IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

The Vicomtech Speech Transcription Systems for the Albayzín-RTVE 2020 Speech to Text Transcription Challenge

Aitor Álvarez, Haritz Arzelus, Iván G. Torre, Ander González-Docasal

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia – San Sebastián (Spain)
[aalvarez,harzelus,igonzalez,agonzalezd]@vicomtech.org

Abstract

This paper describes the Vicomtech's submission to the Albayzín-RTVE 2020 Speech to Text Transcription Challenge, which calls for automatic speech transcription systems to be evaluated in realistic TV shows.

A total of 4 systems were built and presented to the evaluation challenge, considering the primary system along to three constrastive systems. These recognition engines are different versions, evolutions and configurations of two main architectures. The first architecture includes an hybrid DNN-HMM acoustic model, where factorized TDNN layers with and without initial CNN layers were trained to provide posterior probabilities to the HMM states. The language model for decoding correspond to modified Kneser-Ney smoothed 3-gram model, whilst a RNNLM model was used in some systems for rescoring the initial lattices. The second architecture was based on the Quartznet architecture proposed by Nvidia with the aim of building smaller and ligther ASR models with SOTA-level accuracy. A modified Kneser-Ney smoothed 5-gram model was employed to re-score the initial hypothesis of this E2E model. The results obtained for each TV program in the final test set are also presented in addition to the hardware resources and computation time needed by each system to process the released evaluation data.

Index Terms: albayzín evaluations, speech recognition, deep learning, convolutional neural networks, recurrent neural networks.

1. Introduction

The Albayzín-RTVE 2020 Speech to Text Transcription Challenge calls for Automatic Speech Recognition (ASR) systems that are robust against realistic TV shows. Currently, it is a notable trend that aims to approach ASR technology to automate different applications such as subtitling or metadata generation for archive. 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 can 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.

During the last years, ASR systems have positively evolved at acoustic modeling with the integration of DNNs in combination with Hidden Markov Models (HMMs) to outperform traditional approaches [1]. More recently, new attempts have been focused on building E2E ASR architectures [2], which directly map the input speech signal to character sequences and therefore greatly simplify training, fine-tuning and inference [3, 4, 5, 6]. Additionally, deep Transformer or LSTM-RNN based language models have shown better performance than the traditional n-gran models specially during the rescoring of the initial lattices [7].

Driven by the need to reduce the size and complexity of the ASR models, new architectures have recently arisen to make these models lighter, faster and more feasible to deploy on hardware with limited computation capabilities while maintaining the SOTA-level accuracy. In this sense, based on the Jasper architecture [8], Nvidia proposed Quartznet [9], a new E2E neural acoustic model composed of multiple blocks with residual connections in between. Each block consists of one or more modules with 1D time-channel separable convolutional layers, batch normalization, and ReLU layers. They reached near-SOTA error rates on the well-known LibriSpeech [10] and WSJ [11] datasets with models containing fewer than 20 million parameters, in contrast to other larger E2E architectures such as 5x3 Jasper (201 millions) [8] , Deep Speech 2 (38 millions) [2] or Wav2Vec2.0 (95 to 317 millions) [12].

Our systems were built following both DNN-HMM and Quartznet E2E architectures basis, in order to compare the performance of both systems trained with the same corpora as well as their feasibility to be deployed in different platforms, from high-performance servers to embedded systems like Nvidia's Jetson, Google's Coral or Intel's Movidious, among others. We presented a total of 4 ASR engines to the evaluation challenge; three systems based on DNN-HMM hybrid acoustic models, and one system constructed following the E2E Quartznet architecture.

The remainder of this paper is organised as follows: Section 2 describes the corpora used to train the systems; in Section 3 we describe the different speech transcription systems built for the challenge and Section 4 presents the results on the final evaluation test set, in addition to the number of resources and processing time needed per system to process the whole test set. Finally, Section 5 draws the main conclusions.

2. Corpus description

Since in this edition of the Albayzín-RTVE 2020 Speech to Text Transcription Challenge the *open training* condition was only considered, different corpora were used and mixed to train the acoustic and language models.

2.1. Acoustic corpus

The acoustic corpus was composed by annotated audio contents from 7 different datasets, summing up a total of 743 hours and 35 minutes. The following table 1 presents the final number of

10.21437/IberSPEECH.2021-22

hours containing only speech in each of the datasets.

Table 1: *Duration of the speech segments for each dataset*

dataset	duration
RTVE2018	112 h. 30 min.
SAVAS	160 h. 58 min.
IDAZLE	137 h. 8 min.
A la Carta	168 h. 29 min.
Common Voice	158 h. 9 min.
Albayzin	5 h. 33 min.
Multext	0 h. 47 min.
Total	743 h. 35 min.

The *RTVE2018* dataset [13] 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. This dataset was originally composed by 569 hours and 22 minutes of audio with a high portion of imperfect transcriptions and, thus, they could not be used as such for training. Therefore, a forced-alignment was applied in order to recover only the segments transcribed with a high literality, obtaining a total of 112 hours and 30 minutes of nearly correctly transcribed speech segments.

The *SAVAS* corpus [14] is composed of broadcast news contents in Spanish from 2011 to 2014 of 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 *IDAZLE* corpus is integrated by TV shows from the EiTB broadcaster as well, and it comprises a more varied and rich collection of programs of different genres and styles. TV shows are also the contents which compose the *A la Carta*[1] acoustic corpus, including 265 contents broadcasted between 2018 and 2019 by RTVE.

The *Common Voice* dataset [15] is a crowdsourcing project started by Mozilla to create a free and massively-multilingual speech corpus to train speech recognition systems. Finally, the well-known and clean *Albayzin* [16] and *Multext* [17] datasets were also included, mainly to favour the initial training steps and alignments of the systems.

2.2. Text corpus

Regarding text data, different sources were employed to obtain the enough language and domain coverage as close as possible to the contents of the challenge. The following Table 2 presents the number of words provided by each of the text corpus.

Table 2: *Description of the text corpus*

corpus	#words
Transcriptions	7,946,991
RTVE2018	56,628,710
A la Carta	106,716,060
Wikipedia	489,633,255
Total	660,925,016

A total of almost 661 million words were thus compiled and used to estimate the language models for decoding and rescoring purposes. The *Transcriptions* text corpus corresponded to the text transcriptions of the all audio contents used to train the acoustic models. The *RTVE2018* text corpus contains all the text transcriptions and re-spoken subtitles included within the RTVE2018 dataset, whilst the *A la Carta* corpus is integrated by subtitles taken from the "A la Carta" web portal, as a result of a collaboration between RTVE and Vicomtech. Finally, the *Wikipedia* corpus contained texts of the Wikipedia portal gathered in 2017 from Wikimedia[2].

3. Systems description

Two main architectures were employed to build the 4 systems presented to the Albayzín-RTVE 2020 Speech to Text Transcription Challenge: a hybrid DNN-HMM acoustic model for 3 of the systems, and Nvidia's Quartznet architecture for the final system.

3.1. DNN-HMM based systems

The DNN-HMM based systems were built through the *nnet3* DNN setup of the Kaldi recognition system [18], and using the so-called *chain* acoustic model based on optional Convolutional Neural Network (CNN) layers and a factorised time-delay neural network (TDNN-F) [19] which reduces the number of parameters of the network by factorising the weight matrix of each TDNN layer into the product of two low-rank matrices.

Two types of DNN-HMM acoustic models were constructed. The first architecture integrated a CNN-TDNN-F based network, with 6 CNN layers followed by 12 TDNN-F layers. Meanwhile, the second architecture integrated a TDNN-F model and consisted of 16 TDNN-F layers. In both systems, the internal cell-dimension of the TDNN-F layers was of 1536, with a bottleneck-dimension of 160 and a dropout schedule of '0,0@0.2,0.5@0.5,0'. The number of training epochs was set to 4, with a learning rate of 0.00015 and a minibatch size of 64. The input vector corresponded to a concatenation of 40 dimensional high-resolution MFCC coefficients, augmented through speed (using factors of 0.9, 1.0, and 1.1) [20] and volume (with a random factor between 0.125 and 2) [21] perturbation techniques, and the appended 100 dimensional iVectors.

The presented *primary* system was a CNN-TDNN-F based engine with a 3-gram LM for decoding and a 4-gram pruned RNNLM model used for lattice-rescoring following the work presented in [22]. The 3-gram LM was trained with texts coming from the *Transcriptions, RTVE2018* and *A la Carta* corpora presented in Table 2, and the 4-gram pruned RNNLM model was estimated adding the *Wikipedia* text corpus as well.

The first *contrastive* system included the same language models for decoding and rescoring as the *primary* system. The difference relied on the acoustic model, which corresponded to a TDNN-F based network without the initial CNN layers. Finally, the second *contrastive* system was a CNN-TDNN-F based system with a 3-gram LM for decoding and without applying any lattice-rescoring process.

3.2. Quartznet architecture based system

The E2E architecture, which corresponded to the submitted third *contrastive* system, was based on a 5x5 Quartznet system [9] which is completely based on 1D Time-Channel Separable Convolutional layers with residual connections. This design is based on the Jasper architecture [8] but with many modifications focused on considerably reducing the number of parameters and therefore, the computing resources needed.

[1] https://www.rtve.es/alacarta/

[2] https://dumps.wikimedia.org/

Initially there is a 1D Convolutional layer (kernel(k)= 33, output channels (c)= 256) processing the spectrogram input followed by five blocks with residual connections between blocks. Each block is composed of a module repeated five times. Each one of this modules are sequentially composed of (i) a k-sized depthwise convolutional layer, (ii) a pointwise convolution, (iii) a batch normalization layer, and (iv) a ReLU. The configuration of each Block is: B1 ($k = 33$, $c = 256$), B2 ($k = 39$, $c = 256$), B3 ($k = 51$, $c = 512$), B4 ($k = 63$, $c = 512$) and B5 ($k = 75$, $c = 512$). Finally there are three aditional convolutional layers: C1 ($k = 87$, $c = 512$), C2 ($k = 1$, $c = 1025$) and C3 ($k = 1$, $c = labels$). A Connectionist Temporal Classification (CTC) loss function is used for measuring prediction errors and Novograd optimizer with betas 0.8 and 0.5 is used for training with 100 epochs cosine annealing learning rate policy. Initial learning rate was set to 0.015, and minimum to 10^{-5}, weight decay was 10^{-3} and training dataset was computed on three GPUs with batch size of 40 each and mixed precision. Our resulting $5x5$ Quartznet network configuration contains $6, 7$ million parameters.

Additionally, a 5-gram external language model, trained with the *Transcriptions*, *RTVE2018* and *A la Carta* corpora, was used during inference for rescoring the initial hypothesis by using Beam Search CTC Decoder with a beam-width of 1000, $\alpha = 1.2$ and $\beta = 0$. It is worth mentioning that in the previous DNN-HMM based systems, we could not use a 5-gram as LM for decoding due to the lack of memory resources to generate such a large graph with Kaldi.

4. Results and resources

In the following Table 3, the total WER values are presented for each submitted system over all the TV programs in the test set.

Table 3: *Total WER results per system on the whole Albayzin-RTVE 2020 testset*

type	*system*	*tWER*
P	Vicomtech_p-CNN_TDNN_Rescoring	**19.27**
C1	Vicomtech_c1-TDNN_Rescoring	19.98
C2	Vicomtech_c2-CNN_TDNN	19.83
C3	Vicomtech_c3-Quartznet	28.42

As it can be appreciated in Table 3, the *primary* system was correctly selected by the participants since it was the system with the best performance. Likewise, as expected, the Quartznet based experimental system achieved the worse results, even though the quality reached seems promising considering the resources and computing time needed for inference, in addition to the lightness of its E2E model. The robustness of the this E2E model could be improved by adding more training data.

It is also worth noting how the first *contrastive* system performs worst than the second *contrastive* even though the initial results were rescored by a RNNLM model in the former. It seems that, in this case, the acoustic model, which integrated CNN convolutional layers, helped more the ASR engines than rescoring the initial lattices at language model level. It could make sense for this test set, since most of the evaluation contents included spontaneous speech and our rescoring language models were trained with formal language gathered mostly from the *Wikipedia* encyclopedia.

In Table 4, the total WER results obtained by the *primary* system for each TV program in the Albayzin-RTVE 2020 test set are described. The behaviour of the *primary* system regarding the content profiles is similar as expected. In those programs with clean speech, the WER decreases significantly compared to other programs which included adverse acoustic conditions, overlapping or spontaneous speech. More specifically, in TV shows such as *Aquí la Tierra*, *Los desayunos de TV*, *Millenium* and *Versión Española*, with controlled acoustic conditions (*studio*) and long segments with dictate and well-structured speech, the word error rates are below the 20% border in all cases. In contrast, in more complicated contents to process automatically like *Cómo nos Reíamos*, *Imprescindibles* or *Wake-up*, which include many segments with spontaneous and acted speech, acoustically adverse conditions and overlapping, the results degrade appreciably.

Table 4: *Total WER on each test TV program of the Albayzin-RTVE 2020 testset by the primary system*

TV Programs	*tWER*
Aquí la Tierra	16.48
Boca Norte	37.94
Bajo la Red	33.31
Comando Actualidad	24.68
Como nos Reíamos	48.53
Ese Programa del que Usted me Habla	25.67
Imprescindibles (live recordings)	34.45
Los desayunos de TV	10.11
Mercado central	17.83
Millennium	15.98
Never Films Mira Ya	24.21
Si Fueras Tu	29.31
Vaya Crack	19.96
Versión Española	18.10
Wake-Up	33.96
Global	**19.27**

Nevertheless, the global total WER reached 19.27%, an interesting mark considering the difficulty of the test set and that the same engines and models were used within each ASR system to transcribe such different contents from each other in terms of domains, acoustic conditions and speech type. It could have been interesting to check if using different language models fine-tuned to specific domains (debates, news, comedy shows, etc.) and applying them to the corresponding type of contents, the results would have improved.

4.1. Processing time and resources

The decoding processes of the 4 transcription systems were performed on an Intel Xeon CPU E5-2683v4 2.10 GHz 7xGPU server with 256GB DDR4 2400MHz RAM memory. The GPU used for decoding corresponds to an NVIDIA Geforce GTX 1080 Ti 11GB graphics acceleration card.

The following Table 5 presents the processing time and computational resources needed by each submitted system for the decoding of the released *evaluation set* of 55.9 hours of audios. It should be noted that the DNN-HMM based systems were decoded using one CPU core, whilst the Quartzent E2E systems took advantage of a GPU card. In terms of Real-Time Factor (RTF), while the Kaldi-based *primary* system achieved a 0.98 of RTF, the Quartznet based engine reached 0.13 of RTF to process the whole evaluation contents.

Table 5: *Processing time and computational resources needed by each submitted system*

system	*RAM (GB)*	*CPU cores*	*GPU (GB)*	*Time*
p-CNN_TDNN_Rescoring	6.7	1	-	55h
c1-TDNN_Rescoring	6.7	1	-	49h
c2-CNN_TDNN	5.9	1	-	39h
c3-Quartznet	6	1	9.9	7.5h

5. Conclusions

Vicomtech submitted 4 transcription systems; three systems based on more traditional Kaldi's DNN-HMM based engines, using a 3-gram LM for decoding and a RNNLM for lattices rescoring, and a more experimental last system inspired on an optimization of the Nvidia's Quartznet E2E architecture, which aims to be deployed in embedded systems with remarkably accurate results.

As expected, the error rates of the three former systems were notably lower comparing to the E2E based model. However, the participants were motivated to evaluate this novel architecture, designed to be lighter than traditional ASR engines, in order to check their robustness in the same training and evaluation conditions. Nowadays, as larger neural networks with more layers and parameters are built, reducing their complexity and computational cost has becoming critical, specially in real-time applications and scenarios. In addition, the evolution of embedded systems with high computational capacities, triggered great opportunities for researchers to face fundamental challenges in deploying deep learning systems for portable devices with limited resources.

6. 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. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6874–6878.

[8] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *arXiv preprint arXiv:1904.03288*, 2019.

[9] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6124–6128.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[11] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[13] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[14] 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.

[15] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[16] 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.

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

[18] 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.

[19] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[20] 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.

[21] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5929–5933.