Computational Methods for Differential Equations
http://cmde.tabrizu.ac.ir
Vol. 7, No. 4 (Special Issue), 2019, pp. 566-572

Kurdish speaker identification based on one dimensional convolutional neural network

Zrar Khalid Abdul
Department of applied computer,
Charmo University, Sulaymaniyah, Iraq.

Department of computer, Halabja University, Sulaymaniyah, Iraq.
E-mail: eng.zrar1394@gmail.com

Abstract Voice is one of the vital biometrics in human identification and/or verification area. In this paper, two different models are proposed for speaker identification which are a 1D convolutional neural network (CNN) and feature based model. In the feature based model, three global spectral based features including Mel Frequency Cepstral Coefficient (MFCC), Linear Prediction Code (LPC) and Local Binary pattern (LBP) are fed to an SVM and k-NN classifiers. Results show that MFCC is the best feature among the others. Consequently, local MFCC features is extracted from the framed signal and used to both the proposed models. The result shows that the local based MFCC improved the accuracy of the CNN based model.

Keywords. Convolutional neural network, Identification, Machine learing.
1991 Mathematics Subject Classification.

1. Introduction

The speech is an important way of communication among human, and it is the most natural and efficient form of exchanging information among them. There are an ongoing investigations among researchers to improve the interconnection between human and computer [21]. High level of security can be obtained using one of the biometric systems since pin numbers and password can be forgotten any time and forged and are not offer top-level of security. Biometric system is the most potential technology that uses information about a person to recognize that person by relying on specific data about unique biological traits to work effectively. Speaker recognition technology is one of the biometric systems that make our everyday lives more secured. There are two major parts in speaker recognition which are speaker Identification (discovering identity) and speaker verification (authenticating a claim of identity [9]. Speaker identification system can recognize of known voices of speakers. i.e it is a 1: N match where the voice is compared against N templates. Any Error which occurs in the speaker identification system is called a false identification of the speaker. In another hand, Speaker verification is the process of accepting or denying the speaker claiming to be the real one. It is a 1:1 match where one speaker's voice

is matched to one template [19]. Speaker recognition can be classified into two parts in term of the data which are text-dependent and text-independent. In the text-dependent, the SRS builds based on the one or some specific phrases that the speaker says such as password, PIN code and a phrase. However, in text-independent, the system is based on the uncorrelated phrase that the speaker says [8]. Duaglas el at, proposed Gaussian Mixture Model(GMM) for text dependent speaker identification. Their result revealed that GMM works properly for short-term variation of the speakers voice [18] in [15], independent speaker recognition was done by integrating MFCC and phase information. Their result shown that the combining the MFCC and phase information can be reducing the error rate by 44.2Nowadays, deep learning is most interesting subject in machine learning area [22]. Authors in [12] applied two dimensional Convolutional Neural Network (CNN) for speaker identification and clustering. The spectrogram of the voices has been fed to the CNN model. The result shown that CNN is able to recognize unknown speakers with the less computational compared with traditional speaker identification. In this paper, two models are proposed for speaker identification. First one is feature based model and the second is one dimensional convolutional neural network. Framing speech is considered to extract MFCC feature.

2. Background

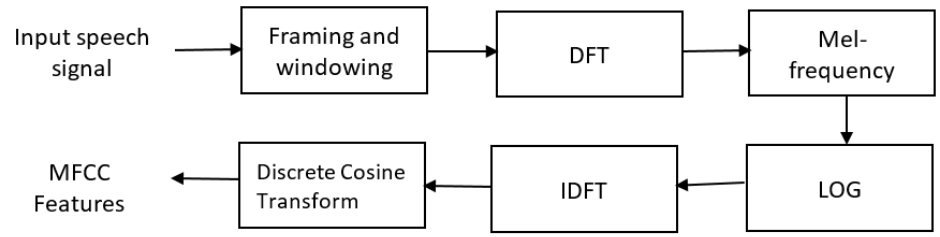
This section will focus on the background of the features, classification technique and CNN used in this work. The presented well-known features in the speech analysis area which are used in this study are: Mel Frequency Cypstrem Coefficients (MFCC), Linear Predicting Coefficients (LPC) and Local Binary Pattern (LBP). While SVM and kNN are the classification technique used which will be presented in an independent section. Moreover, CNN model is clarified in the separated section.

3. Methodology

3.1. **MFCC Feature.** One of the frequently used features in voice processing like voice recognition, speaker recognition and gender identification is the MFCC. In the recent years, it has been used in various application such as the bio-medical area and for the diagnosis of the child's body through the voice while crying [7]. The MFCC can be calculated based on short-term analysis. Therefore, the MFCC is calculated for each overlapped frame with length of 30ms with 15ms overlap. MFCC captures the important characteristic of phonetic in speech and shows the shape of the vocal tract manifests itself in the envelope of the short time power spectrum [14]. The computation of MFCC are summarized in the following steps and shown in Figure 1:

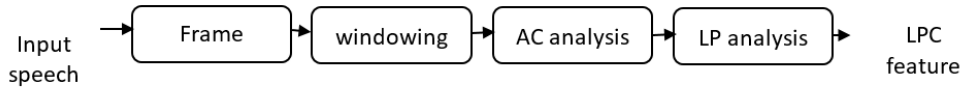
(1) Frame the signal into short frames.
(2) Calculate the power spectrum.
(3) Finding the mel filterbank to the power spectra
(4) Take the logarithm of all filter bank.
(5) Take the IDCT of the log filter bank.
(6) Keep DCT coefficients 2-13, discard the rest.

FIGURE 1. MFCC process.



3.2. **LPC.** Another powerful feature used in this study is the Linear Predicting Coding (LPC), which can determine the basic parameters of speech signal and provides precise estimation of speech parameters. LPC is estimated by the following steps blocking, windowing, auto correlation and Linear prediction respectively [13, 5] as shown in Figure 2.

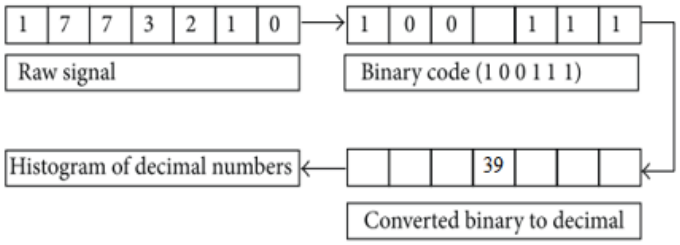
FIGURE 2. LPC Process.



3.3. **Local binary pattern (LBP).** The local binary pattern is a nonparametric operator [17]. The LBP code can explain the data using the differences between a sample and its neighbors. LBPs have been widely used, particularly in face recognition systems and speech analysis [2, 3]. At a fixed pixel position, the LBP operator is described as an ordered set of binary comparisons of pixel intensities between the center pixel and its neighboring pixels. However, LBPs used for images utilize the pixel neighbor in two dimen- sions, which is called 2D LBP. Moreover, 1D LBP can provide similar characteristics as the 2D LBP for the one dimensional signal. In this paper, 1D LBP is proposed to extract feature from the speech signal as shown in Figure 3.

3.4. **Support Vector machine (SVM).** Classification is an important step in any machine learning system and used in various fields as they help us to make decisions about categorizing a data. One of the supervised methods is call Support Vector Machine (SVM) which is one of the most robust and successful classification method. The basic Idea of SVM is minimize the margin between two separated hyperplanes. However, SVM can handle binary classification problem. For multiclass problems different approached need to be followed to perform the classification procedure. Pairwise SVM (followed in this study) which build a machine for each pair of classes is one of the adopted approaches [16].

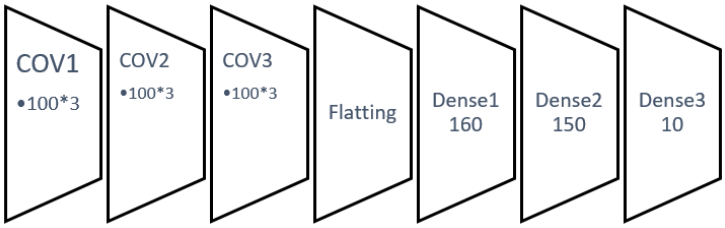
FIGURE 3. Local binary pattern process.



3.5. **K nearest neighbors (kNN).** K nearest neighbors is a simple classification method that isolates data based on its the closest neighbors. KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970. A data is classified by a majority vote of its neighbors, with the data being assigned to the class most common amongst its K nearest neighbors measured by a distance function [4]. There are some ways to measure the distance such as Euclidean, Manhattan and Chessboard distance [20].

3.6. **Convolutional Neural Network).** CNN has become a well-known tool in vision problem solving using machine learning. CNN is inspired by the mechanism of visual detection of living creatures [11], and wildly used in various applications based on feature extraction of vision input data [6]. Image input require a 2D CNN which is the most frequently used form of the CNN. However, 1D CNN is also applicable for one dimensional signal input such as speech signal [1]. The one dimensional nature of some signal emphasise the use of 1D CNN in various application for its consistency to capture useful features. Fusion of both 1D and 2D CNN is also followed in some application when the input include 2D and one D input signal such as emotion detection from image and sound [10]. In this paper we adopted the use of 1D CNN for feature extraction from the framed speech. The 1D CNN extracted feature will be later feeding a fully connected neural network and have ten nodes in the output layer referring to the labels (speakers). The structure of the proposed CNN model is shown in the below.

FIGURE 4. Proposed CNN model.



4. Data acquisition

Data acquisition is an important task in any classification process. The data set used in this study is designed and collected for this work. The number of subjects are 10 (5 men and 5 women). These ages are between (25-35) years old. There are 10 samples per speaker and the length of each sample is 1 second. The total number of the recorded sounds are 100 samples for all classes which are10. Praat is used to record and extract features. eventually, the sounds were kept to be used in the (.wav) format. As a step of processing MFCC, the data are divided to 195 frames. 12 MFCC's confidents are extract from each frame.

5. Result

The speaker identification system is implemented in two ways in term of the models. First, the feature-based model is used with both machine learning SVM and kNN. The second model is CNN model. Moreover, an experiment is done to choose the proper technique for extracting feature.

5.1. **Comparasion of three spectral based features.** Three spectral based features including Mel frequency Cepstral Coefficient (MFCC), Linear Prediction Code (LPC) and Local Binary pattern (LBP) are extracted from the entire speech signal and fed to Support Vector Machine (SVM) and K Nearest Neighbor (kNN). Based on the result that shown in Table 1, it can see clearly that the MFCC has better performance than the others as the obtained P-value less than 0.05 compared with LPC and LBP. Moreover, the error rate in the LPC is 0.29

Table 1. Accuracy rate.

Method	LPC	1D LBP	MFCC
kNN	60	40	88
SVM	71	62	90

5.2. **Feature based Model.** As the length of the speech signal is 1 second and it is divided to 195 frames. Then 12 coefficients are obtained from the speech signal and each of the frame is fed to the machine learning method separately. Based on Table 2 below, kNN has outperform result than the SVM by the P-Value=0.05 because the used data are not high dimensional.

5.3. **CNN model.** The same frames and the speech itself are passed through the CNN model. According to the result in Table 3, significant change can be noticed when the model fed the frame rather the speech directly. Moreover, there is not significant change in result between kNN feature based and CNN model (P-value

Table 2. Recognition rate of feature-based model.

Method	Accuracy
kNN	84
SVM	67

0.132). However, significant different can be seen between the result SVM feature based and the deep learning model with the P-value 0.03.

Table 3. Recognition rate of CNN model.

Method	Accuracy with direct speech	Accuracy with MFFC frames
CNN	46	81

6. Conclusion

The experimental result shows that the local based extracted MFCC from frames can be a useful channel to the CNN model as the amount of the frames can tune the parameters of the CNN properly. In this case, the model can be seen as a data augmentation procedure, because the number of samples increase such that one sample becomes 195 samples. Moreover, the frames can be useful for the feature-based method however there is not significant difference between the feature and the local based features result. Another observation is that kNN classification is still a powerful method for the low dimensional data classification compared to SVM.

Acknowledgment

Authors would like to thank the Kurdstan University in Iran, Charmo University and Halabja University for their support during the preparation of this work.

References

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, *Convolutional neural networks for speech recognition*, IEEE/ACM Transactions on audio, speech, and language processing, *22*(10) (2014), 1533–1545.

[2] T. Ahonen, A. Hadid, and M. Pietikinen, *Face recognition with local binary patterns*, European conference on computer vision, Springer, 2004.

[3] A. Al-Talabani, Z. Abdul, and A. Ameen, *Kurdish Dialects and Neighbor Languages Automatic Recognition*, ARO-The Scientific Journal of Koya University, *5*(1) (2017), 20–23.

[4] T. M. Cover and P. E. Hart, *Nearest neighbor pattern classification*, IEEE transactions on information theory, *13*(1) 1967, 21–27.

[5] N. Dave, *Feature extraction methods LPC, PLP and MFCC in speech recognition*, International journal for advance research in engineering and technology, *1*(6) (2013), 1–4.

[6] J. Gu, Z. Wang, J. Kuenb, L. Ma, A. Shahroudy, B. Shuai, T. Liub, X. Wang, , G. Wang, J. Cai, and T. Chen, *Recent advances in convolutional neural networks*, Pattern Recognition, *77* (2018), 354–377.

[7] S. Gupta, J. Jaafar, W. F. Ahmad, and A. Bansal, *Feature extraction using MFCC*, Signal and Image Processing, An International Journal (SIPIJ), *4*(4) (2013), 101–108

[8] M. R. Hasan and M. Jamil, *Speaker identification using mel frequency cepstral coefficients*, 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh, (2004), 565–568.
[9] W. B. Kheder, D. Matrouf, M. Ajili, and J. Bonastre, *Local binary patterns as features for speaker recognition*, in Odyssey., Proceedings of The Speaker and Language Recognition Workshop Odyssey 2016, 346–351.
[10] M. D. Lewis, *Self-organizing individual differences in brain development*, Developmental Review, *25*(3-4) (2005), 252–277.
[11] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A.W.M. van der Laak, B. Ginneken, C. I. Snchez, *A survey on deep learning in medical image analysis*, Elsevier, Medical Image Analysis, *42* (2017), 60-88.
[12] Y. Lukic, C. Vogt, O. Durr, and T. Stadelmann, *Speaker identification and clustering using convolutional neural networks*, 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP), IEEE, 2016.
[13] L. Muda, M. Begam, and I. Elamvazuthi, *Speaker-dependent-feature extraction*, Recognition and processing techniques, *10*(5-6) (1991), 505–520.
[14] L. Muda, M. Begam, and I. Elamvazuthi, *Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques*, arXiv preprint arXiv:1003.4083, 2010.
[15] S. Nakagawa, L. Wang, and S. Ohtsuka, *Speaker identification and verification by combining MFCC and phase information*, IEEE transactions on audio, speech, and language processing, *20*(4) (2012), 1085–1095.
[16] S. S. Nath, G. Mishra, J. Kar, S. Chakraborty, and N. Dey, *A survey of image classification methods and techniques*, 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), IEEE, 2014.
[17] T. Ojala, M. Pietikinen, and T. Menp, *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*, IEEE Transactions on Pattern Analysis and Machine Intelligence, *7* (2002), 971–987.
[18] D. A. Reynolds and R. C. Rose, *Robust text-independent speaker identification using Gaussian mixture speaker models*, IEEE transactions on speech and audio processing, *3* (1995), 72–83.
[19] K. V. Veena and D. Mathew, *Speaker identification and verification of noisy speech using multitaper MFCC and Gaussian mixture models*, IEEE, 2015 International Conference on Power, Instrumentation, Control and Computing (PICC), (2015).
[20] K. Q. Weinberger, J. Blitzer, and L. K. Saul, *Distance metric learning for large margin nearest neighbor classification*, In Advances in neural information processing systems, 2006.
[21] D. Yu and L. Deng ü, *Automatic speech recognition*, Springer, vol. 34, 2016.
[22] Q. Zhanga, L. T. Yang, Z. Chenc, and P. Li, *A survey on deep learning for big data*, Information Fusion, *42* (2018), 146–157.