IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

The Vicomtech-PRHLT Speech Transcription Systems for the IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge

Haritz Arzelus[1], *Aitor Álvarez*[1], *Conrad Bernath*[1], *Eneritz García*[1],
Emilio Granell[2], *Carlos-D. Martínez-Hinarejos*[2]

[1]Vicomtech, Human Speech and Language Technology Group, Spain
[2]Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, Spain

[harzelus,aalvarez,cbernath,egarciam]@vicomtech.org, [egranell,cmartine]@dsic.upv.es

Abstract

This paper describes our joint submission to the IberSPEECH-RTVE Speech to Text Transcription Challenge 2018, which calls for automatic speech transcription systems to be evaluated in realistic TV shows. With the aim of building and evaluating systems, RTVE licensed around 569 hours of different TV programs, which were processed, re-aligned and revised in order to discard segments with imperfect transcriptions. This task reduced the corpus to 136 hours that we considered as nearly perfectly aligned audios and that we employed as in-domain data to train acoustic models.

A total of 6 systems were built and presented to the evaluation challenge, three systems per condition. These recognition engines are different versions, evolution and configurations of two main architectures. The first architecture includes an hybrid LSTM-HMM acoustic model, where bidirectional LSTMs were trained to provide posterior probabilities for the HMM states. The language model corresponds to modified Kneser-Ney smoothed 3-gram and 9-gram models used for decoding and re-scoring of the lattices respectively. The second architecture includes an End-To-End based recognition system, which combines 2D convolutional neural networks as spectral feature extractor from spectrograms with bidirectional Gated Recurrent Units as RNN acoustic models. A modified Kneser-Ney smoothed 5-gram model was also integrated to re-score the E2E hypothesis. All the systems' outputs were then punctuated using bidirectional RNN models with attention mechanism and capitalized through *recasing* techniques.

Index Terms: speech recognition, deep learning, end-to-end speech recognition, recurrent neural networks

1. Introduction

The IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge calls for Automatic Speech Recognition (ASR) systems that are robust against realistic TV shows. It is a notable trend that aims to approach ASR technology to different applications such as automatic subtitling or metadata generation in the broadcast domain. Although most of this work is still performed manually or through semiautomatic methods (e.g. re-speaking), the current state of the art in speech recognition suggests that this technology could start to be exploitable autonomously without any post-edition task, mainly on contents with optimal audio quality and clean speech conditions.

The use of Deep Learning algorithms in speech processing have made it possible to introduce this technology in such a complex scenario through the use of systems based on Deep Neural Networks (DNNs) or more recent architectures based on the End-To-End (E2E) principle.

Historically, ASR systems have made use of Hidden Markov Models (HMMs) to capture the time variability of the signal and Gaussian Mixture Models (GMMs) to model the HMM state probability distributions. However, numerous works have shown that DNNs in combination with HMMs can outperform traditional GMM-HMM systems at acoustic modeling on a variety of datasets [1]. More recently, new attempts have been focused on building E2E ASR architectures [2], which directly map the input speech signal to grapheme/character sequences and jointly train the acoustic, pronunciation and language models as a whole unit [3, 4, 5, 6]. Nowadays, two main approaches predominate to train E2E ASR models. On the one hand, the Connectionist Temporal Classification (CTC) is probably the most widely used criterion for systems based on characters [2, 7, 8], sub-words [9] or words [10]. It employs Markov assumptions and dynamic programming to efficiently solve sequential problems [2, 3, 7]. On the other hand, attention-based methods employ an attention mechanism to perform alignment between acoustic frames and characters [4, 5, 6]. Unlike CTC, it does not require several conditional independence assumptions to obtain the label sequence probabilities, allowing extremely non-sequential alignments. Additionally, a number of enhancement techniques have been employed to overcome the performance of these systems, such as Data Augmentation [11], Transfer Learning [12], Dropout [13] or Curriculum Learning [14], among others.

Our systems were constructed following both DNN-HMM and E2E architectures basis, given that depending on the available training data, one approach performed more robustly than the other on the *development set*. A total of 6 systems were presented to the evaluation challenge, three systems per condition (*closed* and *open*). Regarding the *closed condition*, in which only the data released by RTVE could be used to train and evaluate systems, two systems based on the DNN-HMM and one E2E system were presented. In contrast, the results obtained from two E2E based systems and one DNN-HMM system were submitted for the *open condition*. In all systems, the raw recognized output was punctuated, capitalized and normalized automatically using Recurrent Neural Models (RNNs), *recasing* techniques and rule-based heuristics respectively.

2. Corpus processing

Depending on the training condition, different datasets were used to train and evaluate the ASR systems.

10.21437/IberSPEECH.2018-56

2.1. RTVE2018 dataset

The RTVE2018 dataset was released by RTVE and comprises a collection of TV shows drawn from diverse genres and broadcast by the public Spanish National Television (RTVE) from 2015 to 2018. The real number of hours provided in the original dataset as *training* and *development sets* is presented in Table 1.

Table 1: *Duration of each partition of the original RTVE2018 dataset*

subset	duration
train	462 h. 9 min.
dev1	62 h. 23 min.
dev2	15 h. 13 min.
total	539 h. 45 min.

The main problem of this dataset was that a great amount of audios had imperfect transcriptions and, therefore, they could not be used as such for training and evaluation purposes. With the aim of recovering only the correctly aligned segments, a highly costly process was carried out, where an alignment and re-alignment techniques were performed first and a manual revision and automatic recognition task afterwards. The alignment and re-alignment processes consist of two steps. In the first step, we tried to align the original audios with their corresponding transcriptions using 4 different beam values (10; 100; 1,000 and 10,000). For the case of the *train* partition, a total of 101 hours and 47 minutes were only aligned after this initial step. Thus, 360 hours and 22 minutes were definitely discarded to be used in any training process. A second step of re-alignment was then performed over this new subset of 101 hours and 47 minutes, using beam and retry-beam values of 1 and 2 respectively, obtaining a total of 86 hours and 29 minutes of audio segments that were considered as nearly correctly aligned. The alignments processes were performed using a feed-forward DNN-HMM acoustic model trained with the Kaldi toolkit [15] and estimated over contents from the broadcast domain. Finally, a small partition of the nearly correctly aligned hours were revised manually, whilst the remaining were recognized using a different recognition architecture as employed for the alignments. In this case, the recognition was performed using an E2E based recognition system trained with the same contents from the broadcast domain. Only the recognition outputs that fit exactly to the reference were tagged as perfect segments. The same *cleaning* methodology was also applied on the *dev1* and *dev2* partitions.

The total number of hours discarded after the first step, and the hours tagged as nearly perfect and completely perfect are summarized in Table 2. These hours correspond to audio segments that lasted more than one second, since the shorter ones were also discarded.

Table 2: *RTVE2018 dataset after the alignment and re-alignment processes*

subset	alignment (discarded)	re-alignment (nearly perfect)	revision+E2E (perfect)
train	360 h. 22 min.	86 h. 29 min.	56 h. 27 min.
dev1	7 h. 43 min.	44 h. 34 min.	29 h. 39 min.
dev2	9 h. 27 min.	5 h. 8 min.	4 h. 3 min.
total	377 h. 32 min.	136 h. 11 min.	90 h. 9 min.

As it can be seen in Table 2, a high number of hours were discarded from the original RTVE2018 dataset. In the end, a total of 136 hours and 11 minutes were considered as nearly perfect audios, whilst only 90 hours and 9 minutes can be thought to be perfectly aligned including the *train, dev1 and dev2* re-aligned partitions. These both subsets were finally used to build and tune the acoustic models (AM). The *development set* for the tuning of the systems was extracted from the completely perfect subset, as it is shown in Table 3.

Table 3: *New perfectly and nearly perfectly aligned subsets for train and development. The 4 hours from dev correspond to the same contents in both subsets.*

subset	*train*	*dev*
Perfectly aligned	86 h. 9 min.	4 h.
Nearly perfect aligned	132 h. 11 min.	4 h.

In terms of text data, a total of 3.5 million sentences and 61 million words were compiled. This data was used to estimate the language models (LM) and the punctuation and capitalization modules.

2.2. Open dataset

The open dataset was used to build the ASR systems for the *open condition*. In addition to the perfectly re-aligned subset from the RTVE2018 dataset, 4 different corpora were prepared for training. The SAVAS corpus [16] is composed of broadcast news contents from the Basque Country's public broadcast corporation EiTB (Euskal Irrati Telebista), and includes annotated and transcribed audios in both clear (studio) and noisy (outside) conditions. The Youtube RTVE Series corpus includes Spanish broadcast contents of RTVE shows and series gathered from the Youtube platform. The audio contents were downloaded along with the automatic transcriptions provided by the platform. These audios and their corresponding automatic transcriptions were then split and re-aligned following the same methodology as it was explained in Section 2.1. Finally, the Albayzin [17] and Multext [18] corpora were also included. The *development set* corresponded to the in-domain new *dev* partition shown in Table 3. The total amount of hours available for the *open condition* are summarized in Table 4.

Table 4: *The open dataset description*

corpus	#hours
RTVE2018	132 h. 11 min.
SAVAS	160 h. 58 min.
Youtube RTVE	197 h. 13 min.
Albayzin	6 h. 5 min.
Multext	53 min.
Total	497 h. 20 min.

Regarding text data, *data selection* techniques were applied on general news data gathered from digital newspapers and using the LM created with the in-domain RTVE2018 text data as a reference. A total of 3.5 million sentences and 71 million words were selected with a maximum perplexity threshold value of 120. Hence, summing the in-domain and new texts data, a total of 132 million words were employed to estimate the LM, and punctuation and capitalization modules for the *open condition*.

3. Main architectures

Two main architectures were employed to build the systems for both *closed* and *open conditions*.

3.1. LSTM-HMM based systems

These systems include a bidirectional LSTM-HMM acoustic model and n-gram language models for decoding and rescoring porpuses. The AMs and final graphs were estimated using the Kaldi toolkit. The AM corresponded to a hybrid LSTM-HMM implementation, where bidirectional LSTMs were trained to provide posterior probability estimates for the HMM states. This model was constructed with a sequence of 3 LSTM layers, using 640 memory units in the cell and 1024 fully connected hidden layer outputs. The number of steps used in the estimation of the LSTM state before prediction of the first label was fixed to 40 in both contexts. Furthermore, modified Kneser-Ney smoothed 3-gram and 9-gram models were used for decoding and re-scoring of the lattices respectively. Both LMs were estimated using the KenLM toolkit [19].

3.2. E2E based systems

The E2E systems were developed following the Deep Speech 2 architecture [2]. The core of the system is basically an RNN model, in which speech spectrograms are ingested and text transcriptions are provided as output.

Initially, a sequence of 2 layers of 2D convolutional neural networks (CNN) are employed as spectral feature extractor from spectrograms. A 2D batch normalization function is then applied to the output of both layers, in addition to a *hard tanh* function as an activation function. The E2E systems were set up using 5 layers of bidirectional Gated Recurrent Units (GRU) [20] layers as RNN networks. Each hidden layer is composed of 800 hidden units. After the bidirectional recurrent layers, a fully connected layer is applied as the last layer of the whole model. The output corresponds to a *softmax* function which computes a probability distribution over the characters. During the training process, the CTC loss function is computed to measure the error of the predictions, whilst the gradient is estimated using backpropagation through time algorithm with the aim of updating the network parameters. The optimizer is the Stochastic Gradient Descent (SGD).

In addition, an external LM was integrated for decoding with the aim of rescoring the initial lattices. To this end, modified Kneser-Ney smoothed 5-grams models were estimated using the KenLM toolkit.

4. Systems descriptions

A total of 6 systems based on the above described architectures were submitted to the challenge, three systems per condition.

4.1. Closed condition

4.1.1. Primary system

The primary system submitted to the *closed condition* was called 'Vicomtech-PRHLT_p-K1_closed' and it is a bidirectional LSTM-HMM based system combined with a 3-gram LM for decoding and a 9-gram LM for re-scoring lattices. The AM was trained for 10 epochs, with an initial and final learning rate of 0.0006 and 0.00006 respectively, using a mini-batch size of 100 and 20,000 samples per iteration. The AM was trained with the nearly perfectly aligned partition (see Table 3), which was 3-fold augmented through the *speed* based augmentation technique. Each audio was transformed randomly depending on a modification parameter ranged between 0.9 and 1.1 values. A total of 396 hours and 33 minutes were therefore used for training. The LMs were estimated with the in-domain texts compiled from the RTVE2018 dataset.

4.1.2. Contrastive systems

The first constrastive system was called 'Vicomtech-PRHLT_c1-K2_closed' and it was set up using the same configuration of the primary system, but the AM was estimated using the 3-fold augmented acoustic data of the perfectly aligned partition (see Table 3). A total of 258 hours and 27 minutes were employed for training.

The same data was used to build the the second contrastive system, tagged as 'Vicomtech-PRHLT_c2-E1_closed'. It was an E2E recognition system which follows the architecture described above, and it was evolved for 30 epochs. The LM was a 5-gram with an *alpha* value of 1.5 and a beam-width of 1000 during decoding.

4.2. Open condition

4.2.1. Primary system

The primary system of the *open condition* was called 'Vicomtech-PRHLT_p-E1_open' and it was based on the E2E architecture described in Section 3.2. This system was an evolution of an already existing E2E model, which was built using the 3-fold augmented SAVAS, Albayzin, and Multext corpora for 28 epochs. This model reached a WER of 7.2% on a 4 hours test set of the SAVAS corpus.

For this challenge, it was evolved for 2 new epochs using the same corpora in addition to the 3-fold augmented nearly perfectly aligned corpus obtained from the RTVE2018 dataset (see Table 3). A total of 897 hours were used for training. The LM was a 5-gram trained with the text data from the open dataset, with an *alpha* value of 0.8 and a beam-width of 1000 during decoding.

4.2.2. Contrastive systems

The first constrastive system was called 'Vicomtech-PRHLT_c1-E2_open' and as the primary system, it was based on the previously explained E2E architecture. This system was also an evolution of the already existing E2E model, but in this case, it was evolved for one epoch using the 3-fold augmented SAVAS, Albayzin, Multext, nearly perfectly aligned partition and Youtube RTVE corpora. The duration of the total amount of training audios was 1488 hours. The LM was a 5-gram trained with the text data from the open dataset, with an *alpha* value of 0.8 and a beam-width of 1000 during decoding.

The second contrastive system was composed by a bidirectional LSTM-HMM acoustic model combined with a 3-gram LM for decoding and a 9-gram LM for re-scoring lattices. The AM was evolved for 10 epochs, with an initial and final learning rate of 0.0006 and 0.00006 respectively, using a mini-batch size of 100 and 20,000 samples per iteration, and it was trained with the same data as the primary system of the *open condition*. The LMs were estimated with the text data from the open dataset.

5. Results

The results obtained over the *development set* shown in Table 3 are presented in the following Table 5. The *development set* is composed by audio segments from all the TV shows included in the original RTVE2018 dataset and lasts a total of 4 hours.

Table 5: *WER results for each submitted system over the generated development set*

type	*system*	*cond.*	*WER*
P	Vicomtech-PRHLT_p-K1_closed	Closed	**22.6**
C1	Vicomtech-PRHLT_c1-K2_closed		22.8
C2	Vicomtech-PRHLT_c2-E1_closed		26.6
P	Vicomtech-PRHLT_p-E1_open	Open	20.7
C1	Vicomtech-PRHLT_c1-E2_open		**20.5**
C2	Vicomtech-PRHLT_c2-K1_open		22.0

5.1. Processing time and resources

The decodings of the 6 recognition systems were performed on an Intel Xeon CPU E5-2683v4 2.10 GHz 4xGPU server with 256GB DDR4 2400MHz RAM memory. Each GPU corresponds to an NVIDIA Geforce GTX 1080 Ti 11GB graphics acceleration card.

The following Table 6 presents the processing time and computational resources needed by each submitted system for the decoding of the released *test set* of almost 40 hours of audios. It should be noted that the LSTM-HMM based systems were decoded using CPU cores, whilst the E2E systems took advantage of the GPU cards.

Table 6: *Processing time and computational resources needed by each submitted system*

system	*RAM*	*CPU cores*	*GPU*	*Time*
Vicom-PRHLT_p-K1_close	12GB	20	-	24h
Vicom-PRHLT_c1-K2_close	12GB	20	-	24h
Vicom-PRHLT_c2-E1_close	4GB	8	7GB	12h
Vicom-PRHLT_p-E1_open	5GB	8	7GB	8h
Vicom-PRHLT_c1-E2_open	5GB	8	7GB	9h
Vicom-PRHLT_c2-K1_open	19GB	12	-	40h

6. Conclusions

In this paper, the ASR systems submitted to the IberSPEECH-RTVE Speech to Text Transcription Challenge 2018 have been presented. In the beginning, one of the most costly task was the processing of the released RTVE2018 dataset, since a high number of transcriptions were imperfect or do not fit exactly to the related spoken audio. Furthermore, the type of contents posed a notable difficulty to the task, given that the TV shows included most of the main challenges for any speech recognition engine, including spontaneous speech, accents, noise backgrounds, and/or overlapped speakers, among others. Thus, the cleaning process of the dataset became a crucial task to exploit the data correctly.

Looking at the results obtained on the internally generated *development test* and presented in Table 5, it can be clearly deduced that LSTM-HMM based systems performed better when fewer training data were available. In fact, the primary system in the *closed condition* achieved an error of 4 percentage points lower than the E2E based second contrastive system. In this condition, it is also remarkable how the primary system, trained with nearly correctly aligned audios, achieved better results than the first contrastive LSTM-HMM based system, which was built with perfectly aligned contents, even if the primary system included more training data. It suggests that in this case, exploiting more data although they were not aligned exactly, helped systems to perform better.

In the *open condition*, the E2E based systems achieved better results than the LSTM-HMM based one. It could be expected since more training data were available to train models. Even if the first contrastive system obtained a slightly better performance than the primary one, a qualitative evaluation of the results gave as the intuition that the primary system was more robust against spontaneous speech. In this sense, the *alpha* value (0.8), which defines the weight of the LM against the AM, of the E2E systems were lower than the *alpha* value (1.5) employed in the E2E system of the *closed condition*, given that the AM performed better and the global system obtained higher precision, especially with spontaneous speech.

Finally, it should be remarked that all the error rates achieved in this work are lower or at least are in the range of the reference WER values given in the evaluation plan. These WER values were obtained by commercial ASR systems over one TV show in the dataset, and ranged between 22% and 27% of word error rate.

7. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *International Conference on Machine Learning*, 2016, pp. 173–182.

[3] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML’14. JMLR.org, 2014, pp. II–1764–II–1772.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[5] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Advances in neural information processing systems*, 2015, pp. 577–585.

[6] L. Lu, X. Zhang, and S. Renals, “On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition,” *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5060–5064, 2016.

[7] Y. Miao, M. Gowayyed, and F. Metze, “Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding,” in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 167–174.

[8] R. Collobert, C. Puhrsch, and G. Synnaeve, “Wav2letter: an end-to-end convnet-based speech recognition system,” *arXiv preprint arXiv:1609.03193*, 2016.

[9] H. Liu, Z. Zhu, X. Li, and S. Satheesh, "Gram-ctc: Automatic unit selection and target decomposition for sequence labelling," *arXiv preprint arXiv:1703.00096*, 2017.

[10] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Nahamoo, "Direct acoustics-to-word models for english conversational speech recognition," *arXiv preprint arXiv:1703.07754*, 2017.

[11] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 2010, pp. 242–264.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[14] S. Braun, D. Neil, and S.-C. Liu, "A curriculum learning method for improved noise robustness in automatic speech recognition," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 548–552.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[16] A. del Pozo, C. Aliprandi, A. Álvarez, C. Mendes, J. P. Neto, S. Paulo, N. Piccinini, and M. Raffaelli, "Savas: Collecting, annotating and sharing audiovisual language resources for automatic subtitling." in *LREC*, 2014, pp. 432–436.

[17] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J. Pardo, and A. Rubio, "Development of spanish corpora for speech research (albayzin)," in *Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy*, 1991, pp. 26–28.

[18] E. Campione and J. Véronis, "A multilingual prosodic database," in *Fifth International Conference on Spoken Language Processing*, 1998.

[19] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011, pp. 187–197.

[20] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.