http://dx.doi.org/10.14500/aro.10837

Kurdish Dialect Recognition using 1D CNN

Karzan J. Ghafoor, Karwan M. Hama Rawf, Ayub O. Abdulrahman and Sarkhel H.Taher Karim

Department of Computer Science, College of Science, University of Halabja, Halabja,
Kurdistan Region - F.R. Iraq

***Abstract*—Dialect recognition (DR) is one of the most attentive topics in the speech analysis area. Machine learning algorithms have been widely used to identify dialects. In this paper, a model that based on three different 1D Convolutional Neural Network (CNN) structures is developed for Kurdish DR. This model is evaluated and CNN structures are compared to each other. The result shows that the proposed model has outperformed the state of the art. The model is evaluated on the experimental data that have been collected by the staff of Department of Computer Science at the University of Halabja. Three dialects are involved in the dataset as the Kurdish language consists of three major dialects, namely Northern Kurdish (Badini variant), Central Kurdish (Sorani variant), and Hawrami. The advantage of the CNN model is not required to concern handcraft as the CNN model is featureless. According to the results, the 1D CNN method can make predictions with an average accuracy of 95.53% on the Kurdish dialect classification. In this study, a new method is proposed to interpret the closeness of the Kurdish dialects using a confusion matrix and a non-metric multi-dimensional visualization technique. The outcome demonstrates that it is straightforward to cluster given Kurdish dialects and linearly isolated from the neighboring dialects.**

***Index Terms*—Convolution neural network; deep learning; dialect recognition; machine learning**

I. Introduction

Dialect is the language variety of a populace set up dependent on different real-life conditions (Chen, Shen, and Campbell, 2010). Every Dialect has its specific patterns of pronunciation, and there are regularly specific words or expressions that are explicitly being spoken among speakers of a specific Dialect (Najafian, *et al*. 2017, Najafian, *et al*. 2014). As of late, Dialect Recognition (DR) has become an intriguing problem for its wide applications in Speech Recognition. The adjusted Speech Recognition framework needs various tools, for example, the recognition of the dialect or the accent to standardize the speech tests for the Speech Recognition framework. For instance, Hirayama, Yoshino, Itoyama, Mori, and Okuno, (2015), build up a programmed Speech Recognition framework that accepts a mixture of various kinds of dialect.

There are many difficulties in the DR research area, for example, the assortment of speech information, which needs to display the different kinds of the examined dialect as well as languages (Diakoloukas, Digalakis, Neumeyer and Kaja, 1997). The results achieved by the researchers on DR are generally limited to the accessible collected data. To come across these challenges, we used a dataset collected by faculty members of the Computer Science Department at the University of Halabja. Subsequently, using the created algorithms beginning with the arrangement of the tested data or the classification techniques is non-persuading. Therefore, a few researches around using collected data under an explicit condition hold the important characteristics of the data. Huang and Hansen (2007) address the problem that considers the situation where no transcripts are accessible for preparing and testing data, and speakers are talking spontaneously.

Another challenge of Multi-dialect Speech Recognition is the growing adoption of smart devices. In particular, Kurdish poses an interesting challenge as the language has many dialects, and dialects do not have standard orthographic rules (Ali, 2018). Whereas the development of such a system faces the problem of the lack of annotated resources and tools, apart from the lack of standardization at all linguistic levels (phonological, morphological, syntactic, and lexical) together with the mispronunciation dictionary needed for Automatic Speech Recognition (ASR) development (Masmoudi, Bougares, Ellouze, Estève, and Belguith, 2018). To address this challenge, the study worked to reach a good result to recognize different dialects in Kurdish language. This will help to improve Speech Recognition for Kurdish dialects.

In this paper, Kurdish DR is studied and an end-to-end model is proposed based on 1D CNN for various classified Kurdish dialects including (Sorani, Badini, and Hawrami). Moreover, closeness of the dialects is presented based on a non-metric multidimensional visualization technique which is built on misclassification among the dialects.

The rest of the paper is organized as follows: Related work is presented in section II. Methodology of the proposed model is explained in section III. Experimental design and hyper parameters are clarified in sections IV and V, respectively. The result is discussed in the penultimate section (VI). Finally, conclusion of this work is pointed out in the last section (VII)

ARO-The Scientific Journal of Koya University
Vol. IX, No.2 (2021), Article ID: ARO.10837, 5 pages
DOI: 10.14500/aro.10837
Received: 21 June 2021; Accepted: 17 September 2021
Regular research paper: Published: 15 October 2021
Correspondent author's e-mail: karzan.ghafor@uoh.edu.iq

II. Related Work

In the literature, some of the studies have been focused on the feature of DR signals. Gaussian mixture model (GMM), for instance. In (Bahari, Dehak, Burget, Ali, and Glass, 2014), a positive factor analysis approach was developed for the GMM weight decomposition and conversion. Their study shows that GMM loads convey less, yet complimentary, data to GMM implies for language and DR. The Bangladeshi dialects using Mel Frequency Cepstral Coefficient (MFCC), its Delta and Delta-delta as main features and GMM to classify the characteristics of a specific dialect, by extracting the MFCCs, Deltas and Delta-deltas from the speech signal (Das, Allayear, Amin and Rahman, 2016). Whereas, in Haines, 2019, an unsupervised bottleneck feature extraction approach is proposed by Chunlei Zhang, to explain the effectiveness of the proposed methods, three datasets have been used (1) a four Chinese dialect dataset, (2) a five Arabic lingo corpus, and (3) multigene communicate challenge corpus (MGB-3) for Arabic. To overcome the utterance degrades. As it has been proposed by (Zhang, et al., 2019) an end-to-end approach to reduce recurrence varieties and extract the global context feature information vector using (CNN), Bidirectional Gated Recurrent Unit (CNN-BiGRU), which is useful to enhance the feature expression of short utterances. In another way, Experiments on ten Chinese dialects showed that the given method achieved 9.93% relative improvement in accuracy than the mainstream i-vector system. An experiment made by Huang and Hansen (2007) addresses novel advances in unsupervised TVDR in English and Spanish.

The researches of language and DR are generally using template-based and/or phonetic centered techniques. The template-based DR receives the utilization of global parameters of the speech signal playing little heed to the special attributes of the accessible phonemes identified with every dialect. This kind of research has been used by Choueiter, Zweig, and Patrick (2008) which found that an absolutely acoustic methodology dependent on a combination of heteroscedastic linear discriminant analysis and maximum mutual information training was exceptionally powerful. However, phonetic-based DR is likewise adopted and compared with acoustic and token-based DR and furthermore saw to be effective as in (Diakoloukas, Digalakis, Neumeyer, and Kaja, 1997). Another methodology embraced for DR is phonetic based recognition of dialect. A study receives a local feature that reflects the appearance of different phonemes in every language or dialect. For instance, both articles (Chen, Shen, and Campbell, 2010, Chen, Shen, Campbell, and Torres-Carrasquillo, 2011) propose supervised and unsupervised learning algorithms to find dialect discriminating phonetic rules and use these rules to apply biphones to identify dialects. Zhang and Hansen, 2018, presented an unsupervised bottleneck feature extraction approach to address the limitation of traditional bottleneck feature extraction, which is derived from the traditional bottleneck structure but trained with estimated phonetic labels.

In some research, a concealed Markov model is used to alter reference phones with dialect-specific pronunciations to describe when and how often replacements, insertions, and deletions occur. In Ying, Zhang, and Deng, 2020, anticipated an ASR system for Sichuan dialect by combining a hidden Markov model and a deep long short-term memory network which can overcomes the problem of using only captured context of a fixed number of information items compared with the deep neural network. To expand the research of DR, in this paper, we used 1D CNN to overcome the problem of feature extraction and recognition of dialects Sorani, Badini, and Hawrami.

III. Methodology

A. Data Description

Data collection is one of the basic requirements for those models developed based on the machine learning algorithm. Machine learning cannot be done without having a dataset. The utilized dataset was collected by some faculty members of the Computer Science Department in University of Halabja. All standard ways and rules were considered for the data acquisition including the consideration of various age groups and genders for those speakers who participated in the dataset. Furthermore, the geographical distribution of the speakers was considered as well. Generalization of the model can be achieved by adopting these considerations in the dataset. Three dialects were considered for the dataset as Kurdish language consists of three main dialects (Sorani, Badini, and Hawrami). Samples were recorded from TV broadcasts and debates. For each dialect, 180 different speakers were involved. Moreover, 299 samples for Sorani, 299 samples for Badini dialect, and 297 samples for Hawrami dialect were recorded. The length of each sample is only 1 s, thus 895 s are the total length of our dataset.

B. CNN

CNNs are a kind of deep neural networks which were originally proposed for 2D input. It is a powerful machine learning tool for learning features from the input raw data and outperforms the traditional machine learning models for image classification (Khan, Sohail, Zahoora, and Qureshi, 2020). One of the modifications of 2D CNNs is the 1D CNN, which has recently been applied in many applications as shown in recent researches (Kiranyaz, Ince, and Gabbouj, 2016, Acharya, et al., 2017, Ince, Kiranyaz, Eren, Askar, and Gabbouj, 2016 , and Kiranyaz, Gastli, Ben-Brahim, Al-Emadi, and Gabbouj, 2018). These researches have clarified that for certain applications, 1D CNNs are more preferable than one dimensional-based applications due to the low complexity, small number of hidden layers and neurons, and low cost of implementation. Typically, CNN models are mainly composed of two parts, feature extraction and classification. The section of feature extraction is responsible for extracting features from raw signals automatically, which normally consists of some layers such as convolution and pooling layers, and the other part (classification part) is in charge of the classification decision (Abdul, 2019).

http://dx.doi.org/10.14500/aro.10837

Furthermore, the classification part is identical to a typical Multi-Layer Perceptron and called as fully-connected layers (Kiranyaz, Ince, Hamila, and Gabbouj, 2015).

The configuration of any 1D-CNN explores some important processes as shown below:

- Initialize weights and biases
- Feed forward process applies from the input layer to the output layer to find outputs of each neuron at each layer. The process is formulated in equation (1)

$$x_k^l = b_k^l + \sum_{i=1}^{Nl-1} Conv1D\left(W_{ik}^{l-1}, s_i^{l-1}\right) \quad (1)$$

Where, x_k^l is the input, b_k^l is bias of the kth neuron at layer l, s_i^{l-1} and W_{ik}^{l-1} is the output of the th neuron at layer l–1 and the kernel from the th neuron at layer $l-1$ to the kth neuron at layer l, respectively. Moreover, *conv1D* is used to perform convolution process between W_{ik}^{l-1} and s_i^{l-1}

- Back propagation process: Start from computing delta error at the output layer and back-propagate it to the first hidden layer to compute the delta errors. The equation (2) below is delta error.

$$\frac{\partial E}{\partial_{W_{ik}^{l-1}}} = \Delta_k^t \, y_i^{l-1} \text{ and } \frac{\partial E}{\partial b_k^l} = \Delta_k^t \quad (2)$$

Where, E is the mean-squared error, y_i^{l-1} is intermediate output, and is Δ_k^t defined as delta error.

- Post-process to compute the weight and bias sensitivities.
- Update the weights and biases

IV. Experimental Design

The proposed model consists of seven layers which can be divided in two main parts, namely: Feature extraction part and classification part as shown in Fig. 1. The feature extraction part includes four sequential convolution layers, one max-pooling, and one drop layer. The convolution layers are different in size of the filters and their activation function for instance relu function is used for the first and the fourth convolution layers whereas tanh function is adopted in the second and the third convolution layer. The max-pooling layer is applied to reduce the dimension of the learned feature and the drop layer is used to overcome the overfitting problem. The classification part is responsible for classifying the learned feature from the previous part and it only consists of one fully-connected layer. Three nodes are involved in the mentioned layer because three dialects have been considered. The softmax function is usually used in a fully-connected layer for classification. More information about the structure of CNN are found in Table I.

V. Hyper Parameters

For CNN, some parameters should be tuned to obtain the optimum value to improve the performance of the model. There are some parameters of CNN that can be tuned such as number of convolution layers, number of fully-connected layers, selecting activation function, and number of epochs. Some of the mentioned parameters are tuned properly using nine folds' cross validation approach. The number of convolutional layers is tuned to four layers, where they include 20 filters for each layer. The number of fully connected layers is set as one with a number of nodes equal to the number of the classes which are three. The optimization technique adopted in this network is Adam Table II.

VI. Results And Discussion

A. DR

The length of the signal is 895 samples and divided into three classes, such as 299, 299, and 297 samples for each Sorani, Badni, and Hawrami, respectively. Based on relevant research in the literature, five-fold cross-validation is adopted

Table I
Description of the Layers

No.	Name	Description
1	Convolution Layer 1	20×5 convolutions with activation (relu), Input shape [44100×1]
2	Convolution Layer 2	20×7 convolutions with activation (tanh)
3	Convolution Layer 3	20×7 convolutions with activation (tanh)
4	Convolution Layer 4	20×3 convolutions with activation (relu),
5	Pooling	MaxPooling (9)
6	Dropout	90% dropout
7	Fully Connected1	fully connected three layers

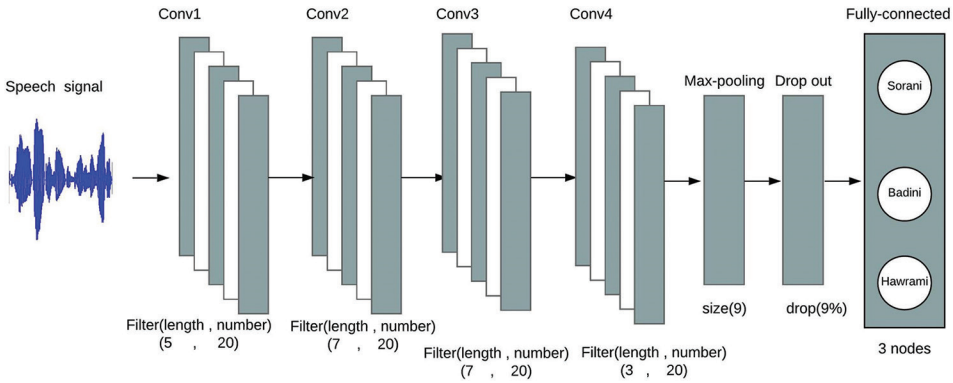

Fig. 1. Structure of the proposed model.

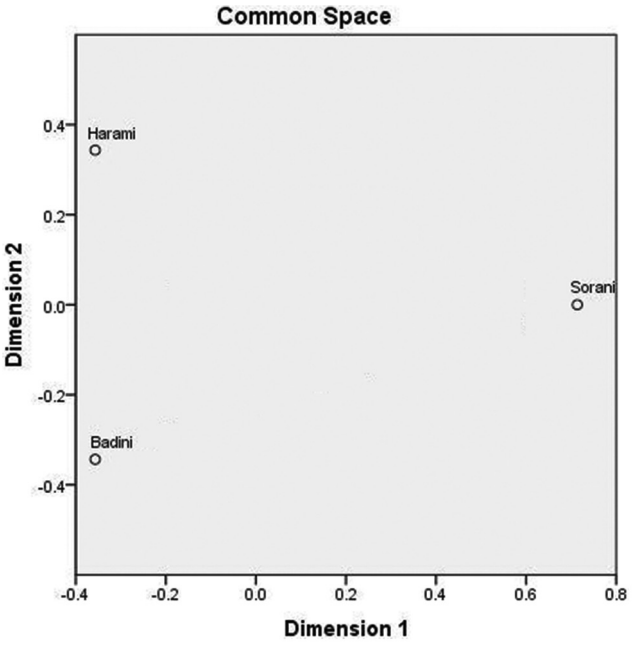

Fig. 2. Closeness of the Kurdish Dialect from each other's.

Table II
Confusion Matrices with Producer Accuracy and User Accuracy

No.	Sorani	Badini	Hawrami	Classification	Producer Accuracy (Precision)	F1-score
Sorani	53	0	2	55	96.364%	96%
Badini	1	58	1	60	96.667%	96%
Hawrami	1	3	60	64	93.75	94.%
Truth	55	61	63	179		
User Accuracy (Recall)	96.364%	95.082%	95.238%			

Table III
Accuracy of kurdish Dialect Recognition Model

Methods	Accuracy
Al-Talabani, Abdul an Ameen.	89.6
proposed model	95.5

to weigh the proposed method. The experimental result in Table III shows a significant improvement in the classification compared to the state of arts with P-value equal to 0.043 As compared to Al-Talabani's work, the study clearly shows that using MFCC feature set (95.5%) in one-dimensional CNN has greater accuracy than using MFCC feature set (71% weak accuracy) and local binary pattern (LBP)-LPC (89.6% best accuracy) fused feature set in one-dimensional LBP (Al-Talabani, Abdul, and Ameen, 2017). Moreover, Precision is 96%, 96%, and 93% for Sorani, Badni, and Hawrami and at least 95% is the proportion of actual positives which are identified correctly as illustrated in Table II.

B. Closeness of the Kurdish Dialects

The whole correct paragraph of Sub-section B should be:

The second purpose of this study is to illustrate how close the Kurdish dialect is to each other and how they impact each other in terms of the phonetic and style. A non-metric multidimensional visualization technique is applied for this purpose which is computed based on misclassification samples between dialects in such a way that a large number of misclassifications indicate high common properties between any two classes. Based on the result, Sorani Dialect has the same distance from both Badini and Hawrami; that means phonetics' Sorani is quite different from the other Dialects. But there are some similarities between Hawrami and Badini in terms of their phonetics and style. (Fig. 2 and Table III)

VII. Conclusion

In this study, 1D CNNs is proposed for Kurdish DR and presents the closeness of the Kurdish Dialects. Based on the experimental result, the performance of the proposed 1D CNN model outperforms the state of arts. Moreover, the proposed model is straightforward to apply because the model is built on the CNN which leads to carelessness about the handcraft feature, but time consumption is high as it requires to learn 23302 parameters including learning parameters of filters and learning weights for fully-connected layers. Another observation in this study is that the phonetic's Badini and Hawrami are closely related compared to the Sorani phonetic.

In addition, for future work some evaluation can be added to the study such as increasing the number of the sample, which is one of the significant points, may impact the results. Furthermore, increasing the number of classes such as Zazaki, Luri, and Bakhtyari have been changed the results and improve the evaluation of the model in the next steps.

References

Abdul, Z.K., 2019. Kurdish speaker identification based on one dimensional convolutional neural network. *Computational Methods for Differential Equations*, 7(4), pp.566-572.

Acharya, U.R., Oh, S.L., Hagiwara, Y., Tan, J.H., Adam, M., Gertych, A. and Tan, R.S., 2017. A deep convolutional neural network model to classify heartbeats. *Computers in Biology and Medicine*, 89, pp.389-396.

Ali, A., 2018. *Multi-Dialect Arabic Broadcast Speech Recognition*. [e-book] University of Edinburgh, Edinburgh p.193. Available from: https://www.era.lib.ed.ac.uk/bitstream/handle/1842/31224/Ali2018.pdf?sequence=1 and isAllowed=y [Last accessed on 2020 Dec 12].

Al-Talabani, A., Abdul Z. and Ameen, A., 2017. Kurdish dialects and neighbor languages automatic recognition. *The Scientific Journal of Koya University*, 5(1), pp.20-23.

Bahari, M.H., Dehak, N., Burget, L., Ali, A.M. and Glass, J., 2014. Non negative factor analysis of gaussian mixture model weight adaptation for language and dialect recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(7), pp.1117-1129.

Chen, N.F., Shen, W. and Campbell, J.P., 2010. A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models. In: 2010 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.5014-5017.

Chen, N.F., Shen, W., Campbell, J.P. and Torres-Carrasquillo, P.A., 2011. Informative Dialect Recognition Using Context-dependent Pronunciation

Modeling, ICASSP. In: IEEE The international Conference on Acoustics, Speech, and Signal Processing. pp.4396-4399.

Choueiter, G., Zweig, G. and Patrick, N., 2008. *An Empirical Study of Automatic Accent Classification*. Microsoft Research One Microsoft Way Redmond, WA 98052, pp.4265-4268.

Das, P.P., Allayear, S.M., Amin, R. and Rahman, Z., 2016. Bangladeshi Dialect Recognition Using Mel Frequency Cepstral Coefficient, Delta, Delta-delta and Gaussian Mixture Model. In: *Proceeding 8th International Conference on Advanced Computational Intelligence*, pp.359-364.

Diakoloukas, V., Digalakis, V., Neumeyer, L. and Kaja, J., 1997. Development of Dialect-specific Speech Recognizers Using Adaptation Methods, ICASSP. *IEEE The International Conference on Acoustics, Speech, and Signal Processing*, 2, pp.1455-1458.

Haines., Goleman, D., Boyatzis, R., Mckee, A., 2019. The meaning and process of communication. *Journal of Chemical Information and Modeling*, 53(9), pp.1689-1699.

Hirayama, N., Yoshino, K., Itoyama, K., Mori, S. and Okuno, H.G., 2015. Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2), pp.373-382.

Huang R. and Hansen, J.H.L., 2007. Unsupervised discriminative training with application to dialect classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 15(8), pp.2444-2453

Ince, T., Kiranyaz, S., Eren, L. Askar, M. and Gabbouj, M., 2016. Real-time motor fault detection by 1-D convolutional neural networks. *IEEE Transactions on Industrial Electronics*, 63(11), pp.7067-7075.

Khan, A., Sohail, A., Zahoora, U. and Qureshi, A. S., 2020. *A Survey of the Recent Architectures of Deep Convolutional Neural Networks, No. 0123456789*. Springer, Netherlands.

Kiranyaz, S., Gastli, A., Ben-Brahim, L., Al-Emadi, N. and Gabbouj, M., 2019. Real-time fault detection and identification for MMC using 1-D convolutional neural networks. *IEEE Transactions on Industrial Electronics*, 66(11), pp.8760-8771.

Kiranyaz, S., Ince, T. and Gabbouj, M., 2016. Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Transactions on Industrial Electronics*, 63(3), pp.664-675.

Kiranyaz, S., Ince, T., Hamila, R. and Gabbouj, M., 2015. Convolutional neural networks for patient-specific ECG classification. *Annual International Conference, IEEE Engineering in Medicine and Biology Society*, pp.2608-2611.

Masmoudi, A., Bougares, F., Ellouze, M., Estève, Y. and Belguith, L., 2018. Automatic speech recognition system for Tunisian dialect. *Language Resources and Evaluation*, 52, pp.249-267.

Najafian, M., DeMarco, A., Cox, S. and Russell, M., 2014. Unsupervised model selection for recognition of regional accented speech. *Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp.2967-2971,

Najafian, M., Hsu, W., Ali, A. and Glass, J., 2017, Automatic speech recognition of Arabic multi-genre broadcast media. In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), pp.353-359.

Ying, W., Zhang, L. and Deng, H., 2020, Sichuan dialect speech recognition with deep LSTM network. *Frontiers of Computer Science*, 14(2), pp.378-387.

Zhang, Q. and. Hansen, J.H.L., 2018. Language/dialect recognition based on unsupervised deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(5), pp.873-882.

Zhang, Q., Ma, Y., Gu, M., Jin, Y., Qi, Z., Ma, X. and Zhou, Q., 2019. End-to-End Chinese Dialects Identification in Short Utterances using CNN-BiGRU. In: 2019 *IEEE 8th Joint International Information Technology and Artificial Intelligence Conference* (ITAIC), 2019, pp. 340-344