American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN: 2320-0936
Volume-15, Issue-8, pp-39-44
www.ajer.org
Research Paper Open Access

A Conformer Based Automated Speech Recognition Model for the Albanian Language

Joan Janku[1]
[1]Faculty of Information Technology, Polytechnic University of Tirana, Albania
Corresponding Author: Joan Janku

ABSTRACT : *Speech to text transcription has always been challenging on low-resource languages.[1] This vpaper aims to solve the problem of speech-to-text transcription in Albanian Language. To achieve this, a deep learning model is implemented, using the DeepSpeech 2 architecture [2,3] in which context odelling is implemented using Conformer Blocks [4]. This architecture is tuned to uniquely accommodate the features and challenges of the Albanian Language.The dataset used to train and evaluate the accuracy of the model consists of 42 hours and 15 minutes of Albanian speech data, which includes different speaking styles and accents. The amount of data in the dataset, the tuning with conformer and the DeepSpeech 2 architecture has made it possible to achieve an accuracy of 49%, in accurately transcribing speech to text in Albanian Language. Our model is measured with WER and reaches a WER of 0.507 at epoch 32 against 0.576 of the standard DeepSpeech 2 architecture trained on the same data.*

KEYWORDS *DeepSpeech 2 architecture, Convolutional Neural Networks, Automatic Speech Recognition, low resource languages, Albanian Language*

Date of Submission: 11-08-2026 Date of acceptance: 24-08-2026

I. INTRODUCTION

The of amount of data available in low resource languages is not big enough thus creating a widely usable model to provide speech-to-text transcription in such languages poses a challenge. However, with the extensive research in the field of Deep Learning and Convolutional Neural Networks, the process of transcribing text from spoken audio signals has been eased a lot. The use of Convolutional Neural Networks to learn from the spectrogram converted audio-signals, has increased the accuracy of deep learning models in not only identifying spoken languages but also transcribing the given audio signals.

Albanian language introduces a few challenges when dealing with problems in ASR such as: rich vowel system, tonal stress, consonant clusters, free word order, particle use [5]. DeepSpeech based systems have already been implemented for other languages such as German [6] , and earlier work on Albanian language has relied on deep learning for the same problem [7] Other works have dealt with low resource and tonal languages through cross lingual adapters [8], tonal acoustic modelling [9] and toolkits such as PyTorch-Kaldi[10].
The aim of this paper, is to present a deep learning model which can be later used as a solid base model towards building a widely available "Automated Speech Recognition" tool, that can address some of the challenges that Albanian language introduces, which can be used for research or even professional purposes.

The dataset that we take into consideration consists of 42 hours and 15 minutes of audio signals in wav format, mono, with a sampling frequency of 16000 Hz, and a written CSV file which maps each audio filename to the real transcribed text. Every utterance carries the age, gender, accent and dialect of the speaker, the audio origin (ex. podcast or lecture) and a validation score between 0 and 1. The validation score of 0 represents an utterance that is not labelled , 0.5 when one user has labelled it and 1 when a second user has confirmed the same text. In this study we kept only the utterances with a validation score above 0.9. The dataset is split into 80:20 ratio, where 80% is used to train the model and the remaining 20% is used to evaluate the model.

Our model relies on converting spoken audio signals into filterbanks (instead of normal spectrograms), which are then inserted into our model which heavily relies in the use of Convolutional Neural Network layers based on the DeepSpeech 2 model. [2,3] The model is trained using these filterbanks and then evaluated, with the remaining 20% of the dataset in spectrograms format. [11]

II. PRESENTED APPROACH

The challenge when dealing with languages such as Albanian are the vowels and the exclamations. [5] In order for our deep learning model to be able to correctly identify the above two points, we added an extra conformer blocks [4]. Introducing Conformer Blocks over this architecture, allows us to make yuse of multi head self attention mechanisms and depthwise convolution inside the same blocks. The self attention mechanism part gives the global context of the utterance while the convolution part models the local context dependent on the phonemems. Both of these mechanisms are trained together instead of being attached together.

We also modified the input and the decoding. Instead of using raw spectrograms we instead use 80 log mel normalized filterbanks as inputs. The normalization is done “per utterance”. During the training “SpecAugment” [14] is applied on the filterbanks with 2 frequency masks and 2 time masks. This serves as data augementation and helps in our case since our dataset is small.
At inference time, the predictions are decoded with a bea search that makes use of a 4-gram language model trained on Albanian text withi KenLM [15].

The model that is studied in this paper, uses the DeepSpeech 2 architecture [2,3] supported by the Keras library [16] based on the implementation published in the Keras Documentation [17]. The model’s architecture is shown in Table 1.

Table 1. *Deep Learning Model Architecture for ASR in Albanian Language*

Layer	Description
Input Layer	Audio is converted into 80n log mel filterbanks, normalized per utterance and reshaped into 4 dimensions. Second and third dimensions represent frequency and time, while the fourth one the number of channels, which is 1 in this case (mono)
Spec Augment	Only during the training, 2 frequency masks and 2 tome masks are applied over the input [14]
Convolutional Layers	Filterbanks are fed into two Convolutional Layers with filter dimensions (11,41) and (11,21) and a step of (2,2) and (1,2) [3,17]. Each layer uses 32 filters and is followed by Batch Normalization and ReLu activation function.
Reshaping Layer	The output from convolutional layers is then converted into a 3D Tensor preserving frequency and number of channels. [3,17]
Conformer Blocks	The output from convolutional layers is then converted into a 3D Tensor preserving frequency and channel numbers. Then a linear layer projects it to 256

	dimensions so it can enter conformer blocks. There are 2 conformer blocks present, each of them containing 4 attention heads, relative positional encoding and a depthwise convolution at a kernal size of 31. The self attention mechanism [12,13] present in this block, captures the global context of the utterances, while the convolution part models the context dependent phonemes.
Recurrent Layers	The model uses Bidirectional GRU Layers [18,19] to capture the time variables from the input layer. Bidirectional architecture makes it easy for the model to gain context since it can capture both input and output i.e. the previous and the following context. GRU units are used instead of plain recurrent units because they are less affected from the vanishing gradient problem [20]. A dropout layer is then added to prevent overfitting [21].
Dense Layers	A dense layer with rnn_units x 2 units and a “ReLu” [22] activation function and a dropout layer transforms the learned features so they can be easily learned from the output layer [11]
Output Layers	Output Layer, which is a dense layer with output_dim + 1 unit, uses “softmax” activation function to allow for probability distribution over the output tags, which can then be used to predict the written text [23]. The extra unit is the blank symbol required by the CTC loss.
Optimizer and Loss Function	The model uses “Adam optimizer” [24] with a learning rate of “1e-4” , and “CTC” loss function [25]. The goal of the training phase is to minimize the loss function by adjusting the weights and predictions during the learning of the model.
Decoder	At inferences the predictions are decoded with a beam search that

uses a 4-gram language model trained on Albanian text with KenLM [15]. The language model is only used at inference and does not affect the training.

In our tests, we studied both the conventional DeepSpeech 2 architecture and the modified it for Albanian Language. We used the WER metric to measure the accuracy whilie evaluating both models, computed with the *jiwer* library at the end of every epoch. Both models were trained on the same dataset for 50 epochs, with a batch size of 32 on a Google Colab machine with an NVIDIA V100 GPU.

III. RESULTS AND DISCUSSION

The objective of this research, was to create an ASR model for a low-resource language such as Albanian Language, which could serve as a base for wider use in the future. We based our implementation on the DeepSpeech 2 architecture [3] and slightly tweaked it to accommodate Albanian Language features such as vowels and tone (exclamations) by adding conformer blocks and modifying the input to be filterbanks instead of spectrograms.

After evaluating the model with 20% of the dataset the achieved accuracy of our custom model was 49% as opposed to the normal DeepSpeech 2 model being 42.4%. Comparing the two architectures directly at epoch 32 our model reaches a WER of 0.507 while the standard DeepSpeech 2 stays at about 0.576 , so our model is 11.9% better at that point of training. The validation loss of our model also starts dropping faster and after the third epoch it stays consistently below the validation loss from the standard architecture. Figures 1 and 2 show the validation loss and model WER respectively.

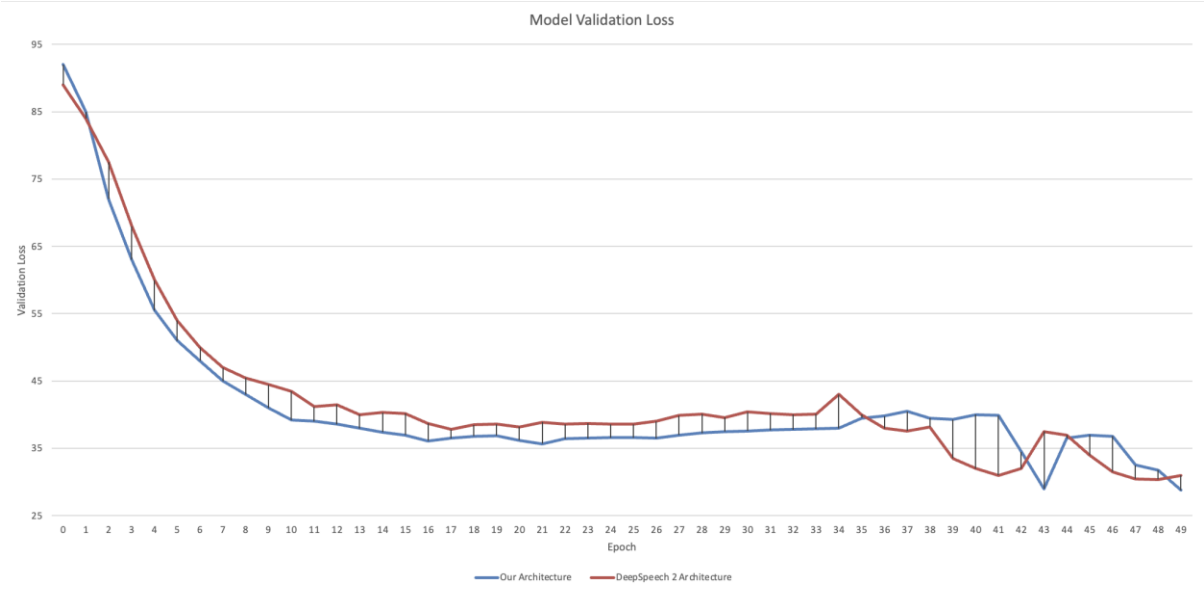

Fig. 1. Comparative models validation loss during training

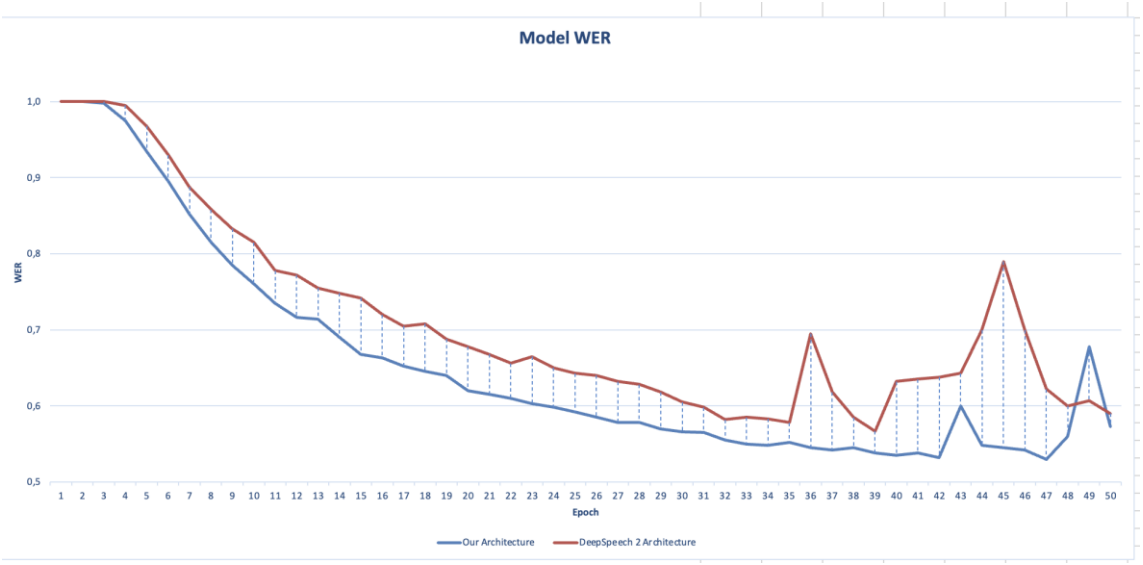

Fig. 2. Comparative model WER during training

Resulting in an accuracy of 49% compared to 42.4% of DeepSpeech 2 and a WER 11.9% lower than the one of DeepSpeech 2 is a good improvement compared to the traditional architecture. However, it is not enough for a widely used model. Further research on this model and especially an even larger dataset should be prepared in order for a more thorough evaluation for professional use.

IV. CONCLUSION

In this paper we studied the DeepSpeech 2 architecture and conformer blocks in Deep Learning for building an "Automated Speech Recognition" Tool in Albanian Language. We noticed that by using *80 mel log filterbanks* instead of plain spectrograms and by adding a Conformer Block at the end of Reshaping Layer for gaining more context and correctly modelling phoneme's local context, we were able to tweak the conventional DeepSpeech Architecture and gain 6.6% higher accuracy than the normal DeepSpeech archicture. These tweaks were implemented to address the existence of vowels and exclamations (tone) in Albanian Language and proved successful. However, as further work a larger dataset is needed in order to be able to perform a more detailed training of the model, also addressing the other challenges that Albanian Language poses such as free word order and consonant clusters.

REFERENCES

[1] Weizhe Wang, Xiaodong Yang, and Hongwu Yang. End-to-End Low-Resource Speech Recognition with a Deep CNN-LSTM Encoder. In 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), pages 158-162, September 2020.
[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep Speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567, 2014.
[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In Proceedings of the 33rd International Conference on Machine Learning (ICML), PMLR vol. 48, pages 173-182, 2016.
[4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In Interspeech 2020, pages 5036-5040, 2020.
[5] L. Newmark, P. Hubbard, and P. Prifti. Standard Albanian: A Reference Grammar for Students. Stanford University Press, Stanford, CA, 1982.
[6] J. Xu, K. Matta, S. Islam, and A. Nürnberger. German Speech Recognition System using DeepSpeech. In Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, pages 102-106. Association for Computing Machinery, New York, NY, USA, February 2021.
[7] A. Rista and A. Kadriu. A Model for Albanian Speech Recognition Using End-to-End Deep Learning Techniques. Interdisciplinary Journal of Research and Development, vol. 9, no. 3, 2022. doi:10.56345/ijrdv9n301.
[8] W. Hou, H. Zhu, Y. Wang, J. Wang, T. Qin, R. Xu, and T. Shinozaki. Exploiting Adapters for Cross-lingual Low-resource Speech Recognition. arXiv:2105.11905, December 2021.
[9] W. Hu, Y. Qian, and F. K. Soong. A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3206-3210, May 2014.
[10] M. Ravanelli, T. Parcollet, and Y. Bengio. The PyTorch-Kaldi Speech Recognition Toolkit. In ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing, pages 6465-6469, May 2019.
[11] U. Kamath, J. Liu, and J. Whitaker. Deep Learning for NLP and Speech Recognition. Springer, Cham, 2019.
[12] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-Based Models for Speech Recognition. In Advances in Neural Information Processing Systems 28 (NIPS), pages 577-585, 2015.

[13] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In 3rd International Conference on Learning Representations (ICLR), 2015. arXiv:1409.0473.
[14] 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. In Interspeech 2019, pages 2613-2617, 2019.
[15] K. Heafield. KenLM: Faster and Smaller Language Model Queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 187-197, 2011.
[16] F. Chollet et al. Keras. 2015. https://keras.io (accessed Aug. 2026).
[17] M. R. Bouadjenek and N. D. Huynh. Automatic Speech Recognition using CTC. Keras code examples, 2021. https://keras.io/examples/audio/ctc_asr/ (accessed Aug. 2026).
[18] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of EMNLP 2014, pages 1724-1734, 2014.
[19] M. Schuster and K. K. Paliwal. Bidirectional Recurrent Neural Networks. IEEE Transactions on Signal Processing, vol. 45, no. 11, pages 2673-2681, November 1997.
[20] S. Hochreiter. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 6, no. 2, pages 107-116, April 1998. doi:10.1142/S0218488598000094.
[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, vol. 15, pages 1929-1958, 2014.
[22] V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (ICML), pages 807-814, 2010.
[23] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, Cambridge, MA, 2016.
[24] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In 3rd International Conference on Learning Representations (ICLR), 2015. arXiv:1412.6980.
[25] 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), pages 369-376, 2006.