Isolated Spoken Word Recognition Using One-Dimensional Convolutional Neural Network

Jihad Anwar Qadir¹, Abdulbasit K. Al-Talabani², and Hiwa A. Aziz¹

¹Department of Computer Science, University of Raparin, Rania, Iraq

²Department of Software Engineering, Faculty of Engineering, Koya University, Koya KOY45, Iraq

Ijfis

Abstract

Isolated uttered word recognition has many applications in human–computer interfaces. Feature extraction in speech represents a vital and challenging step for speech-based classification. In this work, we propose a one-dimensional convolutional neural network (CNN) that extracts learned features and classifies them based on a multilayer perceptron. The proposed models are tested on a designed dataset of 119 speakers uttering Kurdish digits (0–9). The results show that both speaker-dependent (average accuracy of 98.5%) and speaker-independent (average accuracy of 97.3%) models achieve convincing results. The analysis of the results shows that 9 of the speakers have a bias characteristic, and their results are outliers compared to the other 110 speakers.

Keywords: Feature extraction, Classification, One-dimensional CNN

1. Introduction

With the widespread growth in the use of digital electronic objects, the need to communicate with these devices has increased, especially through human friendly instructions such as uttering keywords like spoken digits. This sort of instruction design could be useful in applications such as control access for banking by telephone, voice dialing and database access services and for home appliances such as TVs, lamps, fans, etc. [1, 2]. As regards, for Kurdish language, with a population of over 40 million people worldwide, and up to our knowledge no extensive study for Kurdish keywords is found. The study of recognizing spoken words like the Kurdish digits can have another benefit in Kurdish speech recognition area because each of the spoken Kurdish digits (0–9) is a single syllable, which could be used to introduce a study of a syllable-based speech recognition in Kurdish language.

In this work, we used one-dimensional convolutional neural network (CNN) where no feature extraction step is required before applying CNN. CNN make the relationship between the raw speech signal and the phones able to be directly modeled, which exceed the step of feature extraction where the convolutional layers will be responsible of conducting this step [3]. Feature extraction for speech recognition is a challenging step, because each proposed feature might include some useful and non-useful information for a specific application. Convincing result achieved using CNN, where the feature extraction step is exceeded, is an ongoing and promising in the area speech recognition.

Received: Aug. 28, 2020
Revised : Nov. 11, 2020
Accepted: Nov. 30, 2020

Correspondence to: Jihad Anwar Qadir
(jihad.qadir@uor.edu.krd)
©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

The CNN take the isolated words data and extract a global feature from all of the signal. However, this is not what the traditional speech recognition scheme is followed. Based on the fact that the speech has a time series nature, researchers are mostly using approaches dealing with this sort of data such as hidden Markov model (HMM) and recurrent neural network (HMM) and Recurrent Neural Networks (RNN), which deals with a data sequence. But for limited isolated words especially those that represent a small units of the speech signal such as a single syllable, CNN approach could be promising and performing well.

In today's world, there are a few available datasets for spoken isolated words like digits. However, the need of large and high quality dataset is mandatory to conduct a reliable study. In this work, a dataset is designed to utter Kurdish digits (0–9), where 119 subjects participated in recording 11,872 samples in a quite environment. The achieved results in this study is promising for both the use of CNN in isolated speech recognition (a feature less technique), and as an indication for Kurdish syllabic speech recognition.

The rest of this paper is organized as follows: in Section (2), some related works has been discussed. Section (3) describes the proposed model approach, while Section (4) describes data used in this work, Section (5) present the results and its analysis and finally, Section (6) provides the overall conclusions and future works.

2. Related Work

In any traditional speech recognition application, feature extraction is mandatory to have a well representing vector of the input raw data. For isolated spoken digit recognition, as any speech recognition application many features are proposed in the literature such as: cepstrum features [2], mel frequency cepstral coefficients (MFCCs) [4–7], perceptual linear predictive (PLP) [8] and weighted MFCC (WMFCC) [9]. While for classification purpose, there are many proposed techniques for the same application, for example, Allosch et al. [1] consider two techniques which are pitch detection algorithm (PDA) [2, 6] and cepstrum correlation algorithm (CCA) [7] for digit recognition with a database of spoken Arabic digits (0–9) that consists of (3 males and 3 females) with speech recording in length of 12 seconds. A pre-processing including normalization and zero padding has been adopted, then PDA and CCA were compared and showed that PDA accuracies are 35.8% and 31.8% for males and females, respectively. The research found that the result

of using CCA relatively behaves better than PDA. Rudresh et al. [2] is also adopted vector quantization for feature matching using Euclidian distance metric aiming to recognize a spoken digit. The data used in this paper were a (0–9) digit speech that recorded in a single file with 10-second length with microphone using MATLAB that consist of 15 persons in 3 sessions with time gap of one-week duration. The training set consist of 2 sessions while the third session is used for the test set. The work achieves an average accuracy of 93.3%. While in [5], the authors used dynamic time warping (DTW) to detect speech digit recognition. Firstly, zero crossing and energy parameters were detected for the digit boundary detection. Then MFCCs are used to provide an estimate of the vocal tract filter which are fed into a DTW classification. An isolated English digit (0–9) are recorded and an accuracy of 90.5% was achieved for 100 samples. Kurian and Balakrishnan [8] offer a mechanism to speaker-independent connected digit recognizer for Malayalam language using PLP cepstral coefficient for speech parameterization and continuous density HMM to the recognition process. The used training data consists of 21 subjects with ages of 20 to 40 years old. Sounds recording process were done in a typical office environment and each file contains 20 continuous digits. The training data was 66% and the remaining data used for testing the model. The study achieved recognition accuracy of 99.5%. Fachrie and Harjoko [6] applied their experiments on Indonesian language, by extracting MFCC features and using natural logarithm of frame energy to improve it. The features were an input to Elman recurrent neural network (ERNN) for to recognize Indonesian spoken digits. Data was collected from 20 native persons and consists of 1000-digit utterances, 400 of them used to train the model while the remaining samples is used to test a speaker dependent and speaker independent models that accuracies were 99.3% and 95.17%, respectively. Kalith et al. [7] used CMU Sphinx tools to isolate the connected Tamil digit speech recognition and MFCC feature in addition to HMM to model the speech utterance and Viterbi beam algorithm to decoding process. Data is recorded with Samsung Galaxy j1 in unnoisy condition and consists of 10 speakers (5 males and 5 females) that divided into two parts, Isolated Tamil Digit (0–9) and Connecting Tamil Digit (random digit from 0–100). The result shown in Table 1. The best recognition accuracy obtained were 98.6% and 64.8% for speaker dependent and speaker independent, respectively. Ali et al. [10] have studied the Urdu language and used MFCC, delta and delta-delta features and classified by support vector machine (SVM), a random forest (RF) and a linear discriminant analysis

(LDA) classifiers. A comparison among SVM, RF and LDA has been done and shown that the best performance is achieved using SVM. The data set used consists of (5 males and 5 female) native/non-native speaker with different ages and the data is normalized with zero-mean and variance of 1. The achieved result was 73% for SVM and 63% for RF and LDA. Chapaneri [9] extract WMFCC features and used improved features for DTW (IFDTW) algorithms for speaker-independent isolated spoken digits (0–9). The data is taken from TI-DIGIT dataset and consists of 10 male and 10 female speakers and 400 utterances that uses 240 utterances (60%) for training and 160 utterances (40%) for testing. The obtained result was 98.13%. In another study conducted by Chapaneri and Jayaswal [11] a suggestion is made to reduce the time complexity of the recognition system by time-scale modification using a SOLA-based technique and also by using a faster implementation of IFDTW (FIFDTW). The data is taken also from TI-DIGIT dataset but consists of 15 males and 15 females with 600 utterances, where 360 utterances (60%) is utilized for training and 240 utterances (40%) is used for testing. Consequently, the recognition accuracy is improved to 99.16%.

3. The Proposed Model

The proposed model is designed to recognize isolated spoken Kurdish digits using data recorded for 119 subjects. Both speaker dependent and a speaker independent approaches are followed. One-dimensional CNN is proposed to extract features and classify the input features into their categories. It means that the raw data feeds the proposed model directly without any preprocessing or feature extraction step. A number of convolutional layers and pooling is suggested in this model as shown in Figure(1). Convolutional layer applies a filter masks on a specific window length equal to the length of the proposed filter. The first convolutional layer includes 110 filters of size 10×1 samples, followed by the the second layer that include 100 filters with size of 5×1 , then a max pooling of size 3×1 is

Figure 1. The architecture of the proposed model.

filter is applied. The third and the fourth convolutional layers include 50 and 10 filters of size 5×1 , respectively. Finally, two fully connected layers with number of neurons equal to 1024 is added, in addition to ten neurons in the output layer with softmax activation function.

4. Rail Data Set

Raparin Artificial Intelligent Lab (RAIL) data set is designed by the computer science department at University of Raparin. The dataset includes the recording uttered Kurdish digits (0–9) of different subject ages (18–46 years). The process of sound recording is implemented at different sessions. The number of participated subjects is 119 speakers (65 males and 54 females). A software is designed using MATLAB to record, save and prepare the data in a proper way. The sound recording is done in a quietude atmosphere and equal time for each person to utter the numbers (0–9), the sound of each subject has been recorded for about ten times. The duration of each isolated utterance is one second. The total amount of the recorded sounds is 11,928 samples which has been recorded with 16,000 sample rate and saved as a .wav file. The data set will be available for researchers to be utilized in any scientific study.

5. Results and Discussion

The first experiment aims to investigate how accurate the designed model is for speaker-dependent isolated spoken digit recognition. A 15-fold cross-validation is adopted; the mean of the accuracy is 98.52% with standard deviation of 0.004%. The box plot of all 15 experiments results is presented in Figure 2, which none of the experiment accuracy values is an outlier. The average confusion matrix (Table 1) shows how each single digit (class) has been recognised. The digit zero (sifr) is the digit with the highest incorrect classified samples. The most other two digits that confused with zero (sifr) is three (sey) and four (chwar). The confusion may be due to the diversity of stress per each subject or trial on the beginning phone /s/ in both (sifr) and (sey) and ending with /r/ in (sifr) and (chwar). The extracted features on these stressed phones could occupy more than it is real capacity which may dismiss the other phone features. However, still 96.2% of the samples regarding the digit zero is correctly classified. the speaker independent task consists of 119 experiments which is equal to the number of subjects participated in the dataset. Leave-one-speaker-out cross-validation is followed. The mean of the accuracies of all of the folds is

Figure 2. Boxplot for 15 speaker-dependent experiment accuracies.

Table 1. The average confusion matrix of 15 speaker-dependent experiments

Digits	0	1	2	3	4	5	6	7	8	9
0	96.2	0.0	0.1	0.7	2.7	0.0	0.3	0.0	0.0	0.0
1	0.0	99.0	0.0	0.1	0.1	0.1	0.0	0.1	0.1	0.7
2	0.0	0.0	98.0	0.3	0.1	0.0	0.0	0.1	0.0	1.6
3	0.4	0.3	0.0	98.2	0.9	0.0	0.2	0.0	0.0	0.1
4	0.5	0.0	0.0	0.0	98.2	0.0	0.6	0.2	0.0	0.4
5	0.0	0.3	0.3	0.1	0.0	99.0	0.0	0.0	0.1	0.3
6	0.0	0.0	0.0	0.0	0.8	0.1	98.5	0.0	0.6	0.0
7	0.0	0.0	0.2	0.0	1.0	0.0	0.0	98.2	0.3	0.3
8	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.4	99.4	0.0
9	0.0	0.2	1.5	0.0	0.3	0.1	0.0	0.0	0.0	98.0

97.38% with standard deviation of 3.8%. The boxplot presented in Figure (3) shows that the accuracy of 9 out of 119 speakers are an outlier and below 91%. In an analysis for the speakers that get outlier accuracies, we have found that the total number of misclassified samples is 119, which is almost 1% of the total number of the samples involved in this study. Adding this 1% to the speaker-independent accuracy make the result close the difference between speaker-dependent and -independent accuracies equal to only 0.14%. The confusion matrix of the speaker-independent results is shown in Table (2). Similar to the speaker-dependent the classification of the digit zero (sifr) is confused with both of the digits three and four which has been discussed previously. However, the digit nine (no) get worse result than zero, which is confused with the digit two (du). The diversity of uttering both of the vowels /o/ and /u/ among the

Figure 3. Boxplot for 119 speaker-independent experiment accuracies.

Table 2. The average confusion matrix of 119 speaker-independent experiments

Digits	0	1	2	3	4	5	6	7	8	9
0	95.5	0.4	0.2	1.7	1.1	0.1	0.8	0.2	0.1	0.0
1	0.1	96.7	0.0	0.3	0.3	0.4	0.0	0.1	0.0	2.0
2	0.0	0.0	98.5	0.3	0.0	0.0	0.0	0.2	0.0	1.1
3	0.3	0.4	0.3	98.0	0.6	0.3	0.0	0.0	0.0	0.2
4	0.3	0.1	0.1	0.3	98.1	0.0	0.8	0.4	0.1	0.0
5	0.4	0.7	0.5	0.2	0.0	97.6	0.0	0.2	0.3	0.0
6	0.3	0.0	0.0	0.0	0.9	0.2	97.7	0.0	0.9	0.0
7	0.1	0.0	0.2	0.0	0.3	0.1	0.0	98.8	0.5	0.0
8	0.1	0.1	0.0	0.0	0.1	0.5	0.7	0.4	98.2	0.0
9	0.0	0.8	4.1	0.0	0.1	0.1	0.0	0.2	0.0	94.7

speakers may be the reason of this confusion.

6. Conclusion

We can conclude that extracting learned features using CNN can stand instead of other techniques of feature extraction, and hence the feature extraction step could be exceeded or combined with the classification step in one process. The results show that the CNN is capable to recognize isolated utterances. Speaker independent based classification shows high accuracy all of the speakers, although, the accuracy for 9 speakers out of 119 where an outlier compared to other speakers.

The time series nature of the speech signal makes the expansion of the current application to larger number of spoken words challenging. For future work, learned features using CNN could be used as an input to a useful classifier for a sequence data such as long short-term memory (LSTM).

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

- [1] A. Allosh, N. Zlitni, and A. Ganoun, "Speech recognition of Arabic spoken digits," *Conference Papers in Science*, vol. 2013, article no. 130473, 2013. <https://doi.org/10.1155/2013/130473>
- [2] M. D. Rudresh, A. S. Latha, J. Suganya, and C. G. Nayana, "Performance analysis of speech digit recognition using cepstrum and vector quantization," in *Proceedings of 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, Mysuru, India, 2017, pp. 1-6. <https://doi.org/10.1109/ICEECCOT.2017.8284580>
- [3] D. Palaz and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," *Idiap Research Institute*, Martigny, Switzerland, 2015.
- [4] S. M. Omer, J. A. Qadir, and Z. K. Abdul, "Uttered Kurdish digit recognition system," *Journal of University of Raparin*, vol. 6, no. 2, pp. 78-85, 2019. [https://doi.org/10.26750/Vol\(6\).no\(2\).paper5](https://doi.org/10.26750/Vol(6).no(2).paper5)
- [5] S. D. Dhingra, G. Nijhawan, and P. Pandit, "Isolated speech recognition using MFCC and DTW," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, no. 8, pp. 4085-4092, 2013.
- [6] M. Fachrie and A. Harjoko, "Robust Indonesian digit speech recognition using Elman recurrent neural network," in *Proceedings of Konferensi Nasional Informatika (KNIF)*, Bandung, Indonesia, 2015, pp. 49-54.
- [7] I. M. Kalith, D. Ashirvatham, and S. Thelijagoda, "Isolated to connected Tamil digit speech recognition system based on hidden Markov model," *International Journal of New Technologies in Science and Engineering*, vol. 3 no. 4, pp. 1-11, 2016.
- [8] C. Kurian and K. Balakrishnan, "Connected digit speech recognition system for Malayalam language," *Sadhana*, vol. 38, no. 6, pp. 1339-1346, 2013. <https://doi.org/10.1007/s12046-013-0160-2>
- [9] S. V. Chapaneri, "Spoken digits recognition using weighted MFCC and improved features for dynamic time warping," *International Journal of Computer Applications*, vol. 40, no. 3, pp. 6-12, 2012. <https://doi.org/10.5120/5022-7167>
- [10] H. Ali, A. Jianwei, and K. Iqbal, "Automatic speech recognition of Urdu digits with optimal classification approach," *International Journal of Computer Applications*, vol. 118, no. 9, pp. 1-5, 2015. <https://doi.org/10.5120/20770-3275>
- [11] S. V. Chapaneri and D. J. Jayaswal, "Efficient speech recognition system for isolated digits," *International Journal of Computer Science & Engineering Technology*, vol. 4, no. 3, pp. 228-236, 2013.

Jihad Anwar Qadir was born in Sulaymaniyah, Iraq, in 1989. He received his B.S. degree in computer science from the University of Sulaymaniyah, Sulaymaniyah, Iraq, in 2011, and the M.E. degree in electronics computer engineering from the Institute of Natural and Applied Sciences at Hasan Kalyoncu University, Gaziantep, Turkey, in 2016. In 2011, he joined the Department of Computer Science, University of Raparin, as an assistant programmer and in 2017 became an assistant lecturer. His current research interests include deep learning, machine learning, computer vision, image processing, and speech processing.

E-mail: jihad.qadir@uor.edu.krd

Abdulbasit K. Al-Talabani was born in Kirkuk city, Iraq in 1977. He received the B.S. in mathematics from Salahaddin University, and M.S. and Ph.D. in computer science from Koya University, Kurdistan Region, Iraq in 2006 and the University of Buckingham, UK, in 2016, respectively. From 2003 to 2006, he was a research assistant at the Education College, Koya University. Since 2006, he has been an assistant lecturer, then a

lecturer at the Software Engineering Department, Koya University. His research interests include machine learning, speech analysis, deep learning, and vision.

E-mail: abdulbasit.faeq@koyauniversity.org

Hiwa A. Aziz was born in Mahabad, West Azerbaijan, Iran in 1983. He received his bachelor's degree in computer science from Payam-E-Noor University (PNU), Mahabad, Iran 2009. He received the master's degree in software engineering from PNU,

Tehran, Iran in 2012. He joined the University of Raparin from Iraq as an assistant lecturer in Basic Education Department in 2013, and he received a lecturer degree in 2019. His primary research interests are machine learning, deep learning, signal processing, and improve algorithm.

E-mail: hiwa.ahmad@uor.edu.krd