The Vicomtech-UPM Speech Transcription Systems for the Albayzín-RTVE 2022 Speech to Text Transcription Challenge

Haritz Arzelus[1], *Iván G. Torre*[2 3], *Juan M. Martín-Doñas*[1],
Ander González-Docasal[1 4], *Aitor Álvarez*[1]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia – San Sebastián (Spain)
[2]Departamento de Matemática Aplicada, ETSIAE Universidad Politécnica de Madrid,
Plaza Cardenal Cisneros 3, 28040 Madrid
[3]Language and Speech Laboratory, Universidad del País Vasco, 01006, Vitoria-Gasteiz
[4]University of Zaragoza, Department of Electronics, Engineering and Communications,
Pedro Cerbuna 12, 50009 Zaragoza, Spain

[harzelus,jmmartin,agonzalezd,aalvarez]@vicomtech.org,ivan.gonzalez.torre@upm.es

Abstract

This paper describes the Vicomtech-UPM submission to the Albayzín-RTVE 2022 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 alongside three contrastive systems. Each system was built on top of one different architecture, with the aim of testing several state-of-the-art modelling approaches focused on different learning techniques and typologies of neural networks.

The primary system used the self-supervised Wav2vec2.0 model as the pre-trained model of the transcription engine. This model was fine-tuned with in-domain labelled data and the initial hypothesis re-scored with a pruned 4-gram based language model. The first contrastive system corresponds to a pruned RNN-Transducer model, composed of a Conformer encoder and a stateless prediction network using BPE word-pieces as output symbols. As the second contrastive system, we built a Multistream-CNN acoustic model based system with a non-pruned 3-gram model for decoding, and a RNN based language model for rescoring the initial lattices. Finally, results obtained with the publicly available Large model of the recently published Whisper engine were also presented within the third contrastive system, with the aim of serving as a reference benchmark for other engines. Along with the description of the systems, the results obtained on the Albayzin-RTVE 2020 and 2022 test sets by each engine are presented as well.

Index Terms: albayzín evaluations, speech recognition, deep learning, self-supervised learning, sequence-to-sequence.

1. Introduction

The Albayzín-RTVE 2022 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 in the media, such as subtitling or metadata generation for archive and information retrieval. The use of Deep Learning algorithms in speech processing has made it possible to introduce this technology in such complex scenarios through the use of systems based on Deep Neural Networks (DNNs), currently trained with a huge amount of (non-)labelled acoustic data using different learning techniques.

During the last years, ASR systems have positively progressed in acoustic modelling with the integration of DNNs 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 simplify training, fine-tuning and inference [3, 4, 5, 6]. Nowadays, driven by the increasing availability of data in major languages, novel approaches have emerged on training big neural models through self-supervised learning using hundreds of thousands of unlabelled acoustic data [7]. Today, 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 [8]. Nevertheless, well-known architectures trained with hundreds of thousands of multilingual annotated data that defy the performance of the most novel approaches have emerged as well [9].

Our systems were built following different strategies, learning techniques and neural architectures. The primary system was built on top of the self-supervised Wav2vec2.0 pre-trained model [7], which was fine-tuned with in-domain labelled data and helped by a pruned 4-gram language model for the final hypothesis. The first contrastive system was composed of a pruned RNN Transducer (RNN-T) E2E model [10], which integrated a Conformer [11] encoder and a stateless (non-recurrent) prediction network using Byte-Pair Encoding (BPE) [12] as output symbols. The second contrastive system was constructed on the Multistream CNN architecture [13] designed for robust acoustic modelling, processing input speech with various temporal resolutions by having stream-specific dilation rates to Convolutional Neural Networks (CNNs) across multiple streams. The last contrastive system corresponds to the recently published Whisper ASR engine [9], for which the publicly available Large model was used to decode the results. In addition, some architectures benefited from an initial acoustic segmentation provided by a Voice Activity Detection (VAD) module, based on the GPVAD convolutional recurrent model [14] trained with 5,000 hours from the Audio Set [15] dataset.

The remainder of this paper is organised as follows: Section 2 describes the corpora used for training; in Section 3 we describe the speech transcription systems built for the challenge and Section 4 presents the results on the Albayzín-RTVE 2020 and 2022 test sets. Finally, Section 5 draws the conclusions.

2. Corpora description

In this section, the different acoustic and text corpora employed to train the systems are described in detail.

2.1. Acoustic corpus

The acoustic corpus was composed by annotated audio contents from 9 different datasets, summing up a total of 1,927 hours and 47 minutes, as it is presented in Table 1.

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

dataset	duration
RTVE2018	112 h. 30 min.
SAVAS	160 h. 58 min.
IDAZLE	778 h. 21 min.
RTVE Play 2020	168 h. 29 min.
RTVE Play 2022	296 h. 15 min.
RTVE YouTube	18 h. 6 min.
Common Voice	386 h. 48 min.
Albayzin	5 h. 33 min.
Multext	0 h. 47 min.
Total	1,927 h. 47 min.

The *RTVE2018* dataset [16] 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 originally comprised 569 hours and 22 minutes of audio with a high portion of imperfect transcriptions. Therefore, a forced-alignment was applied to recover only the segments transcribed with a high literalness, obtaining a total of 112 hours and 30 minutes of nearly correctly transcribed speech segments. The *SAVAS* corpus [17] 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 *RTVE Play 2020* and *2022* acoustic corpus, including programs broadcasted between 2018 and 2022 by RTVE. Additionally, we gathered transcribed contents of RTVE from the YouTube platform[1,2] as well.

The *Common Voice* dataset [18] 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* [19] and *Multext* [20] datasets were also included.

2.2. Text corpus

Regarding text data, different sources were employed, as it is presented in number of words in Table 2.

Table 2: *Description of the text corpus*

corpus	#words
Transcriptions	20,299,703
RTVE2018	56,628,710
RTVE Play	241,330,497
Generic news	76,276,831
RTVE news	180,611,376
Total	575,147,117

A total of almost 575.2 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 all audio contents used to train the acoustic models. The *RTVE2018* text corpus contained all the text transcriptions and re-spoken subtitles included within the RTVE2018 dataset, whilst the *RTVE Play* corpus was integrated by subtitles taken from the "RTVE Play"[3] web portal, and the *Generic news* corpus incorporated news gathered from digital newspapers in the Internet. Finally, the *RTVE news* corpus was composed of news collected from the RTVE website[4].

3. Systems description

This section describes each of the neural architectures built for the Albayzín-RTVE 2022 S2T Transcription Challenge.

3.1. Wav2vec2.0 based system

Wav2vec2.0 [7] is a self-supervised E2E architecture based on a CNN feature extractor and Transformer layers for the encoder and decoder. The Wav2vec2.0 model maps speech audio through a multi-layer convolutional feature encoder $f : \chi \rightarrow Z$ to latent speech representations $z_1, \dots, z_T$, which are fed into a Transformer network $g : Z \rightarrow C$ to output context representations $c_1, \dots, c_T$. These context representations are then quantised to $q_1, \dots, q_T$ to represent the targets in the self-supervised learning objective [7]. 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.

As the main baseline Wav2vec2.0 model, for this work we selected the pretrained Wav2Vec2-XLS-R-1B model [21], which corresponds to one of the different versions of the Facebook AI's XLS-R multilingual model [22] composed by one billion of parameters. This model was initially trained through self-supervised learning methods using 436k hours of unlabelled speech data in 128 languages. This data was collected from the VoxPopuli [23], MLS [24], CommonVoice [18], BABEL[5] , and VoxLingua107 [25] corpora.

This Wav2Vec2-XLS-R-1B pre-trained model was then adapted with 300 hours of in-domain data from the *RTVE2018* and *RTVE Play 2020* datasets during 50,000 updates, in which the CTC layer was trained only during the initial $10,000$ steps. The batch-size was set to 50 and the model was trained with a tri-stage learning rate policy, in which after the 10% of the warm-up updates, the learning rate was set to $8 \cdot 10^{-6}$ during the following 40% of the updates, then linearly decaying for the rest of the training. Finally, a pruned 4-gram language model, trained with all the corpora detailed in Table 2 except *RTVE news*, was used for decoding. We employed a Bayesian Optimisation procedure to find the best decoding hyper-parameters over the *dev* partition of the Albayzin-RTVE 2020 dataset. The

[1] https://www.youtube.com/user/rtve/videos

[2] https://www.youtube.com/c/RTVEArchivo/videos

[3] https://www.rtve.es/play/

[4] https://www.rtve.es/

[5] Corpus collected under the IARPA BABEL research program

decoding was performed using a beam-size of 1024, a language model weight of 0.95, a word score weight of 1.27 and a silence weight of -0.18.

3.2. RNN-Transducer based system

RNN-Transducer framework has become very popular in the Industry due to its high accuracy in online and streaming applications [10]. Nevertheless, due to its architecture, its loss function can be relatively slow to compute, making use of a high GPU memory when the vocabulary size is too large. In order to accelerate training, we estimated a pruned RNN-T [10], which reduces the memory usage and speeds up the training process.

The inputs to the neural network were 80-dimensional log Mel-filter banks with a processing window size of 25 ms and a window shift of 10 ms. Speed perturbation with 0.9 and 1.1 factors and SpecAugment [26] techniques were also applied to make the training more robust, whilst the model outputs were 500 word-pieces with BPE as the segmentation algorithm. The training of the model was configured for 40 training epochs with an initial learning rate of 0.003, which started decaying after epoch 6. We used Noam optimiser to learn parameters and it was estimated over all the acoustic corpora shown in Table 1.

The encoder of the RNN-T model was a Conformer [11] with 18 layers and 8 self-attention heads per layer. The attention dimension and the feed-forward dimension were 512 and 2048, respectively. We employed a stateless decoder [27] as the prediction network, which consisted of an embedding layer with a dimension of 512, followed by a 1-D convolutional layer with a kernel size of 2. The decoding was performed on the speech segments generated by our VAD module, applying a beam-size of 16 and without using any external language model.

3.3. Multistream-CNN based system

The Multistream-CNN based ASR engine was built on top of the Kaldi toolkit [28] through the *nnet3* DNN setup. The acoustic model is composed by an initial set of five 2D-CNN layers in charge of processing the given input speech frames dynamically augmented through the SpecAugment [26] 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 6, 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. The acoustic model was trained with all the acoustic corpora described in Table 1.

This system included a non-pruned 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, RTVE Play* and *Generic News* corpora presented in Table 2, and the 4-gram pruned RNNLM model was estimated adding the *RTVE news* text corpus. In order to gain effectiveness in rescoring, the decoding was performed on the speech segments previously generated by our VAD module.

3.4. Whisper based system

Whisper [9] is a recently proposed ASR model that leverages a large amount of weakly labelled audio data to train a multilingual transcription model. The architecture is based on the well-known encoder-decoder Transformer sequence-to-sequence model [32]. This network includes an acoustic encoder fed with log-Mel spectrograms, whose outputs are used for conditioning an auto-regressive text decoder via cross-attention mechanisms. The network is trained with 680k hours of speech data, where only 2.6% represents Spanish audio. Among the released models, we chose the Large version, which has a total of 1.55B parameters.

For long-audio decodings, the model works with 30-second audio chunks, and outputs time-coded text segments. The decoder is pre-conditioned on the previous predictions to keep consistency among segments, with the risk that the network is prone to loop errors given the same results regardless of the encoded audio features. This phenomenon is especially critical in audios with long non-speech segments. To alleviate this issue, we looked at the text outputs of the decoding. For transcriptions where more than 20% of segments were text repetitions, we remade the decoding by first segmenting the audio through our VAD module, discarding non-speech segments longer than 2 seconds. The decoding process of the rest of the speech segments was performed by resetting the pre-condition of the text decoder. In addition, we sought for text phrases repeated three times or more, keeping only one appearance. Garbage outputs such as “subtitle” or “subscribe” were deleted as well. Finally, the output was de-normalised by removing the capitalisation and punctuation marks.

4. Results and resources

In Table 3, the total WER values over the Albayzín-RTVE 2020 and 2022 test sets are presented for each submitted system.

Table 3: *Total WER results per system on the Albayzin-RTVE 2020 (WER_20) and 2022 (WER_22) test sets*

type	*system*	*WER_20*	*WER_22*
P	VICOM-UPM_Wav2vec2.0	13.77	15.30
C1	VICOM-UPM_RNN-T	14.32	**14.78**
C2	VICOM-UPM_Multistream-CNN	17.10	17.29
C3	VICOM-UPM_Whisper-Large	**12.15**	14.87

As it can be observed in Table 3, the Whisper based system obtained the best WER on the Albayzin-RTVE 2020 test set, followed by the Wav2vec2.0 based system, which achieved a very competitive 13.77 of total error rate. The other two systems reached worse results; 14.32 and 17.10 for the RNN-T and Multistream-CNN based systems, respectively. These values were definitive to select the primary and contrastive systems for the 2022 challenge. It is worth mentioning that considering that the engine and recognition models of the Whisper engine were not developed and/or adapted by the authors, we decided to leave this engine as the last contrastive system.

On the other hand, if we observe the results obtained by these systems on the Albayzin-RTVE 2022 test set, we realise that the RNN-T based system managed to generalise better for the new contents, obtaining the best results with a very interesting 14.78 of total error rate. This WER was even better than the one obtained by the Whisper engine, which reached a 14.87 with the Large model after applying the decoding and cleaning strategies explained in subsection 3.4. Our primary system based on the Wav2vec2.0 pre-trained model scored the third position within our submitted systems with a 15.30 of WER, whilst the Multstream-CNN based system reached a performance similar to that obtained with the 2020 test set.

In Table 4, the total error rates obtained per system at word level for each TV show in the 2022 test set are presented.

Table 4: *Total WER of the ASR systems for each TV program of the Albayzín-RTVE 2022 test set. The name of the programs are presented as acronyms.*

TV program	P	C1	C2	C3
3x4	**12.60**	13.37	33.58	14.78
AG	7.70	6.72	6.74	**6.16**
APB	63.15	60.24	**52.14**	67.05
AT	12.78	11.65	13.68	**9.60**
ATE	10.50	9.07	10.96	**7.92**
CA	**18.12**	19.09	21.12	21.30
CCA	12.75	**9.51**	12.08	11.93
CO	10.12	**7.89**	8.68	9.88
CPE	17.14	**13.46**	16.66	14.57
EC	13.90	14.32	15.64	**13.45**
ED	14.09	14.21	16.11	**13.36**
EE	27.87	25.16	29.05	**22.20**
ERA	18.99	19.88	21.17	**18.80**
GR	**24.79**	29.02	32.07	31.08
IU	23.16	19.79	23.21	**19.50**
JYS	11.51	12.04	**10.69**	11.74
NN	10.70	**10.20**	13.01	10.49
RD	**18.18**	23.30	23.81	20.84
SYG	10.35	10.24	10.33	**10.07**
TO	23.57	**20.61**	23.34	24.91
YR	22.36	25.33	28.79	**21.48**
Global	15.30	**14.78**	17.29	14.87

As it can be observed in Table 4, the systems perform consistently along all the contents in the Albayzín-RTVE 2022 test set considering the characteristics and difficulty of each content. It is interesting to observe how the RNN-T based system (C1) obtained the best total WER even though the Whisper based system (C3) achieves better results in more TV shows (10) than the former (6).

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 with adverse acoustic conditions, overlapping or spontaneous speech. More specifically, in TV shows such as AG (Agroesfera), ATE (Ateneo), CCA (Conversatorios en Casa América), JYS (Jara y Sedal) or SYG (Saber y Ganar) with more controlled acoustic conditions and many segments with formal and well-structured speech, the error rates are below the 13% border for all the systems, which demonstrates the good performance of the 4 systems in this type of contents. In contrast, in more complicated TV shows like EE (Entrevistas en estudio), GR (Grasa), TO (Toros) or YR (Yrreal), which include more segments with spontaneous and acted speech, acoustically adverse conditions and overlapping, the results are between 20% and 30% of WER, as expected. Finally, the worse results were achieved with the APB (A pedir de Boca) content, which includes very poor acoustic conditions. Interestingly, it was the C2 system with the worst overall WER which achieved the best result with this program.

In terms of decoding, the transcription hypotheses were computed on two different servers and GPU acceleration cards. The RNN-T, Multistream-CNN and Whisper based systems were run on an Intel Xeon CPU E5-2683v4 2.10 GHz 7xGPU server with 256 GB DDR4 2400 MHz RAM memory, using an NVIDIA Titan RTX 24 GB graphics acceleration card. On the other hand, the decoding of the Wav2vec2.0 based system was performed on an AMD Ryzen 7 5800X 3.8 GHz server with 128 GB DDR4 3200 MHz RAM memory, using an NVIDIA RTX 3090 24GB graphics acceleration card.

The Table 5 presents the processing time and resources needed per system to decode the 54 hours and 2 minutes of media contents from the Albayzín-RTVE 2022 test set.

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

system	*RAM (GB)*	*CPU cores*	*GPU (GB)*	*Time*
VICOM-UPM_Wav2vec2.0	18.5	1	8.5	7.3h
VICOM-UPM_RNN-T	11.6	3	5.6	1.9h
VICOM-UPM_Multistream-CNN	11.9	1	5.7	39.9h
VICOM-UPM_Whisper-Large	8.6	1	11.2	21.7h

As it is shown in Table 5, the Wav2vec2.0 based system was the engine that occupied the most RAM memory (18.5 GB), mainly due to the 1 billion parameters of the pre-trained model. Nevertheless, it was the second fastest system in decoding. It took 7.3 hours to decode the whole test set. Within the 4 submitted systems, the RNN-T based system was the fastest transcription engine. It required only 1.3 hours to generate the hypotheses of the 54 hours and 2 minutes of the test set using 3 CPU cores, which supposed a very competitive Real-Time Factor (RTF) of 0.035. Despite the good quality performance offered by the Whisper based engine, the RTF of this system grew up to 0.40 using the Large model and 1 CPU core and 1 acceleration card only. Finally, the Multistream-CNN based system was the engine with the worst results in terms of quality and performance.

5. Conclusions

In this work, the 4 systems submitted by the Vicomtech-UPM team to the Albayzín-RTVE 2022 S2T Challenge were presented. Although the best results on the previous Albayzín-RTVE 2020 test set were achieved by the Large model of the Whisper based system, the RNN-T based model reached the lowest WER on the 2022 test set with a very interesting and competitive 14.78 of total word error rate. Furthermore, this system was the fastest engine generating the recognition hypotheses, needing less than 2 hours to process more than 54 hours of content.

As future work, the authors plan to continue exploring the possibilities of the RNN-T based system, mainly to be applied in a more challenging real-time or streaming domain. Besides, it would be interesting to fine-tune the Large model of the Whisper based system in order to be adapted to the application domain, in addition to continuing to control unexpected recognition outputs. Regarding the Wav2vec2.0 based systems, new training configurations will be studied, including the possibility of improving the rescoring process with more sophisticated neural language models.

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] 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.

[8] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subramanian, T. Wang, S.-w. Yang, Y. Tsao, H.-y. Lee *et al.*, "An exploration of self-supervised pretrained representations for end-to-end speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 228–235.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," Technical report, OpenAI, 2022. URL https://cdn. openai. com/papers/whisper. pdf, Tech. Rep., 2022.

[10] F. Kuang, L. Guo, W. Kang, L. Lin, M. Luo, Z. Yao, and D. Povey, "Pruned RNN-T for fast, memory-efficient ASR training," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2068–2072. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10340

[11] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[12] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[13] K. J. Han, J. Pan, V. K. N. Tadala, T. Ma, and D. Povey, "Multi-stream cnn for robust acoustic modeling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6873–6877.

[14] H. Dinkel, S. Wang, X. Xu, M. Wu, and K. Yu, "Voice activity detection in the wild: A data-driven approach using teacher-student training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1542–1555, 2021.

[15] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[22] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[23] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[24] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[25] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[27] M. Ghodsi, X. Liu, J. Apfel, R. Cabrera, and E. Weinstein, "Rnn-transducer with stateless prediction network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7049–7053.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.