Delve deep into End-To-End Automatic Speech Recognition Models

Maria Labied
Hassan II University, Ben M'sik Faculty of Sciences, Laboratory of Information Technology and Modeling
Casablanca - Morocco
mr.labied@gmail.com

Abdessamad Belangour
Hassan II University, Ben M'sik Faculty of Sciences, Laboratory of Information Technology and Modeling
Casablanca - Morocco
belangour@gmail.com

Mouad Banane
Hassan II University, Faculty of Legal, Economic and Social Sciences, Laboratory of Artificial Intelligence & Complex Systems Engineering,
Casablanca - Morocco
mouad.banane-etu@etu.univh2c.ma

***Abstract*— Automatic Speech Recognition (ASR) has experienced significant advancements in recent years, with end-to-end approaches emerging as a promising paradigm shift. Unlike traditional ASR systems that rely on a pipeline of separate components, end-to-end models aim to directly transcribe speech inputs into text using deep learning architectures. In this paper, we conduct a comprehensive study on end-to-end ASR models. We review various architectures employed in end-to-end ASR, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models. We explore different training methodologies, loss functions, and optimization algorithms used in end-to-end ASR. Additionally, we discuss large-scale datasets commonly used for training and evaluate the performance of end-to-end models using established evaluation metrics such as word error rate (WER). Furthermore, we analyze the strengths and weaknesses of end-to-end ASR models, highlight their applications in real-world scenarios, and discuss open challenges and future directions. By providing this comprehensive study, we aim to facilitate a deeper understanding of end-to-end ASR models and their potential for driving advancements in speech recognition technology.**

Keywords*— *Automatic Speech Recognition, Transformer, End-to-End, RNN-Transducer, Attention based encoder decoder, Transformer-AED, RNN-AED, Conformer, ContextNet.

I. INTRODUCTION

Deep Neural Network (DNN) speech recognition models, more precisely hybrid Automatic Speech Recognition (ASR) models, have replaced the traditional ASR models, but they still retain all disjoint components of traditional models (fig.1) such as the lexicon, the acoustic, and the language models [1][2]. Speech recognition modeling has made a major leap from hybrid speech recognition models to End-to-End models (E2E)[3][4][5][6][7]. Compared to hybrid models, E2E models are composed of a single block[8] that jointly optimizes acoustic, lexical, and language models simultaneously. With E2E architectures, a single model needs training (fig.1), with the possibility of using only the speech signals and their target transcripts. Without the need for word alignments or lexicon dictionaries to train the models. These E2E models are the most groundbreaking as they overturns all the traditional ASR system modeling components that have been used for so many years. Now, with E2E models, we can directly perform the transcription of an input speech into an output text using just a single neural network, which was impossible with hybrid models. With these E2E architectures, we will be able to train speech recognition models for some languages and dialects that have not been trained due to a lack of data resources, as in the case of speech recognition for the Moroccan dialect 'Darija' [9]. E2E will significantly reduce the time needed to train speech recognition models for different languages and their variants compared to hybrid models.

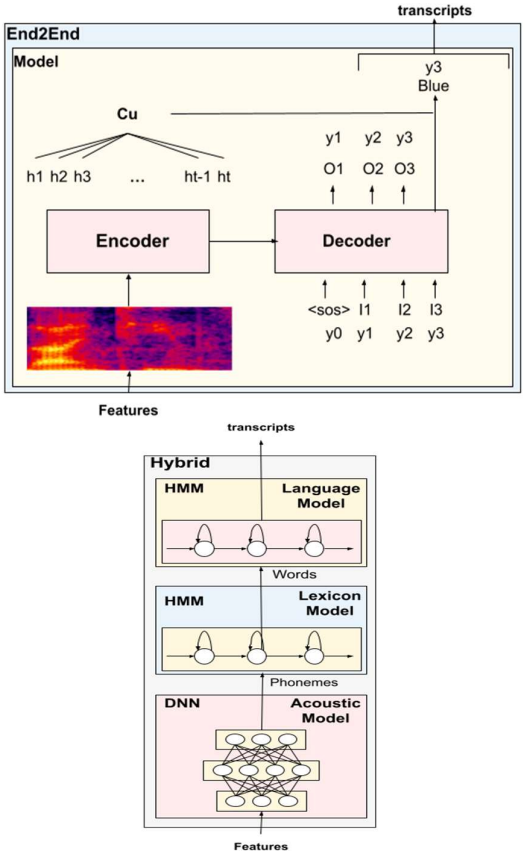

Fig. 1. End-to-End vs Hybrid ASR models

Given the fast evolution of E2E speech recognition approaches, it is opportune to benchmark the most promising and popular E2E models in the ASR field. Some of those widely used E2E approaches there is Recurrent Neural Network-Transducer (RNN-T) [13][14][15], Connectionist Temporal Classification (CTC) [10][11][12], attention-based encoder-decoder (AED) ; mainly RNN-AED and Transformer-AED [16][17]. CTC was the earliest E2E approach that could match the input voice signal to the target tags with no need for external pre-alignments. But it assumes

that frames are independent. RNN-T extends the modeling philosophy of CTC and changes the model architecture and the objective function to consider the dependence of frames, also it was able to replace hybrid models, specifically in streaming contexts [18][19]. AED models were primarily proposed for machine translation [19] but have also been successfully used in speech recognition [20][21][22][3]. Recently, Transformer-AED with self-attention has gained prominence and currently is used as the fundamental block of encoder and decoder models [23].

The rest of this paper is organized as follows: In Section 2, we give an overview of the most popular end-to-end (E2E) speech recognition architectures. In Section 3, we present a benchmarking of the different E2E models in the speech recognition field. Finally, we conclude the paper in Section 4.

II. Speech Recognition E2E models

E2E models have achieved great results in the majority benchmarks tests in terms of ASR accuracy and efficiency, we give in this section an overview of the three popular categories of speech recognition E2E models, namely CTC, RNN-T, and AED which consist of Transformer-AED and RNN-AED. The architecture of these models shares similarities in the encoding part while they differ in the decoding part where each model uses a specific decoding mechanism.

A. Connectionist temporal classification models

CTC models were the earliest E2E approach that has matched the input voice signal to its target tags without needing to align the signal to a reference transcription, by assuming that frames are independent. CTC-based models are popular among the speech recognition community[24][25][26][27] due to their ease of training and efficiency in decoding [28].

CTC was specially designed for temporal classification sequence labeling tasks without knowledge of any prior alignment between input and output sequences[29]. CTC-based models, allow repetition of labels and work by adding a special blank label to distinguish the less informative frames. Also, it removes the state alignment step in training by automatically inferring the frame alignment between speech and label. Meanwhile, CTC E2E models give competitive results when used in conjunction with sequence-to-sequence attention-based models [17] [30][27][31].

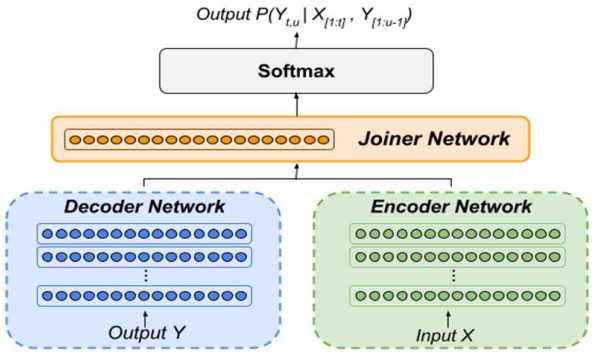

Fig. 2. RNN-T architecture

B. Recurrent Neural Network-Transducer models

RNN-T models are the most popular E2E speech recognition models and the predominant in the ASR industry nowadays [18][32][33][34]. RNN-T is a sequence-to-sequence model that was initially proposed by Graves in 2012 [13]. Since then, no real usage of these RNN-T models was considered, but currently, big attention is going toward these kinds of E2E models after the achievements of google research that have confirmed the low latency and enhancement of speech recognition using the RNN-T transducer [33].

RNN-T model architecture consists of an encoder network, a predictor network, and a joiner network as illustrated in figure 2(fig.2). RNN-T succeeded to removes the conditional independence assumption problem presented in CTC-based models by adding both the predictor and the joiner networks. The encoder network starts by converting the acoustic features in a time step t into a high-level representation H_t^{enc}, then the predictor network, called also the decoder takes the previous outputs as input for predicting in an autoregressive manner a high-level representation of the next output H_u^{pre}. The joiner network is a simple neural network that takes the encoder output H_t^{enc} and the predictor network output H_u^{pre} as input, and then combines them to produce a joint representation matrix $H_{t,u}$. This joint representation is then used to calculate the softmax output (Eq.1).

$$H_{t,u} = f^{joint}(H_t^{enc}, H_u^{pre}) \quad (1)$$

C. Attention-encoder-decoder models

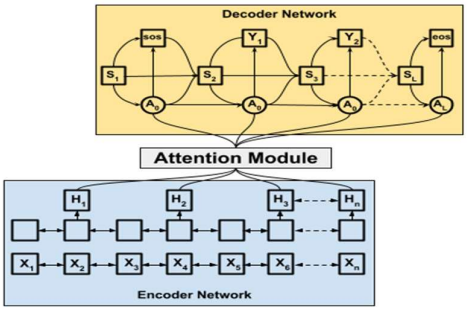

Fig. 3. AED models Architecture

The AED models are another type of E2E ASR model [35] [20] known for their attention structure. This category of E2E models shares the same architecture, mainly consists of an encoder network, an attention module, and a decoder network as illustrated in figure 3(fig.3). The encoder performs the conversion of the input features into a hidden feature sequences. While attention module produces a context vector through calculating the attention weights between the previous decoder output and each frame of the encoder output. The decoding network then uses the preceding output label as well as the context vector to produce its output in an autoregressive manner based on the antecedent label outputs without the conditional independence presumption.

D. RNN-AED

Speech recognition AED-based models are mainly bisecting into two categories. The first one is the **RNN-AED** which uses Long-short-term memory RNN for the encoder and decoder output. The encoder part of RNN-AED models is similar to the encoder part of RNN-T models. However, the decoder is enhanced by the attention mechanism. The attention mechanism is used to calculate a context vector,

which is a representation of the encoder output that is relevant to the current decoder state. The context vector is then used to generate the next token in the output sequence as illustrated in Eq.2.

$$H_u^{dec} = LSTM(C_u, Y_{u-1}, H_{u-1}^{dec}) \quad (2)$$

E. Transformer-AED

The second category of AED models is **Transformer-AED** which is mainly based on the Transformers concept in both encoder and decoder parts. Transformers are known for their long terms dependencies and they outperform the RNNs at this level. In the Transformer AED models, the encoder consists of a stack of Transformer blocks, In which each block has a feedforward layer and a multi-head self-attention layer, where the connection between different layers and blocks is performed using layer normalization [36] and residual connections. A third layer in the decoder part, is used to perform multi-head attention on the encoder output.

Transformer models based on the attention mechanism have been extensively adopted for sequence modeling due to their training efficiency and long-range interaction capturing[37], however they are not very effective at extracting local feature patterns. Recent works show that combining transformers with convolution improves their capabilities compared to using them alone[38]. In [39] a combination of transformer and convolutional neural network have been proposed to benefit from the best of both architectures under the name of Conformer. Transformers with attention learn the global interaction while a convolutional neural network captures local correlations based on the relative offset. Conformer-based models have shown a competitive accuracy compared to existing transformer-based attention end-to-end speech recognition models.

III. BENCHMARKING

Advances in deep learning techniques have enhanced the performance of ASR systems. End-to-end ASR approaches take advantage of these advances by benefiting from independent intermediate models (acoustic, pronunciation, and language models) and the ASR model training process reduced complexity. Several easy to use and easy to update E2E models have emerged, based on the main categories of E2E ASR architectures; CTC, RNN-T, and AED. These end-to-end models require the accessibility to a massive training dataset, to train extensive complex deep architectures. The findings revealed by many works dedicated to E2E ASR models rely on the scenarios and the availability of datasets. Most of these end-to-end ASR models have reached state-of-the-art accuracy on the LibriSpeech dataset.

The first E2E models based on RNN-T architecture used as encoders the Long Short-term Memory models (LSTMs). By the time replacing the LSTM encoders with Transformer encoders gives a competitive model named Transformer-transducer. The experiments done in this work [37] found that with an equal number of parameters, Transformer-Transducer models trained much faster than the RNN-T models based on LSTM. Also, the proposed model can be improved by applying the RNN-T loss function, which is suitable for synchronized decoding and efficiently marginalizes all possible alignments. This Transformer-Transducer is suitable for streaming ASR through limiting the context of label and audio used in self-attention. The results obtained while training the Transformer-Transducer model on the LibriSpeech [40] dataset show that training this model on LibriSpeech clean/other test sets with 139M parameters without a language model achieves a WER of 2.4%/5.6%. With the usage of an external language model, the Transformer-Transducer model on the same dataset achieves a WER of 2%/4.6%.

Transformer and Convolution neural networks (CNNs) models have achieved promising results in ASR. By Benefiting from the fact that transformers are good at capturing content-based global interactions and from the effective exploitation of local features by CNN, new E2E transformers-CNN-based models have been proposed. In this context, a convolution-augmented transformer model for speech recognition subscribed under the name of Conformer has been introduced.

Conformers significantly outperform the previous ASR models based on Transformers and CNN, trained on the widely used speech recognition dataset Librispeech Conformer without a language model on Librispeech test-clean set to achieve 2.1% as a WER which mean an accuracy of accuracy of 98% and a WER of 3% on test-other set. while with an external language model Conformer achieved a WER of 1.9% on the test-clean set and 3.9% on the test-other set. This result was achieved with a large version of this model with about 118M parameters. Other competitive performances of Conformer have been achieved by the medium and the small version of this model with about 30M and 10M parameters respectively. Conformer show 15% improvement compared to transform-transducer-based models.

Inspired by the wav2letter approach, another family of neural architectures for E2E speech recognition has been presented. Named Jasper, this model consists in replacing the acoustic and pronunciation models with a convolutional neural network. Jasper models consist of a block architecture like this; Jasper BxR with B as the number of blocks, and R as the number of sub-blocks within each block. Where each block consists of one 1D convolutions, batch normalization, ReLU, dropout layers, and residual connections to enable the depth architecture. The smaller version of Jasper uses 34 convolutional layers with about 201M parameters, while the deepest version of Jasper uses 54 convolutional layers with about 333M parameters. For more training efficiency, a smaller memory footprint NovoGrad optimizer has been used with this model, which represents a new variant of the Adam optimizer. Evaluated on LibriSpeech Jasper show competitive results, With Jasper10x5 architecture, with about 201parameters on LibriSpeech clean/other test sets, this model achieves a 2.95%/8.79% WER using an external language model with a beam-search decoder, and 3.86%/11.95% WER with a greedy decoder without a language model.

Large E2E models have achieved very good accuracy but at the cost of high computational and memory requirements. Some research has focused on building E2E ASR models that can achieve the same accuracy but are faster to train and require fewer parameters while providing a higher inference rate and easy to deploy on hardware with limited computing memory.

In this work [41] an end-to-end neural acoustic model having fewer parameters was proposed, named QuartzNet. This model is subscribed under CNN E2E models and

designed based on jasper E2E model architecture [42], with the replacement of the one-dimensional convolutions with one-dimensional time-channel separable convolutions. The model architecture consists mainly of multiple blocks with residual connections between them. Each block consists of one or more modules with 1D time-channel separable convolutional layers, batch normalization, ReLU layers, and uses the CTC loss as the training loss function. This model is one of the most accurate speech recognition models on the LibriSpeech dataset. It achieves a Word Error Rate (WER) of 3.9% and 11.28% on the clean and other LibriSpeech test sets, respectively, without using an external language model. With the usage of an external language model, the WER is further improved to 2.69% and 7.25%. The small size of this model offers new scope for speech recognition on embedded and mobile devices.

Despite the promising results obtained by the CNN E2E speech recognition models, these models are by no means equal to the performance of transformer-based or RNN-based models. Recent research [43] has focused on creating a fully convolutional encoder capable of incorporating global contextual information into the convolution layers by adding squeeze-and-excite modules, this proposed model has been called ContextNet, which is based on the RNN-Transducer architecture. Trained on clean/other LibriSpeech test sets, ContextNet with about 112M parameters achieves a WER of 2.1%/4.6% without a language model and a 1.9%/4.1% when using an external language model. Also, ContextNet showed competitive results with its 10M and 30M parameters versions.

Sequence-to-sequence models and RNN-T models are autoregressive which makes them much slower to train or evaluate, CTC base models are non-autoregressive which makes them more stable and much easier to train. Nevertheless, Seq2Seq and RNN-T models outperform CTC models' accuracy. Due to the CTC conditional independence assumption, it is necessary to use a language model when CTC is used. In [44] a new deep convolutional CTC model named CitriNet, has been introduced to overcome the CTC model's weaknesses and benefit from the advances in neural network architectures.

CitriNet is a non-autoregressive CTC-based model proposed to bridge the gap between CTC and the best Seq2Seq and Transducers models, by introducing an encoder that integrates the squeeze-and-excite mechanism of the ContextNet model and the 1D time-channel separable convolutions of QuartzNet model. Citrinet-1024 the large version of Citrinet with about 142M parameters Trained on LibriSpeech clean/other test sets achieved a WER of 2%/4.69% without a language model and a WER of 2.52%/6.22% with an external language model. Thus, CitriNet model accuracy on LibriSpeech dataset without any external language model is close to the autoregressive models' accuracy, which goes against the popular notion that CTC models need an external language model to output accurate results.

IV. DISCUSSION

In this work, we have collected the most promising E2E models in the speech recognition fields. Each of these models has its advantages and drawbacks. We realize that it is always necessary to think about the combination of the advantages of one or more architectures. which can lead to a more efficient and less expensive model.

TABLE I. E2E AUTOMATIC SPEECH RECOGNITION MODELS

Model name	Params (M)	Without LM		With LM	
		clean	other	clean	other
Conformer(S)[39]	10.30	2.7	6.3	2.1	5
Conformer(M)[39]	30.70	2.3	5	2	4.3
Conformer(L)[39]	118.80	2.1	4.3	1.9	3.9
ContextNet(S)[43]	10	2.9	7	2.3	5.5
ContextNet(M)[43]	30	2.4	5.4	2	4.5
ContextNet(L)[43]	112	2.1	4.6	1.9	4.1
Transformer-Tranducer[37]	139	2.4	5.6	2	4.6
QuartzNet 15x5[41]	19	3.9	11.28	2.69	7.25
CitriNet-256[44]	9.80	2.52	5.95	3.78	9.6
CitriNet-512[44]	36.50	2.19	5.5	3.11	7.82
CitriNet-768[44]	81	2.04	4.79	2.57	6.35
CitriNet-1024[44]	142	2	4.69	2.52	6.22
Jasper DR 10x5[42]	201	3.86	11.95	2.95	8.79
JasperDR 10x5 (+ Time/Freq Masks)[42]	333	4.32	11.82	2.84	7.84

From the benchmark above we deduce that the main advantage of a CNN model is its parameter throughput; to improve both the speed and accuracy of the CNN models the depth-separable convolutions [45] have been used [46], however, the overall WER obtained by the first CNNs models, Jasper and QuartzNet [41], is still higher than that of the RNN/transformer-based models [47][37], which is argued by the small length of CNN models context. RNN/Transformer models Benefit from the bidirectional nature of RNN models, which allows information access in the whole context, and from the attention mechanism of transformers models. ContextNet incorporates both the advantages of CNN models and RNN/ transformer models, the state of art results of this model shows a competitive WER compared to the existing RNN/Transformer models with 112M parameters on LibriSpeech clean/other test sets this model achieved a WER 2.1/4.6 without incorporating a language model and 1.9/4.1 with the inclusion of a language model which is clearly under the WER achieved by the RNN/Transformer models on the same LibriSpeech test sets.

By comparing the different WER displayed on the TABLE.I we can see that the training method impact the performance of the E2E ASR models. Except CitriNet model the WER achieved when train the E2E ASR model with a language model is lower than the one achieved when no language model is used. Also, we can see that the models WER decrease when using more parameters except the case of JasperDR (10x5).

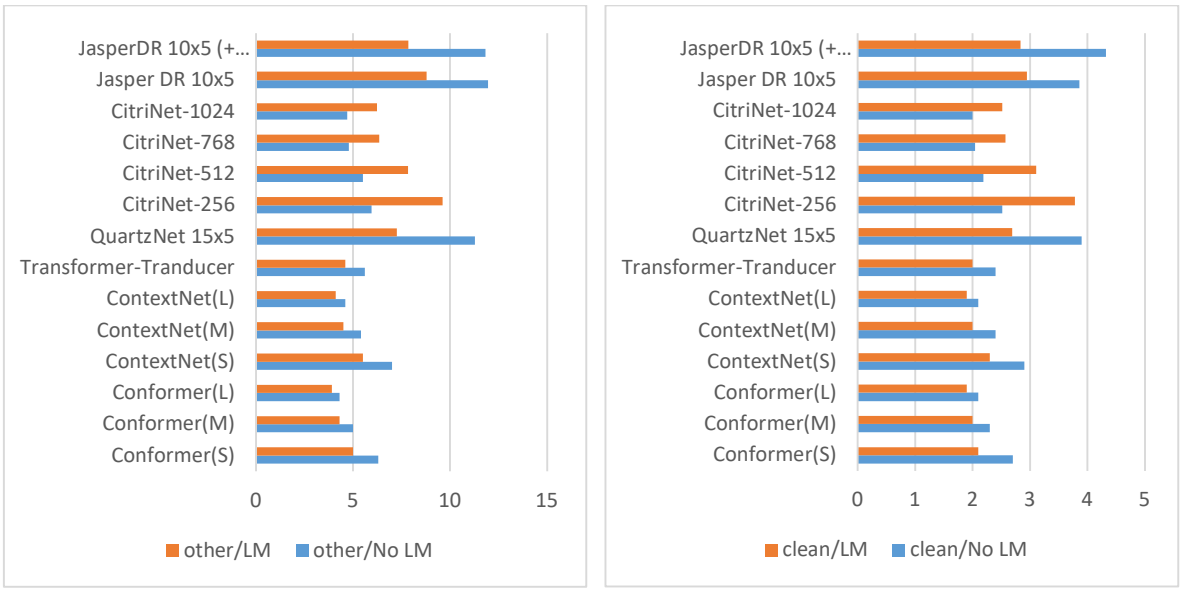

Fig. 4. End-to-End models training with and without a language model comparison

The charts displayed in fig.4 show that between the presented E2E models, the Conformer model [39] achieved the best WER followed by ContextNet[43] and Transformer-Tranducer[37] models(more details shown in TABLE I). Nevertheless, in noisy, under-resourced, or multichannel contexts, end-to-end models still are far from reaching the performance of HMM-based models. Up to now, E2E models have not been able to outperform the CNN hybrid models under these challenging circumstances. Meanwhile, to enhance the performance of E2E models, Data augmentation techniques have been used to increase the variety and quality of training inputs by satisfying some criteria to enhance E2E models robustness.

This study has shown that convolutions and Transformers-based models are the most promising models in the speech recognition research field, recent research confirms that the combination of CNN modules and transformers in one architecture results in successful E2E speech recognition models [39][43]. However, these CNN and transformer-based models need to be trained with a variety of datasets, the best results have only been obtained on the most popular datasets in the field of speech recognition, such as LibriSpeech. further work needs to be done to train these models on other challenging datasets, such as Arabic datasets and dialectal datasets.

V. Conclusion

In this paper, we benchmarked the different E2E models used for speech recognition, We explored the specifics of each E2E model and how convolutions and transformers have enhanced the performance of the speech recognition E2E models. The findings of this work highlight that the usage of a language model has a big impact on reducing the WER. Also, we deduce that Conformer and ContextNet E2E models are close to being competitive. However, the performance of these models has not been tested in the challenging contexts of speech recognition in which speech enhancement techniques are involved. In future work, we will revisit these two models on the Moroccan dialectal Arabic "Darija" Speech Dataset.

References

[1] X. Tang, "Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition," in *2009 Pacific-Asia Conference on Circuits, Communications and Systems*, May 2009, pp. 682–685, doi: 10.1109/PACCS.2009.138.

[2] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2013, pp. 273–278, doi: 10.1109/ASRU.2013.6707742.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 4960–4964, doi: 10.1109/ICASSP.2016.7472621.

[4] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2017, pp. 193–199, doi: 10.1109/ASRU.2017.8268935.

[5] C.-C. Chiu *et al.*, "State-of-the-Art Speech Recognition with Sequence-to-Sequence Models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4774–4778, doi: 10.1109/ICASSP.2018.8462105.

[6] N. Moritz, T. Hori, and J. Le, "Streaming Automatic Speech Recognition with the Transformer Model," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6074–6078, doi: 10.1109/ICASSP40776.2020.9054476.

[7] T. Hori, N. Moritz, C. Hori, and J. Le Roux, "Advanced Long-context End-to-end Speech Recognition Using Context-expanded Transformers," *arXiv Prepr. arXiv2104.09426*, 2021.

[8] J. M. Perero-Codosero, F. M. Espinoza-Cuadros, and L. A. Hernández-Gómez, "A Comparison of Hybrid and End-to-End ASR Systems for the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge," *Appl. Sci.*, vol. 12, no. 2, p. 903, Jan. 2022, doi: 10.3390/app12020903.

[9] M. Labied and A. Belangour, "Moroccan Dialect 'Darija' Automatic Speech Recognition: A Survey," in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, Jul. 2021, pp. 208–213, doi: 10.1109/PRML52754.2021.9520690.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 369–376, doi: 10.1145/1143844.1143891.

[11] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[12] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv Prepr. arXiv1412.5567*, 2014.

[13] A. Graves, "Sequence Transduction with Recurrent Neural Networks," 2012, [Online]. Available: http://arxiv.org/abs/1211.3711.

[14] J. Li, R. Zhao, H. Hu, and Y. Gong, "Improving RNN Transducer Modeling for End-to-End Speech Recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019, pp. 114–121, doi: 10.1109/ASRU46091.2019.9003906.

[15] H. Hu, R. Zhao, J. Li, L. Lu, and Y. Gong, "Exploring Pre-Training with Alignments for RNN Transducer Based End-to-End Speech Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7079–7083, doi: 10.1109/ICASSP40776.2020.9054663.

[16] L. Lu, X. Zhang, and S. Renais, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5060–5064, doi: 10.1109/ICASSP.2016.7472641.

[17] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved Training of End-to-end Attention Models for Speech Recognition," in *Interspeech 2018*, Sep. 2018, pp. 7–11, doi: 10.21437/Interspeech.2018-1616.

[18] Y. He *et al.*, "Streaming End-to-end Speech Recognition for Mobile Devices," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6381–6385, doi: 10.1109/ICASSP.2019.8682336.

[19] G. Saon, Z. Tuske, D. Bolanos, and B. Kingsbury, "Advancing RNN Transducer Technology for Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 5654–5658, doi: 10.1109/ICASSP39728.2021.9414716.

[20] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 2017-December.

[21] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-Based End-to-End Speech Recognition on Voice Search," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4764–4768, doi: 10.1109/ICASSP.2018.8462492.

[22] M. Sperber, G. Neubig, J. Niehues, and A. Waibel, "Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation," *Trans. Assoc. Comput. Linguist.*, vol. 7, pp. 313–325, Nov. 2019, doi: 10.1162/tacl_a_00270.

[23] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5884–5888, doi: 10.1109/ICASSP.2018.8462506.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 369–376, doi: 10.1145/1143844.1143891.

[25] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[26] S. Kim, M. L. Seltzer, J. Li, and R. Zhao, "Improved training for online end-to-end speech recognition systems," *arXiv Prepr. arXiv1711.02212*, 2017.

[27] A. Das, J. Li, R. Zhao, and Y. Gong, "Advancing Connectionist Temporal Classification with Attention Modeling," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4769–4773, doi: 10.1109/ICASSP.2018.8461558.

[28] T. Zhao, "A Novel Topology for End-to-end Temporal Classification and Segmentation with Recurrent Neural Network," *arXiv Prepr. arXiv1912.04784*, 2019.

[29] A. Graves, "Connectionist Temporal Classification," 2012, pp. 61–93.

[30] T. Hori, S. Watanabe, and J. R. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 518–529.

[31] C.-X. Qin, W.-L. Zhang, and D. Qu, "A new joint CTC-attention-based speech recognition model with multi-level multi-head attention," *EURASIP J. Audio, Speech, Music Process.*, vol. 2019, no. 1, p. 18, Dec. 2019, doi: 10.1186/s13636-019-0161-0.

[32] S. Punjabi *et al.*, "Joint ASR and Language Identification Using RNN-T: An Efficient Approach to Dynamic Language Switching," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 7218–7222, doi: 10.1109/ICASSP39728.2021.9413734.

[33] J. Li *et al.*, "Developing RNN-T models surpassing high-performance hybrid models with customization capability," 2020.

[34] T. Makino *et al.*, "Recurrent Neural Network Transducer for Audio-Visual Speech Recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019, pp. 905–912, doi: 10.1109/ASRU46091.2019.9004036.

[35] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 4945–4949, doi: 10.1109/ICASSP.2016.7472618.

[36] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," *arXiv Prepr. arXiv1911.07013*, 2019.

[37] Q. Zhang *et al.*, "Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7829–7833, doi: 10.1109/ICASSP40776.2020.9053896.

[38] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.

[39] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv Prepr. arXiv2005.08100*, 2020.

[40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210, doi: 10.1109/ICASSP.2015.7178964.

[41] S. Kriman *et al.*, "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)*, May 2020, pp. 6124–6128, doi: 10.1109/ICASSP40776.2020.9053889.

[42] J. Li *et al.*, "Jasper: An end-to-end convolutional neural acoustic model," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 71–75, 2019, doi: 10.21437/Interspeech.2019-1819.

[43] W. Han *et al.*, "ContextNet: Improving convolutional neural networks for automatic speech recognition with global context," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, no. 1, pp. 3610–3614, 2020, doi: 10.21437/Interspeech.2020-2059.

[44] S. Majumdar, J. Balam, O. Hrinchuk, V. Lavrukhin, V. Noroozi, and B. Ginsburg, "Citrinet: Closing the Gap between Non-Autoregressive and Autoregressive End-to-End Models for Automatic Speech Recognition," pp. 1–5, 2021, [Online]. Available: http://arxiv.org/abs/2104.01721.

[45] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1800–1807, 2017, doi: 10.1109/CVPR.2017.195.

[46] A. Hannun, A. Lee, Q. Xu, and R. Collobert, "Sequence-to-sequence speech recognition with time-depth separable convolutions," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 3785–3789, 2019, doi: 10.21437/Interspeech.2019-2460.

[47] S. Karita *et al.*, "A Comparative Study on Transformer vs RNN in Speech Applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019, pp. 449–456, doi: 10.1109/ASRU46091.2019.9003750.