
- Trained on multiple publicly available sources.
- No optimization/adaptation to IberSpeech-RTVE challenge data.

FINAL RESULTS

- WER on *dev2* and test partitions, per TV show and globally.

TV shows	Closed	Open
dev2 (Millennium & La noche en 24H)	20.0	15.6
Al filo de lo imposible (AFI)	25.4	15.9
Arranca en Verde (AV)	36.2	23.9
Dicho y Hecho (DH)	50.8	34.4
España en comunidad (EC)	16.7	11.4
Latinomérica en 24H (LA24H)	12.0	7.4
La Mañana (LM)	26.8	21.9
La tarde en 24H Tertulia (LT24HTer)	19.1	19.0
Saber y Ganar (SG)	18.8	16.0
Global	22.0	**16.4**