applied sciences

MDPI

Article

Evaluating Novel Speech Transcription Architectures on the Spanish RTVE2020 Database †

Aitor Álvarez [1,*,‡], Haritz Arzelus [1,‡], Iván G. Torre [1,‡] and Ander González-Docasal [1,2]

[1] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain; harzelus@vicomtech.org (H.A.); igonzalez@vicomtech.org (I.G.T.); agonzalezd@vicomtech.org (A.G.-D.)

[2] Department of Electronics Engineering and Communications, University of Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, Spain

* Correspondence: aalvarez@vicomtech.org

† This paper is an extended version of our paper published in IberSPEECH2020, Valladolid, Spain, 24–25 March 2021 .

‡ These authors contributed equally to this work.

Abstract: This work presents three novel speech recognition architectures evaluated on the Spanish RTVE2020 dataset, employed as the main evaluation set in the Albayzín S2T Transcription Challenge 2020. The main objective was to improve the performance of the systems previously submitted by the authors to the challenge, in which the primary system scored the second position. The novel systems are based on both DNN-HMM and E2E acoustic models, for which fully- and self-supervised learning methods were included. As a result, the new speech recognition engines clearly outperformed the performance of the initial systems from the previous best WER of 19.27 to the new best of 17.60 achieved by the DNN-HMM based system. This work therefore describes an interesting benchmark of the latest acoustic models over a highly challenging dataset, and identifies the most optimal ones depending on the expected quality, the available resources and the required latency.

Keywords: automatic speech recognition; deep learning; Spanish; convolutional neural networks; recurrent neural networks; embedded systems; quartznet; Wav2vec2.0; self-supervised learning

Citation: Álvarez, A.; Arzelus, H.; Torre, I.G.; González-Docasal, A. Evaluating Novel Speech Transcription Architectures on the Spanish RTVE2020 Database. *Appl. Sci.* **2022**, *12*, 1889. https://doi.org/10.3390/app12041889

Academic Editors: Francesc Alías, Valentín Cardeñoso-Payo, David Escudero-Mancebo, César González-Ferreras and António Joaquim da Silva Teixeira

Received: 29 December 2021
Accepted: 2 February 2022
Published: 11 February 2022

1. Introduction

The Albayzín-RTVE 2020 Speech to Text Transcription Challenge (http://catedrartve.unizar.es/s2tc2020.html, accessed on 28 December 2021) called for Automatic Speech Recognition (ASR) systems that were robust against realistic TV shows. Nowadays, applying speech to text technologies to the broadcast domain is a growing trend that aims to approach ASR technology to automate different applications such as subtitling or metadata generation for archives. Although most of this work is still performed manually or through semiautomatic methods (e.g., re-speaking), the current state of the art (SoA) in speech recognition suggests that this technology can be exploitable autonomously without or with minor post-edition effort, mainly on contents with optimal audio quality and clean speech conditions. The use of Deep Learning algorithms along with the increasingly availability of speech data 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.

Besides the broadcast domain, the significant increase in the ASR field has brought special interests to integrate this technology in many other applications and devices. For instance, considering speech as the most natural means of communication between humans, conversational assistants have acquired great relevance in our daily lives, both in the personal and professional environments [1]. In addition, other main sectors such as Industry, Healthcare or Automotive have already discovered the usability of speech technologies

mainly with the use of voice control applications integrated in machines, medical instruments or technical devices. These interests have triggered special challenges for the current ASR technology, mainly related to the need to optimise and reduce neural models in order to be integrated in devices with low computational power but without a noticeable loss of quality. With the aim of meeting the requirements of embedded systems, the most common optimisation techniques rely on architecture and format optimisation as well as quantisation [1].

In addition to the compression of neural models, another big topic in the ASR field today corresponds to the use of large pre-trained models generated with self-supervised learning methods and a big amount of unlabelled data. The main idea behind is that these models can provide a solution in those challenging scenarios in which the amount of labelled data are scarce in any language or domain. Self-supervised models like Wav2vec2.0 [2] and HuBERT [3] learn first from a large scale unlabelled speech corpus in different languages. The knowledge acquired during this initial training is usually then transferred to downstream tasks by using the model as a feature extractor or by fine-tuning it with in-domain labelled data in the target language.

In this paper, we explore and evaluate three novel ASR architectures on the Spanish RTVE2020 database [4]. The first system is based on the Kaldi toolkit [5], and it was constructed on the novel Multistream CNN architecture designed for robust acoustic modelling in speech recognition tasks. This architecture processes input speech with diverse temporal resolutions by having stream-specific dilation rates to convolutional neural networks (CNNs) across multiple streams. The second system presented in this paper is based on the NVIDIA's Quartznet E2E architecture [6], designed by the need to reduce the size and complexity of the ASR E2E models and to make them lighter, faster and more feasible to deploy on embedded hardware while maintaining the SoA-level accuracy. The Quartznet architecture is 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 normalisation, and ReLU layers. Finally, for the last architecture presented, we explored the application of the self-supervised E2E model Wav2vec2.0 proposed by Facebook AI [2] as the pretrained model. This model was then fine-tuned with the same in-domain labelled data employed to train the first two ASR systems described above. The Wav2vec2.0 model encodes speech audio via a multi-layer convolutional neural network and then it masks spans of the resulting latent speech representation, which are fed to a Transformer network to build contextualised representations.

Throughout the paper, we compared the performance of these three novel architectures with the ones obtained by the primary system and the third contrastive system previously presented by the authors [7] to the Albayzín-RTVE 2020 Speech to Text Transcription Challenge. The primary system, which reached the second position in the competition, was trained with the Kaldi toolkit and included a hybrid CNN-TDNN-F acoustic model, whilst the third contrastive system was based on a lighter Quartznet acoustic model [6]. The comparison presented in this work demonstrated not only an evolution and improvement of similar architectures, such as the DNN-HMM or Quartznet systems, but also the inclusion of a new architecture based on self-supervised learning methods evaluated over the same evaluation set. All the systems were built by using the same speech labelled data to train and/or fine-tune the corresponding acoustic models. Likewise, we used the same text data to construct the different language models to perform the decoding and rescoring processes of the initial hypothesis.

The remainder of this paper is organised as follows: Section 2 introduces recent advances in speech recognition technology. Section 3 describes the corpora used to train and evaluate the systems, whilst in Section 4 we describe the different speech transcription systems, including the three novel architectures and the two systems previously submitted to the challenge. Section 5 presents the results of the systems on the same evaluation set of the RTVE2020 database, in addition to the number of resources and processing time needed

per system to process the whole test set. Finally, Section 6 draws the main conclusions and the lines of future work.

2. Recent Advances in Speech Recognition

During the last few years, ASR systems have positively evolved at acoustic modelling with the integration of DNNs in combination with Hidden Markov Models (HMMs) to outperform traditional approaches [8]. More recently, new attempts have been focused on building E2E ASR architectures [9], which directly map the input speech signal to character sequences and therefore greatly simplify training, fine-tuning and inference [10–14]. Once the potential of the E2E architectures for speech recognition was demonstrated [9], and considering the need to reduce the size and complexity of the ASR models due to their large hardware requirements, new architectures arose to make these models more optimal to be deployed on embedded hardware without loss of quality. More recently, driven by the increasing availability of data in major languages and the scarcity of annotated data in minority languages or specific domains, novel approaches have emerged focused on training big neural models through self-supervised learning methods and making use of hundreds of thousands of unlabelled acoustic data. Nowadays, most of the efforts in the field seem to be focused on this last direction, given the availability of pre-trained models and their high performance when being used as a feature extractor or when they are fine-tuned with in-domain data [15].

Nevertheless, the more conventional hybrid acoustic models also continue to evolve, improving benchmark results over well-known datasets like LibriSpeech [16]. Within the Kaldi community, a novel neural network architecture was recently presented, known as Multistream CNN [17]. The Multistream CNN acoustic model was inspired in the work presented in [18] but leaving out the multi-headed self-attention layers. This model processes input speech with diverse temporal resolutions by having stream-specific dilation rates to CNNs across multiple streams to achieve the robustness. A factorised time-delay neural network (TDNN-F) is stacked in each stream, while the dilation rate for the TDNN-F layers in each stream is chosen from multiples of 3 of the default sub-sampling rate for both training and decoding. The output embedding vectors from each stream are concatenated and followed by ReLU, batch normalisation and a dropout layer, which is finally projected to the output layer via a couple of fully connected layers. This hybrid model achieved very competitive results on the Librispeech test-clean/other sets in combination with the efficient *self-attentive SRU* [19] language model. Specifically, WER values of 1.75 and 4.46 were reported on the test-clean/other partitions of the Librispeech dataset, respectively [19].

With the aim of building lighter but competitive E2E models for speech recognition, NVIDIA proposed Quartznet [6], based on the main Jasper architecture [20]. This architecture consists of 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 normalisation, and ReLU layers. They reached near-SoA error rates on the LibriSpeech dataset, for which WER values of 2.69 and 7.25 were achieved on the test-clean/other sets [6], respectively. These results were reached with the model 15 × 5 Quartznet, which contained 18.8 million parameters, in contrast to other larger E2E architectures such as 10 × 5 Jasper (333 million) [20], PaddlePaddle Deep-Speech2 (49 million) [9] or Wav2Vec2.0 (95 to 317 million) [2].

Nowadays, many of the recent works in the ASR field seem to be focused on taking advantage of big acoustic models trained with self-supervised learning methods and a large amount of unlabelled data. Facebook AI demonstrated for the first time that learning powerful representations through self-supervised methods from big amounts of unlabelled speech and then fine-tuning on transcribed speech can enhance neural models trained with semi-supervised techniques [2]. Their experiments demonstrated the great potential of using large pre-training models estimated with unlabelled data and the impact of employing a different amount of labelled data for fine-tuning the model. For instance, using a large model pre-trained on 60,000 h and fine-tuned on 10 min of labelled training

data, they reached Word Error Rates (WER) of 4.8 and 8.2 on test-clean/other of Librispeech, respectively, using a Transformer language model (LM). Increasing the labelled data to 100 h and using the same LM, they achieved a WER of 2.0 and 4.0 on the same test set.

Nevertheless, self-supervised approaches are challenging mainly because there is not a predefined lexicon for the input sound units during the pre-training phase. Moreover, sound units can be variable in length since an explicit segmentation is not provided [21]. In order to deal with these problems, Facebook AI released HuBERT (Hidden-Unit BERT), a new approach for learning self-supervised speech representations [3]. Unlike Wav2vec2.0, the HuBERT model learns not only acoustics, but also language models from continuous inputs. First, the model encodes unmasked audio inputs into meaningful continuous latent representations, which map to the classical acoustic modelling problem. In a next step, the model captures the long-range temporal relations between learned representations to reduce the prediction error. This work mainly focuses on the consistency of the *k*-means mapping from audio inputs into discrete targets, which enables the model to focus on modelling the sequential structure of input data. HuBERT has shown that it can overcome other SoA approaches on speech representation learning for speech recognition. In the work presented in [21], WER values of 1.8 and 2.9 were reported on test-clean/other of the LibriSpeech dataset, respectively, using a Transformer LM.

On the other hand, at the language model level, deep Transformers or LSTM-RNN based language models have shown better performance than the traditional *n*-gram models especially during the re-scoring of the initial lattices [22]. More recently, the novel network architecture Self attentive simple Recurrent Unit (SRU) shows interesting improvements on the rescoring of the initial hypothesis [19].

The ASR engines presented in this work were built following the hybrid DNN-HMM, Quartznet, and Wav2vec2.0 architecture basis, in order to compare the performance of the systems trained with the same corpora as well as their feasibility to be deployed in different platforms, from high-performance servers to embedded systems.

3. Corpora Description

In this section, the training and evaluation corpora are described in detail. With respect to the training corpora, all the systems presented in this work shared the same acoustic and text data to train and/or fine-tune the initial models. On the other hand, the evaluation data are composed entirely of the RTVE2020 database [4].

3.1. Training Corpora

3.1.1. Acoustic Corpus

The acoustic corpus was composed by annotated audio contents from seven different datasets, summing up a total of 743 h and 35 min. The following Table 1 presents the final number of hours containing only speech in each of the datasets.

Table 1. Duration of the speech segments for each dataset with their corresponding licenses.

Dataset	Duration	License
RTVE2018	112 h 30 min	Non-Commercial
SAVAS	160 h 58 min	Commercial/Research
IDAZLE	137 h 8 min	Non-Commercial
A la Carta	168 h 29 min	Non-Commercial
Common Voice	158 h 9 min	Mozilla Public License 2.0
Albayzin	5 h 33 min	Commercial/Academic
Multext	0 h 47 min	Commercial/Academic
Total	743 h 35 min	

The RTVE2018 dataset [23] 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 h and 22 m 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 h and 30 min of nearly correctly transcribed speech segments.

The *SAVAS* corpus [24] 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* (https://www.rtve.es/alacarta/, accessed on 28 December 2021) the acoustic corpus, including 265 contents broadcasted between 2018 and 2019 by RTVE.

The *Common Voice* dataset [25] 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* [26] and *Multext* [27] datasets were also included, mainly to favour the initial training steps and alignments of the systems.

3.1.2. 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 RTVE2020 database. The following Table 2 presents the number of words provided by each of the text corpora.

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 corresponds 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 contains texts of the Wikipedia portal gathered in 2017 from Wikimedia (https://dumps.wikimedia.org/, accessed on 28 December 2021).

3.2. Evaluation Data

RTVE2020 Database

The RTVE2020 database served as the principal evaluation test of this work and the Albayzín Speech To Text Transcription Challenge 2020. It is composed of a series of TV shows of different genres which were broadcast by the public Spanish Television (RTVE) from 2018–2019. The database is composed of a total of 55 h and 40 min of audio, and it was fully transcribed by humans to obtain literal references. These references are presented in STM format, which contain time marked segments each including the waveform's filename and channel identifier, the speaker, the begin and end time, optional subset label and the literal transcription of the segment. The type of TV shows included in the database are presented in Table 3.

Table 3. TV shows included in the RTVE2020 dataset. This information was partially gathered from [4].

TV Program	Duration	Description
Ese programa del que Ud. habla	01:58:36	A TV program that reviews daily political, cultural, socialand sports news from the perspective of comedy.
Los desayunos de RTVE	10:58:34	The daily news, politics, interviews and debate program.
Neverfilms	00:11:41	A webseries that parody humorously trailers of series and movies well-known to the public.
Si fueras tú	00:51:14	Interactive series that tells the story of a young girl.
Bajo la red	00:59:01	A youth fiction series whose plot is about a chain of favours on the internet.
Comando actualidad	04:01:31	A show that presents a current topic through the choral gaze of several street reporters.
Boca norte	01:00:46	A story of young people who dance to the rhythm of trap.
Wake-up	00:57:28	A story that combines science fiction, a post-apocalyptic Madrid and lots of action inspired in video games.
Versión española	02:29:12	Program dedicated to the promotion of Spanish and Latin American cinema.
Aquí la tierra	10:26:02	A magazine that deals with the influence of climatology and meteorology both personally and globally.
Mercado central	08:39:47	A Spanish soap opera set in a today's Madrid market.
Vaya crack	05:06:00	A contest where contestants take multiple quiz designed to test their abilities in several disciplines.
Cómo nos reímos	02:51:42	A program dedicated to the great comedians and their work on RTVE programs.
Imprescindibles	03:12:31	A documentary series on the most outstanding figures of Spanish culture in the 20th century.
Millennium	01:56:11	Debate show for the spectators of today, accompanying them in the analysis of everyday events.
Total duration	**55:40:16**	

As it can be observed in the description of the programs shown in Table 3, most of the TV shows include content with spontaneous speech, thus significantly increasing the difficulty of automatically transcribing this database. Despite the fact that the database is not correctly balanced with respect to the duration of each TV program, most of them share similar artefacts typical of informal speech.

Los desayunos de RTVE and *Aquí la tierra* are the two shows with the longest duration in the database. Although in the former some segments with orderly and formal speech can be found, it also includes interviews and political debates. Similarly, *Aquí la tierra* include formal and informal speech segments as well, combining weather reports usually performed by one speaker with interviews in which spontaneous speech is much more present. The soap opera *Mercado central* is the third in duration, and its main difficulty is related to the acted and emotional speech, which ASR engines are not usually trained with. Nevertheless, probably the most complex contents correspond to *Cómo nos reímos*, *Vaya crack* and *Ese programa del que usted me habla*, which together sum up a total of almost ten hours in the database. These contents are composed entirely of spontaneous speech, including sketches, overlapping speech, artistic performances, laughs, fillers, among other artefacts. *Comando actualidad* adds the difficulty of including spontaneous interviews and reports on the street, whilst *Imprescindibles* incorporates multi-channel and far-field low quality recordings. Finally, programs like *Millenium* and *Versión española*, which sum up to less than five hours, expose a priori a lower difficulty, since they include content with more formal speech, although the latter could integrate parts of movies.

In summary, the RTVE2020 database is a challenging evaluation set of Spanish TV shows in which the types of content that generate the greatest problems to the ASR engines today predominate.

4. Systems Description and Configuration

In this section, all the systems presented in this work are described in more detail. In the first subsection, the three novel ASR architectures are described. These architectures include systems built on the top of (1) the Kaldi toolkit, (2) the Quartznet Q15×5 architecture of NVIDIA, and (3) the Wav2vec2.0 model.

On the other hand, the two systems selected as baseline systems are presented in the second subsection. These baseline systems correspond to the DNN-HMM based *primary* system presented to the Albayzín S2T Challenge 2020, which scored the second position in the competition, and the third contrastive system based on the Quartznet Q5×5 architecture.

4.1. Novel ASR Architectures

4.1.1. Multistream CNN Based System

The Multistream CNN based ASR engine was built on the top of the Kaldi toolkit through the *nnet3* DNN setup and following the `egs/librispeech/s5/local/chain/run_multistream_cnn_1a.sh` recipe from the ASAPP Research repository (https://github.com/asappresearch, accessed on 28 December 2021).

The Multistream CNN architecture is illustrated in Figure 1. The acoustic model is composed by an initial set of five 2D-CNN layers in charge of processing the given input speech frames augmented dynamically through the SpecAugment [28] technique. Each embedding vector outputted from the single-streamed set of CNN layers in each time step is then inserted as the input of each of the three stacks of TDNN-F layers, combined with a dilation rate configuration of `6-9-12`. Each stack is composed of 17 TDNN-F layers, with an internal cell-dimension of 512, a bottleneck-dimension of 80 and a dropout schedule of `'0,0@0.20,0.5@0.5,0'`. The number of training epochs was set to six, with an initial and final learning rates of 10^{-3} and 10^{-5}, respectively, and a mini-batch 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) [29] and volume (with a random factor between 0.125 and 2) [30] perturbation techniques, and the appended 100 dimensional *i*-Vectors.

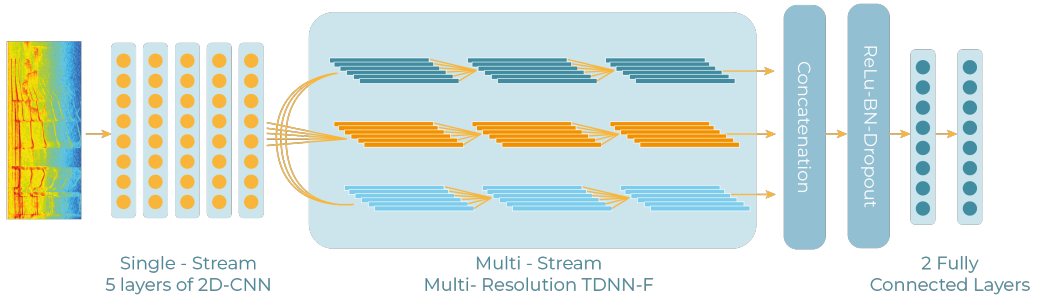

Figure 1. Architecture of the Multistream CNN acoustic model composed by an initial set of five 2D-CNN layers followed by three stacks of 17 TDNN-F layers. This figure is inspired by the one presented in [17].

This system included a 3-gram language model for decoding and a 4-gram pruned RNNLM model for lattice-rescoring following the work presented in [31]. 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.

4.1.2. Quartznet Q15×5 Based System

The Quartznet family of models are E2E ASR architectures completely based on 1D Time-Channel Separable Convolutional layers with residual connections, as it is illustrated in Figure 2. This design is based on the Jasper architecture [20] but with many modifications focused on considerably reducing the number of parameters and, therefore, the computing resources needed. As the most novel architecture, we trained and evaluated the Q15×5 architecture, which is described as follows.

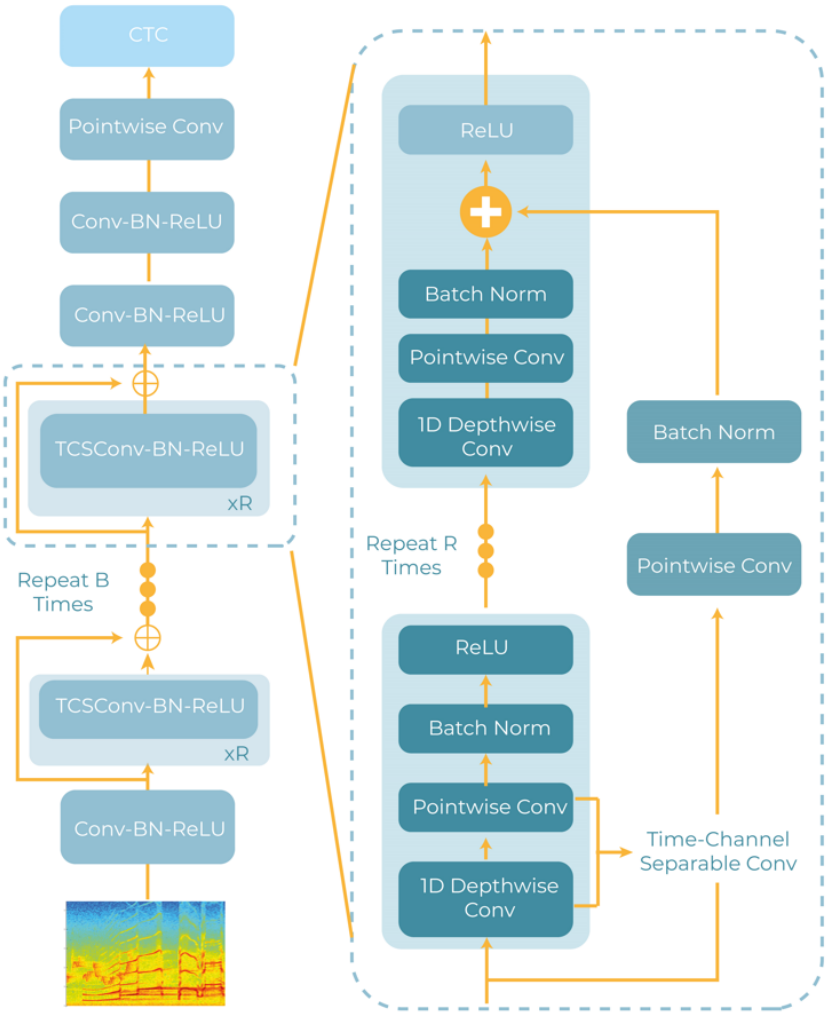

Figure 2. Architecture of the Quartznet ASR model of NVIDIA. The input spectrogram of the audio signal is first processed by a 1D CNN layer and then fed into a series of five groups of blocks. Each group is composed of three blocks with residual connection between them in the Q15×5 architecture and one block for the Q5×5 architecture. Blocks are formed by a module composed of a Time-Channel Separable CNN repeated five times. Finally, a sequence of 3 CNN feed the output to a trainable CTC layer. This figure is based on the one presented in [6].

The model initially integrates a 1D Convolutional layer (kernel(k) = 33, output channels (c) = 256) processing the spectrogram input from each speech frame. This layer is then followed by five groups of blocks. On the Q15×5 architecture, each group of block is composed of a block B_i repeated three times with residual connections in between. Each B_i is composed of a module repeated five times and composed of (i) a k-sized depth-wise convolutional layer, (ii) a point-wise convolution, (iii) a batch normalisation layer, and (iv) a ReLU. The configuration of the CNN for each B_i was: 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 additional convolutional layers C1 ($k = 87$, $c = 512$), C2 ($k = 1$, $c = 1025$) and

a point-wise convolutional layer ($k = 1$, $c = labels$), followed by a Connectionist Temporal Classification (CTC) layer.

During training, a CTC loss function was employed to measure the prediction errors, in addition to the Novograd optimiser with beta values of 0.8 and 0.5 and a triangular cyclical learning rate policy during five cycles of 60 epochs for a total of 300 epochs, as it was described in [32]. The initial and minimum learning rates were set to 0.015 and 10^{-5}, respectively, whilst the weight decay was set to 10^{-3}. The training process was performed on four GPU cards, applying a batch size of 20 each and mixed precision training. Our resulting Q15×5 network configuration contained 18.9 million parameters.

Additionally, a 5-gram language model was trained with the *Transcriptions*, *RTVE2018* and *A la Carta* corpora. This language model was employed during decoding with a CTC beam-search decoder based on the `pyctcdecode` library [33]. The decoding was performed with a *beam-width* of 256, a $\alpha = 0.5$ and $\beta = 1$.

4.1.3. Wav2vec2.0 Based System

Wav2vec2.0 [2] is a self-supervised E2E architecture based on CNN and Transformer layers schematically represented in Figure 3. The Wav2vec2.0 model maps speech audio through a multi-layer convolutional feature encoder $f : \chi \rightarrow Z$ to latent speech representations $z_1, \ldots, z_T$, which are fed into a Transformer network $g : Z \rightarrow C$ to output context representations $c_1, \ldots, c_T$. These context representations are then quantised to $q_1, \ldots, q_T$ to represent the targets in the self-supervised learning objective [2,34]. The feature encoder contains seven blocks and the temporal convolutions in each block include 512 channels with strides $(5, 2, 2, 2, 2, 2, 2)$ and kernel widths $(10, 3, 3, 3, 3, 2, 2)$. The Transformer used was composed by 24 blocks, a model dimension of 1024, an inner dimension of 4096 and a total of 16 attention heads.

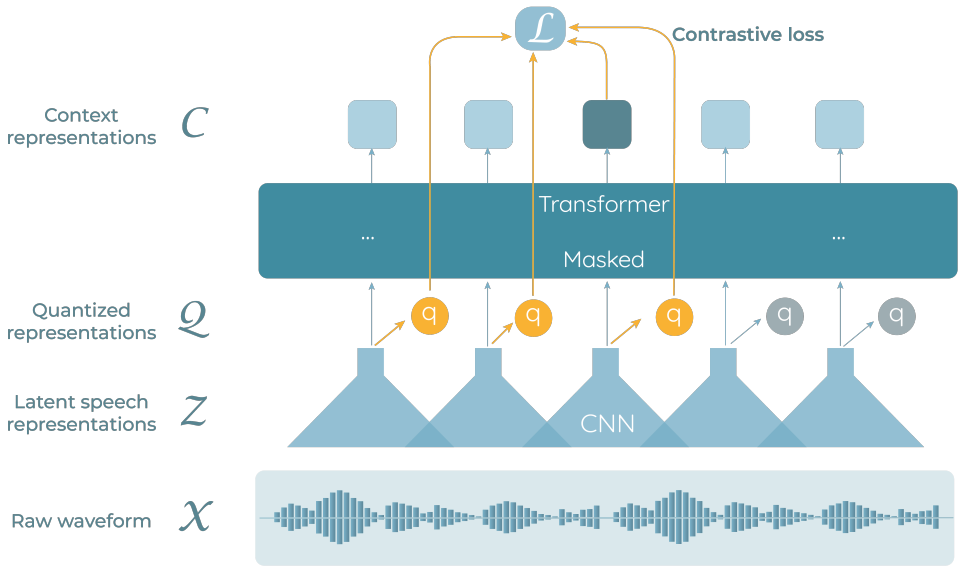

Figure 3. Wav2vec2.0 architecture representation. The raw audio signal is mapped to speech representations that are fed into a transformer network to output context representations. Context representations are then quantised to represent targets in the self-supervised task. This figure is based on the one presented in [2].

As the main baseline Wav2vec2.0 model, in this work, we selected the pretrained Wav2Vec2-XLS-R-300M model (https://huggingface.co/facebook/wav2vec2-xls-r-300m, accessed on 28 December 2021), which was self-supervised pre-trained with 436k h of unlabelled speech data in 128 languages from the VoxPopuli [35], MLS [36], Common-Voice [25], BABEL (corpus collected under the IARPA BABEL research program), and

VoxLingua107 [37] corpora. The Wav2Vec2-XLS-R-300M is one of the different versions of the Facebook AI's XLS-R multilingual model [38] composed by 300 million of parameters.

With the aim of adapting this pre-trained model to the domain, we fine-tuned the Wav2Vec2-XLS-R-300M model for a total of 50 epochs in two steps, and, using the acoustic corpus defined above, augmented dynamically during training through the SpecAugment technique. First, we evolved the pre-trained model for 30 epochs with a maximum learning rate of 10^{-4}. We used the cosine function as the learning rate annealing function with a warm up during the initial 10% of the training, whilst the batch-size was set to 256. This model was later fine-tuned for 20 additional epochs by modifying the maximum learning rate to 5×10^{-5}.

Finally, the decoding was performed with the same 5-gram language model employed for the decoding of the previous Q15×5 based system, but with a *beam-width* of 256, $\alpha = 0.3$ and $\beta = 1.6$.

4.2. Baseline ASR Architectures

4.2.1. DNN-HMM Based System

This first baseline system that we consider in this work corresponds to the primary system [7] presented in the Albayzín-RTVE 2020 Speech to Text Transcription Challenge. This ASR engine was built through the *nnet3* DNN setup of the Kaldi recognition system, and using the so-called *chain* acoustic model based on Convolutional Neural Network (CNN) layers and a factorised time-delay neural network (TDNN-F) [39], 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.

The acoustic model integrated a CNN-TDNN-F based network, with six CNN layers followed by 12 TDNN-F layers. 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 1.5×10^{-4} and a mini-batch 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) [29] and volume (with a random factor between 0.125 and 2) [30] perturbation techniques, and the appended 100-dimensional iVectors.

This system included a 3-gram language model for decoding and a 4-gram pruned RNNLM model for lattice-rescoring following the work presented in [31]. 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.

4.2.2. Quartznet Q5×5 Based System

This ASR engine was presented as the third contrastive system to the Albayzín-RTVE 2020 Speech to Text Transcription Challenge, and we selected it as the second baseline system in this work in order to compare its performance with the previously presented Quartznet Q15×5 based system.

The Quartznet Q5×5 architecture is similar to the Q15×5 architecture explained in Section 4.1.2, but, in this case, each group of block is composed of a block B_i repeated only one time, instead of three, thus decreasing the total number of parameters from 18.9 to 6.7 M.

In contrast to the Q15×5 based system, the training of the Q5×5 was performed for 100 epochs, with a cosine annealing learning rate using a batch-size of 40.

The decoding was realised using the same 5-gram language model described above. The parameters of the Beam Search CTC decoder corresponded to a *beam-width* of 1000, $\alpha = 1.2$ and $\beta = 0$.

5. Results and Resources

In the following Table 4, the total WER values are presented for each system over all the TV programs in the RTVE2020 database.

Table 4. Total WER results per system on the whole Albayzin-RTVE 2020 testset. The results of the novel and evolved ASR engines are marked in bold.

AM Architecture	Model	Type	WER
DNN-HMM	Multistream CNN	novel	**17.60**
	CNN-TDNN-F	baseline	19.27
Quartznet	Q15×5	novel	**22.95**
	Q5×5	baseline	28.42
Wav2vec2.0	Wav2Vec2-XLS-R-300M (fine-tuned)	novel	**20.68**

The first objective of this work was focused on improving our initially best CNN-TDNN-F based ASR engine, presented as the primary system to the Albayzín-RTVE 2020 S2T Transcription Challenge and which scored the second position among all the systems presented to competition. Although these improvements could be focused on improving both acoustic and language models, we decided to put our efforts into enhancing the acoustic model, which probably poses the most challenging task, while maintaining the same Kaldi based ASR architecture. Moreover, one of the main conclusions of the previous study of the authors [7] was that improving acoustic models helped our ASR engines more than language models, since most of the evaluation contents included spontaneous speech and our text corpus was mainly integrated by text contents with formal language. Given the difficulty of the RTVE2020 evaluation dataset at the acoustic and phonemic level, we decided to evaluate a more complex CNN and TDNN-F based neural network. This way, the Multistream CNN architecture clearly improved the performance of the CNN-TDNN-F acoustic model by tripling the stack of TDNN-F layers with diverse temporal resolutions. Therefore, replacing the CNN-TDNN-F acoustic model by the novel Multistream CNN one and maintaining the same lexicon and language models for decoding (3-gram model) and rescoring (4-gram based RNNLM model) the initial lattices, the results improved from 19.27 to a very competitive 17.60 of WER.

The second objective of this work was to see how we could improve the performance of the Nuance's Quartznet Q5×5 based system. However, analysing the state of the art, currently these E2E systems do not perform as well as the Kaldi based engines, the Quartznet architectures present an interesting proposal to integrate ASR functionalities in embedded systems considering the scarce HW resources and the low inference times required. With the aim of improving the results obtained by the Q5×5 based system, we included three principal improvements. First, we extended the architecture by adding two more groups of blocks and thus building a bigger Q15×5 architecture based ASR system. The second improvement corresponded to the inclusion of a triangular cyclical learning rate policy. This method [32] lets the learning rate vary cyclically between reasonable boundary values instead of monotonically decreasing it during training, thus improving classification accuracy. Finally, we included more training epochs by extending the 100 epochs employed for the Q5×5 network to the 300 training epochs used for the new Q15×5 acoustic model. Applying these evolutions, we managed to improve the error rate in 5.47 points, considering the 28.42 of WER achieved by the Q5×5 baseline and the new 22.95 of WER obtained by the Q15×5 based novel architecture. The 5-gram based external language model remained the same for both systems.

The last objective of this study was to evaluate the performance of the self-supervised Wav2vec2.0 model in these challenging scenarios. Although these types of models seem to be more focused on being applied in situations in which there is not enough annotated in-domain data to train acoustic models from scratch, its advantages such as (i) the clarity of the architecture, (ii) taking advantage of the extensive acoustic knowledge obtained by the pre-trained model, and (iii) its low latency in inference, make it a very interesting ASR system to explore. To this end, we selected the Wav2Vec2-XLS-R-300M pre-trained model, which was pre-trained with 436k hours of unlabelled speech data from diverse corpora

and conditions, as it was described in Section 4.1.3. In this case, it is worth mentioning how a fine-tuned Wav2Vec2-XLS-R-300M model improved the performance of the Q15×5 model, even though the former was fine-tuned for only 50 epochs and the Quartznet model was trained for a total of 300 epochs. This way, the Wav2Vec2-XLS-R-300M model fine-tuned with the in-domain data reached a very promising 20.68 of WER. This result clearly demonstrates the power of the self-supervised models trained with huge amounts of unlabelled data, while maintaining similar inference latencies in comparison with the lighter Quartznet architectures (see Table 4). For the language model, the same 5-gram based external language model was employed.

In Table 5, the total WER results obtained by the systems for each TV program in the RTVE2020 dataset are presented.

Table 5. Total WER of the ASR systems on each TV program of the RTVE2020 test set.

TV Program	Multistream CNN	CNN-TDNN-F	Q15×5	Q5×5	Wav2vec2.0
Ese programa del que Ud. habla	23.64	25.67	29.65	36.15	26.81
Los desayunos de RTVE	9.26	10.11	12.14	14.68	11.08
Neverfilms	19.81	24.21	29.03	37.82	28.05
Si fueras tú	24.57	29.31	36.76	46.73	36.43
Bajo la red	22.41	33.31	32.99	41.06	32.33
Comando actualidad	22.58	24.68	27.34	32.70	25.6
Boca norte	32.07	37.94	43.16	52.92	40.37
Wake-up	30.87	33.96	40.81	47.71	38.19
Versión española	16.14	18.10	19.15	25.66	18.06
Aquí la tierra	14.90	16.48	19.69	24.68	17.67
Mercado central	16.44	17.83	25.43	34.05	21.91
Vaya crack	19.22	19.96	28.43	30.16	20.80
Cómo nos reíamos	46.17	48.53	54.33	61.41	53.20
Imprescindibles	30.44	34.45	37.12	44.94	29.52
Millenium	16.02	15.98	17.30	18.82	17.57
Global	**17.60**	19.27	**22.96**	28.42	**20.68**

As it can be observed in Table 5, the systems perform consistently along all the contents in the RTVE2020 dataset. The Multistream CNN based system obtained the best results for all the contents except for *Millenium*, for which the baseline CNN-TDNN-F system performed the best, and for *Imprescindibles* where Wav2vec2.0 performed slightly better. In contrast, the Quartznet Q5×5 is the system that gives the worst performance in all the contents. The rest of the systems maintain consistency for most TV programs with respect to each particular total WER. The Wav2vec2.0 based system outperforms the Quartznet Q15×5 based ASR engine in all cases except one (*Millenium*), although the difference of 0.27 can be considered as negligible. In the rest of the contents, the differences between these two E2E systems are remarkable, although the Wav2vec2.0 system was fine-tuned with six times fewer epochs than the Quartznet Q15×5 based system. The greatest differences between both E2E systems are achieved in contents *Vaya crack* and *Imprescindibles*, which were classified as two of the most complex programs given their spontaneous style and far-field low quality, respectively.

In general, the behaviour of the systems regarding the content profiles is as expected. In those programs with cleaner 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 RTVE*, *Millenium* and *Versión Española*, with controlled acoustic conditions (studio) and many segments with formal and well-structured speech, the error rates are below the 20% border and lower

compared to the other TV shows. In contrast, in more complicated contents like *Cómo nos Reíamos*, *Boca Norte* or *Wake-up*, which include many segments with spontaneous and acted speech, acoustically adverse conditions and overlapping, the results degrade appreciably in all ASR engines.

Processing Time and Resources

The decoding processes of the transcription systems were performed on an Intel Xeon CPU E5-2683v4 2.10 GHz 7xGPU server with 256 GB DDR4 2400 MHz RAM memory. The GPUs used for decoding correspond to an NVIDIA Geforce GTX 3090Ti 24 GB graphics acceleration card.

The following Table 6 presents the processing time and computational resources needed by each ASR system for the decoding of the 55 h and 40 min of the RTVE2020 dataset. It should be noted that the first four systems were decoded using only one CPU core, whilst the Wav2vec2.0 based system took advantage of 40 CPU cores.

Table 6. Processing time and computational resources needed by each submitted system in terms of RAM memory, CPU cores, GPU memory, decoding time and Real-Time Factor (RTF).

ASR System	RAM (GB)	CPU Cores	GPU (GB)	Time	RTF
Multistream CNN	6.7	1	12	4.9 h	0.09
CNN-TDNN-F	6.7	1	12	4.9 h	0.09
Quartznet Q15×5	6	1	12	8 h	0.14
Quartznet Q5×5	6	1	12	7.5 h	0.13
Wav2vec2.0	30	40	18	8 h	0.14

In terms of RAM memory, the Kaldi based and the Quartznet based architectures occupied a similar memory space, mainly related to the size of the language model. In contrast, the Wav2vec2.0 based system took much more memory space since the pre-trained model was composed of 300 million parameters. Even though the Multistream CNN based system outperforms the CNN-TDNN-F in quality (see Table 5), it is worth noting that it took almost as long to decode the whole RTVE2020 database, even though the acoustic model was much bigger and the decodings were performed using only one CPU core. It could be sped up by segmenting the original waveform through a Speech Activity Detection (SAD) module and processing the segments in parallel using several CPU threads. However, the SAD module could have additionally introduced new errors when performing this segmentation that would have irreparably impacted the final result. Finally, it should be mentioned how competitive the latency of the Wav2vec2.0 based system was (RTF of 0.14) despite the big size of the model.

6. Conclusions and Future Work

In this work, three novel ASR architectures have been presented and evaluated on the RTVE2020 database. These systems correspond to an evolution of two similar ASR engines, previously constructed and evaluated on the same database and which were selected as the baseline systems.

The RTVE2020 database is the result of an interesting and necessary initiative to collect real broadcast Spanish speech data with the aim of building competitive ASR engines in this language. In conjunction to the RTVE2018 database [40], both constitute the larger, most complete and challenging speech corpus available for the community in the Spanish language.

Over this interesting and challenging dataset, we explored different alternatives to outperform the initial DNN-HMM and Quartznet Q5×5 based systems submitted to the Albayzín S2T Transcription Challenge 2020. In total, we presented three novel ASR engines, which clearly improved the performances of these baseline systems. The first system, based

on the novel Multistream CNN acoustic model, reached the best results for almost all the contents, while the Quartznet Q15×5 outperformed the Q5×5 model at almost six points of WER, by extending the size of the model, including more training epochs and applying a triangular cyclical method for the optimal learning rate calculation. In addition, we also evaluated the performance of the Wav2vec2.0 self-supervised model, which achieved better results than the Quartznet based systems applying a fine-tuning process of only 50 epochs over the pre-trained model. In summary, this work constitutes an interesting and complete benchmark of several architectures in order to select the optimal ASR engine depending on the required quality, the available HW resources and the latency expected.

As future work, once the acoustic models have been considerably enhanced, we will focus on improving the language models, by incorporating new rescoring processes based on Transformers or Self attentive simple Recurrent Unit architectures. It should also be interesting to explore strategies to reduce the size of the Multistream CNN acoustic model, by reducing the amount of TDNN-F layers and/or the internal cell dimensions without loss of quality. These techniques would probably reduce the size of the model and would allow for increasing the inference times. Moreover, bigger Wav2vec2.0 pre-trained models will also be explored, which contained one and two billion parameters, with longer fine-tuning processes in order to allow the model to better adapt to the application domains. Finally, new self-supervised models like HuBERT [3] and SUPERB [41] will be also studied over the same dataset.

Author Contributions: Conceptualisation, A.Á.; methodology, A.Á.; software, H.A. and I.G.T.; validation, A.Á., H.A. and I.G.T.; formal analysis, A.Á. and I.G.T.; investigation, A.Á., H.A. and I.G.T.; resources, A.Á., H.A. and I.G.T.; data curation, H.A.; writing—original draft preparation, A.Á.; writing—review and editing, A.Á., H.A., I.G.T., A.G.-D.; visualisation, A.Á.; supervision, A.Á.; project administration, A.Á.; funding acquisition, A.Á. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Eunate Yániz for preparing the figures.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RTVE	Corporación de Radio y Televisión Española, S. A.
ASR	Automatic Speech Recognition
SoA	State of the Art
DNN	Deep Neural Network
E2E	End-To-End
CNN	Convolutional Neural Networks
HMM	Hidden Markov Model
TDNN-F	Factorised Time-Delay Neural Network
SRU	Self attentive simple Recurrent Unit
WER	Word Error Rate
LM	Language Model
AM	Acoustic Model
HW	Hardware
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Model
EiTB	Euskal Irrati Telebista
MFCC	Mel-Frequency Cepstral coefficients
GPU	Graphical Processing Unit
SAD	Speech Activity Detection
RTF	Real-Time Factor

References

1. Georgescu, A.L.; Pappalardo, A.; Cucu, H.; Blott, M. Performance vs. hardware requirements in state-of-the-art automatic speech recognition. *Eurasip J. Audio Speech Music. Process.* **2021**, *2021*, 1–30. [CrossRef]
2. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv* **2020**, arXiv:2006.11477.
3. Hsu, W.N.; Tsai, Y.H.H.; Bolte, B.; Salakhutdinov, R.; Mohamed, A. HuBERT: How much can a bad teacher benefit ASR pre-training? In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–12 June 2021; pp. 6533–6537.
4. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. Rtve2020 Database Description. 2020. Available online: http://catedrartve.unizar.es/reto2020/RTVE2020DB.pdf (accessed on 28 December 2021).
5. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
6. Kriman, S.; Beliaev, S.; Ginsburg, B.; Huang, J.; Kuchaiev, O.; Lavrukhin, V.; Leary, R.; Li, J.; Zhang, Y. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6124–6128.
7. Alvarez, A.; Arzelus, H.; Torre, I.G.; González-Docasal, A. The Vicomtech Speech Transcription Systems for the Albayzın-RTVE 2020 Speech to Text Transcription Challenge. In Proceedings of the IberSPEECH 2021, Valladolid, Spain, 24–25 March 2021; pp. 104–107.
8. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
9. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 173–182.
10. Graves, A.; Jaitly, N. Towards End-to-end Speech Recognition with Recurrent Neural Networks. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing, China, 21–26 June 2014; Volume 32, pp. 1764–1772.
11. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4960–4964.
12. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 577–585.
13. Lu, L.; Zhang, X.; Renals, S. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Shanghai, China, 20–25 March 2016; pp. 5060–5064.
14. Yao, Z.; Wu, D.; Wang, X.; Zhang, B.; Yu, F.; Yang, C.; Peng, Z.; Chen, X.; Xie, L.; Lei, X. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv* **2021**, arXiv:2102.01547.
15. Chang, X.; Maekaku, T.; Guo, P.; Shi, J.; Lu, Y.J.; Subramanian, A.S.; Wang, T.; Yang, S.w.; Tsao, Y.; Lee, H.y.; et al. An exploration of self-supervised pretrained representations for end-to-end speech recognition. *arXiv* **2021**, arXiv:2110.04590.
16. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
17. Han, K.J.; Pan, J.; Tadala, V.K.N.; Ma, T.; Povey, D. Multistream CNN for robust acoustic modeling. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Ontario, Canada, 6–12 June 2021; pp. 6873–6877.
18. Han, K.J.; Prieto, R.; Ma, T. State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 54–61.
19. Pan, J.; Shapiro, J.; Wohlwend, J.; Han, K.J.; Lei, T.; Ma, T. ASAPP-ASR: Multistream CNN and self-attentive SRU for SOTA speech recognition. *arXiv* **2020**, arXiv:2005.10469.
20. Li, J.; Lavrukhin, V.; Ginsburg, B.; Leary, R.; Kuchaiev, O.; Cohen, J.M.; Nguyen, H.; Gadde, R.T. Jasper: An end-to-end convolutional neural acoustic model. *arXiv* **2019**, arXiv:1904.03288.
21. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv* **2021**, arXiv:2106.07447.
22. Wang, Y.; Mohamed, A.; Le, D.; Liu, C.; Xiao, A.; Mahadeokar, J.; Huang, H.; Tjandra, A.; Zhang, X.; Zhang, F.; et al. Transformer-based acoustic modeling for hybrid speech recognition. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6874–6878.

23. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. Albayzin 2018 evaluation: The iberspeech-RTVE challenge on speech technologies for spanish broadcast media. *Appl. Sci.* **2019**, *9*, 5412. [CrossRef]
24. del Pozo, A.; Aliprandi, C.; Álvarez, A.; Mendes, C.; Neto, J.P.; Paulo, S.; Piccinini, N.; Raffaelli, M. SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling. 2014. pp. 432–436. Available online: https://d1wqtxts1xzle7.cloudfront.net/47591506/SAVAS_Collecting_Annotating_and_Sharing_20160728-14677-ftvov1.pdf?1469710441=&response-content-disposition=inline%3B+filename%3DSAVAS_Collecting_Annotating_and_Sharing.pdf&Expires=1644576491&Signature=U30UJaJhWnxGGyeWM1nrbsH8X6OmBzgCFnHR-6~yPtA1zWbi~QwWg9nfIXFGc7-iRF5q4FvKblsf0o5-O665DKoLwhtMpAVwuEX71lTmY9qjRRaSMAa3AfEFKyrCNwuKgEWWRmlkaftiUVOTZRyt8T2S3z9Y0TmDTFtJ7Nyvd4~~096vT3y7sdjgd5j~R54br24q13pXlHh64yJozRV41xGvb76yjYF1~yS3oivFYiP0GjZ1jckXgSRB0WzwVkDkve6JThKuxKwO58VP3~WfRIIb2DUrtKYPO-C8EUkPY7e2ZrpLNTjCuTO0JqfIaXmvdKk4XMn7T0KqZA-~fbHABQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA (accessed on 21 December 2021).
25. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv* **2019**, arXiv:1912.06670.
26. Casacuberta, F.; Garcia, R.; Llisterri, J.; Nadeu, C.; Pardo, J.; Rubio, A. Development of Spanish corpora for speech research (Albayzin). In Proceedings of the Workshop on International Cooperation and Standarization of Speech Databases and Speech I/O Assesment Methods, Chiavari, Italy, 26 September 1991; pp. 26–28.
27. Campione, E.; Véronis, J. A multilingual prosodic database. In Proceedings of the Fifth International Conference on Spoken Language Processing, Sydney, Australia, 30 November–4 December 1998.
28. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
29. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
30. Peddinti, V.; Povey, D.; Khudanpur, S. A time delay neural network architecture for efficient modeling of long temporal contexts. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
31. Xu, H.; Chen, T.; Gao, D.; Wang, Y.; Li, K.; Goel, N.; Carmiel, Y.; Povey, D.; Khudanpur, S. A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5929–5933.
32. Smith, L.N. Cyclical learning rates for training neural networks. In Proceedings of the 2017 IEEE Winter conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 464–472.
33. Kucsko, G. Pyctcdecode. Available online: https://github.com/kensho-technologies/pyctcdecode (accessed on 28 December 2021).
34. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *arXiv* **2020**, arXiv:2006.13979.
35. Wang, C.; Rivière, M.; Lee, A.; Wu, A.; Talnikar, C.; Haziza, D.; Williamson, M.; Pino, J.; Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv* **2021**, arXiv:2101.00390.
36. Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G.; Collobert, R. Mls: A large-scale multilingual dataset for speech research. *arXiv* **2020**, arXiv:2012.03411.
37. Valk, J.; Alumäe, T. Voxlingua107: A dataset for spoken language recognition. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 652–658.
38. Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; von Platen, P.; Saraf, Y.; Pino, J.; et al. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv* **2021**, arXiv:2111.09296.
39. Povey, D.; Cheng, G.; Wang, Y.; Li, K.; Xu, H.; Yarmohammadi, M.; Khudanpur, S. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 3743–3747.
40. Lleida, E.; Ortega, A.; Miguel, A.; Bazán, V.; Pérez, C.; Zotano, M.; de Prada, A. *RTVE2018 Database Description*; Vivolab and Corporación Radiotelevisión Espanola: Zaragoza, Spain, 2018.
41. Yang, S.W.; Chi, P.H.; Chuang, Y.S.; Lai, C.I.J.; Lakhotia, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.T.; et al. SUPERB: Speech processing Universal PERformance Benchmark. *arXiv* **2021**, arXiv:2105.01051.