SELCUK UNIVERSITY JOURNAL OF ENGINEERING SCIENCES

Journal homepage: http://sujes.selcuk.edu.tr

ISSN: 2757-8828

International Open Access

Volume 20
Issue 03

December, 2021

Research Article

Text independent speaker recognition based on MFCC and machine learning

Serhat Hizlisoy [a,*], ***Recep Sinan Arslan*** [b]

[a]*Kayseri University, Faculty of Engineering Architecture and Design, Department of Computer Engineering, Kayseri 38280, Turkey*
[b]*Kayseri University, Faculty of Engineering Architecture and Design, Department of Computer Engineering, Kayseri 38280, Turkey*

ARTICLE INFO

Article history:
Received 08 November 2021
Accepted 12 December 2021

Keywords:
Speaker Recognition
Text-Independent
Human Voice
MFCC
Performance Analysis
Machine Learning

ABSTRACT

Speaker recognition (SR) is the process of recognizing the voice of human from a group of speech samples with artificial intelligence. SR models are used in various human-voice based security platforms and authentication problems. In this paper, a text-independent speaker recognition model was developed for the problem with 60 different speakers. Obtaining the distinctive features of speaker expressions during the model design phase is an important point. In this study, the MFCC algorithm, which is the most common method used to obtain short-time features, is used to extract features of speech signals. The classification performance of the proposed model and commonly used 11 different machine learning methods has been evaluated on Audio-MNIST dataset, and the results were shown comparatively. As a result, 97.1% classification rate was achieved with SVM classifier. In addition, precision, recall and f-score values are 98.0%, 97.1% and 97.4%, respectively. The results show that the proposed model produces successful results for all classes and is a widely applicable approach to different types of speaker datasets.

This is an open access article under the CC BY-SA 4.0 license.
(https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Audio can be defined as vibrations that can be perceived by hearing organs, formed by shaping the air coming from the lungs with the help of various organs. Audio is a complex signal that can contain a lot of information. Speaking is the most effective way people use to communicate their ideas and opinions to the other side. The speaker's voice is unique to him like a fingerprint, and the structure of the larynx is determined by the physiological structure of that person, like the vocal cords; however, it is also affected by factors such as gender, age and origin [1].

Speaker recognition can also be called the process of detecting who is speaking by using information such as the frequency, speed, amplitude, quality, pause of the voice signal. Research on speaker recognition has increased significantly in recent years with the development of technology. There are many problems encountered in the design of the speaker recognition system. For example, the change of the sound signs over time is one of them. Changes in the voice may occur due to various factors such as illness, smoking and alcohol use, stress, emotion, speed of expression, speaking style. Technical issues such as different types of microphones and headphones used during the recording of words may potentially cause alterations in speech signals. The system to be created must be sensitive to these.

The fact that speech recognition systems are easily affected by external factors such as noise in the environment also makes the speaker recognition problem more difficult than other problems [2]. In order to design a good model, the training data is prepared in a noise-free environment and the test data is prepared in a noisy environment. Because background noise is one of the important factors affecting accuracy in speaker recognition and reduces the performance of the system. If you like music, traffic etc. If the speaker's voice is recognizable even in an environment with background noise, this is the most important indicator of the robustness of the system.

Another problem that may affect the performance of speaker recognition systems is the lack of sufficient data to

* Corresponding author. E-mail address: *serhathizlisoy@kayseri.edu.tr*
This paper was presented at the International Conference on Engineering Technologies (ICENTE'21), 18-20 November 2021, Konya, Turkey.

make the right decision by training the models. Because the performance of the system changes the amount of training data. It is believed that the more data the system is trained with, the more successful the speaker identification system will be if the test data fits the distribution represented by the speaker model.

The speaker recognition process is classified as text-dependent or text-independent. Different words or sentences are utilized in the training and testing stages of text-independent speaker identification systems, but in text-dependent systems, the text is known in advance and the same words or sentences are used. The training process of the text-based system is fast as it has a fixed set of inputs where the speaker has to say the same words each time to validate it. The training phase of the text-independent system, on the other hand, takes a long time because the model does not consider the spoken word, instead attempting to transform the sound into feature vectors in order to reliably identify the speaker.

Speaker recognition systems consist of three important stages as preprocessing, feature extraction and modeling of the system. Preprocessing is the first stage in the processing of the speech signal and converts an analog signal to a digital one. Its purpose is to prepare the speech signal for feature extraction. Negative factors such as noise in the speech signal to be used in the classification phase are cleaned at this stage, allowing more accurate features to be extracted. In other words, a very important and critical step prevents the degradation of the classification performance of the system. In the feature vector extraction stage, various information representing people from the voices are obtained and digital vectors are created. These feature vectors are then used to train the generated model. The most used features also in emotion recognition [3] and speech recognition [4] studies are Linear Prediction Cepstral Coefficients (LPCCs), Mel Frequency Cepstral Coefficients (MFCC), and Linear Prediction Coding (LPC). At the end of the system, a decision mechanism will do the classification [5]. Thanks to this process, using the input test vector and the trained model, the speaker in the test data is found. In an ideal speaker identification system, the classification methods used should achieve high accuracy while still producing fast results. Mel Frequency Cepstral Coefficients (MFCC) features that are widely used in text-independent speaker recognition tasks, and SVM, logistic regression, extra tree and KNeighbors, which are the classification algorithms that produce the fastest results among existing classifiers, are used in this study.

In the section II, explains of the previous studies on the subject. Section III gives information about the database, how the features are obtained, how they are selected, the classifiers used and the stages of their implementation. In section IV. you can find the results and discussions about the study and finally in section V. shows conclusions obtained in the study.

2. Related Works

Humans have various unique biometric features that distinguish people from each other, such as face, retina, iris, and signature. Speech signal is one of these features. Many scientists have carried out various studies on this subject over the years. As a result of these research, it has been concluded that sound is one of the most useful technologies in the field of security due to its easy implementation and low cost, as well as providing a much higher level of security. In our daily life, speaker recognition systems are used in many places such as telephone banking, telephone shopping, remote voice control of computers. Forensic events are one of the most important areas of application. If a speech sample was recorded during the crime, for example, the case could be solved by comparing it to the suspect's voice using voice sample matching. Again, the voice of a wanted criminal can be matched with the voice in the database to be located and captured. Authentication, on the other hand, is one of the most widely used biometric applications since it allows users to recognize someone by their speech. It's also common to employ a variety of ways to authenticate the speaker, such as a password or facial recognition.

In one of the first speaker recognition studies, Uzuncarsili (2005) [6] designed a text-independent speaker recognition system based on Vector Quantization techniques using MFCC feature vectors and achieved success at different rates depending on the number of speakers. Kral (2010) [7], on the other hand, achieved a recognition success of 98% by working on 2 different datasets consisting of Czech speeches of 10 and 50 speakers using the MLP classifier and DWT features. Nakagawa et al. (2012) [8], speaker recognition studies were conducted in which there were 35 and 270 different speakers in two different datasets, phase information and MFCCs were used as attributes, and GMM was used as a classifier, and 98.8% success was achieved.

Later, Hanilci (2017) [9] applied the GMM-UBM classifier on a Turkish speech dataset of 26 female and 47 male speakers with Mel-Frequency Cepstral Coefficient (MFCC) features to examine the speaker recognition problem. In another study, Celiktas and Hanilci (2017) [10] used the Gaussian Mixture Model-Universal Background Model method on the Turkish text dependent speaker identification system on a Turkish dataset consisting of 46 speakers (10 females 36 males) with a 5.73% error rate achieved. Soleymanpour and Marvi (2017) [11] proposed a technique by applying neural network (NN) to Mel-frequency cepstral coefficients to recognize over 22 speakers in the ELSDSR dataset, and they achieved a 93.2% recognition success. Jagiasi et al. (2019) [12], DNN and CNN methods were used in a text-free system using MFCCs for comparison purposes. Results from 50 speakers from the SRL82 dataset showed that CNN achieved better result with 71% compared to DNN, which only achieved 61%. Chowdhury and Ross, (2020) [13] to increase the quality of

the input voice signal, they were given a 1D-Triplet CNN with a combination of MFCC and LPC characteristics. Experiments on the NIST SRE 2008 and 2010 datasets, the Fisher dataset, and TIMIT dataset revealed strong performance against a variety of audio distortions. In the study by Jahangir (2020) [14], the LibriSpeech dataset consisting of 100 speakers was used. In this system, MFCC and time-based (MFCCT) features are fed to the DNN as input to create the speaker identification model. Better performance than the other five known classification algorithms with 92.9% accuracy. Hourri et al. (2021) [15], aimed to develop a customized CNN by proposing a new approach to extract speaker features by creating speaker-dependent CNN filters for speaker recognition. They also proposed new vectors, which they called conv-Vectors (convolutional neural network vectors), to identify speaker characteristics. They performed the experiments with a sex-linked dataset under three noise conditions: 0 db, 9 db, and clean. Comparing the proposed method with the latest methods, the results show that the conv-Vectors method is the most robust method and improves the basic system by an average of 43%.

3. Proposed Method

In this study, a text-independent speaker recognition model was developed for the problem with 60 different speakers. The flow-chart diagram of the proposed model was as shown in Figure-1. In the process of developing proposed model, first the dataset should be obtained. Audio-MNIST [16] dataset, which is openly distributed on the internet, was used. The dataset contains recordings of 60 different speakers. They consist of 48 male and 12 females between the ages of 23-31 with different accents such as German, South Korean or Spanish. None of them are native speakers for the English. All recordings are collected in a studio environment and do not contain any noise. As a result, there are a total of 30000 records in this dataset, in which 60 different speakers each repeat 10 digits 50 times. The reason for using this dataset is that there are quite a lot of audio recordings, and these show a heterogenous distribution. Thus, it is aimed to provide objectivity in the results of the proposed model. After the dataset was obtained, it was not possible to produce features with many feature extraction methods such as contrast, cq etc. Because the records were very short (<1 sec.). For this reason, features were extracted with MFCC algorithm. MFCC method flow diagram is as shown in Figure-1.

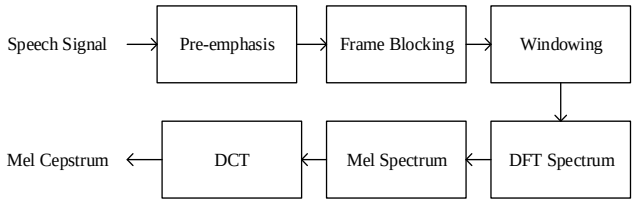

Figure 1. Steps in MFCC Feature Extraction

There are six steps in calculating the MFCC coefficients. In the first stage, the pre-emphasis process is performed on the audio signal. It refers to the filtering of higher frequencies. Its function is to compensate the spectrum of speech sounds with radical changes in the high frequency regions. The speech signal is a slowly changing or semi-stationary characteristic. For stable acoustic properties, it is necessary to study speech for a sufficient short period of time. Speech sound analysis should be done in parts where the speech signal is more stable. These processes are carried out in frame blocking and windowing stages. DFT is applied on each window and converted to magnitude spectrum. After DFT, the mel spectrum is calculated by passing the Fourier Transformed signal through the mel filter bank. In the final stage, the energy levels in adjacent bands tends to be close to each other. Cepstral coefficients were produced by applying DCT to the transformed mel frequency coefficients [17].

As a result of feature extraction process, a feature vector of 30000x28000 dimensions was obtained from which 28000 features were produces for each audio recording. In order to avoid problems that may occur during training and testing due to the size being too large, all float type data is converted to integer type. Thus, processing has been made easier. After this stage, the feature selection process was carried out. The reason for this is the elimination of features that negatively or not at all affect the classification success. Thus, it is aimed to achieve high accuracy rate. Principal component analysis (PCA) method was used for this. PCA is a powerful data representation method that can capture the most variable data components of samples [18]. It provides a lower dimensional representation of the feature vector by transferring it to new space. As a result of this step, an integrated reflection of the original feature components of the samples is obtained. With the PCA process, the feature size of each speech recording has been reduced to 500. After PCA, the last step of the pre-processing stage is data standardization using the normal distribution. Thus, it was ensured that the data were taken into a certain range. After all step, 30000 records are reserved at 70%-30% to be used in training and testing phases. Accordingly, the classifier trained with 24000 records and were tested with 6000 records. Obtained results were evaluated comparatively according to accuracy, precision, recall and f-score values calculated as given in Equation 1-4.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (3)$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ P. + True\ N. + False\ P. + False\ N.} \quad (4)$$

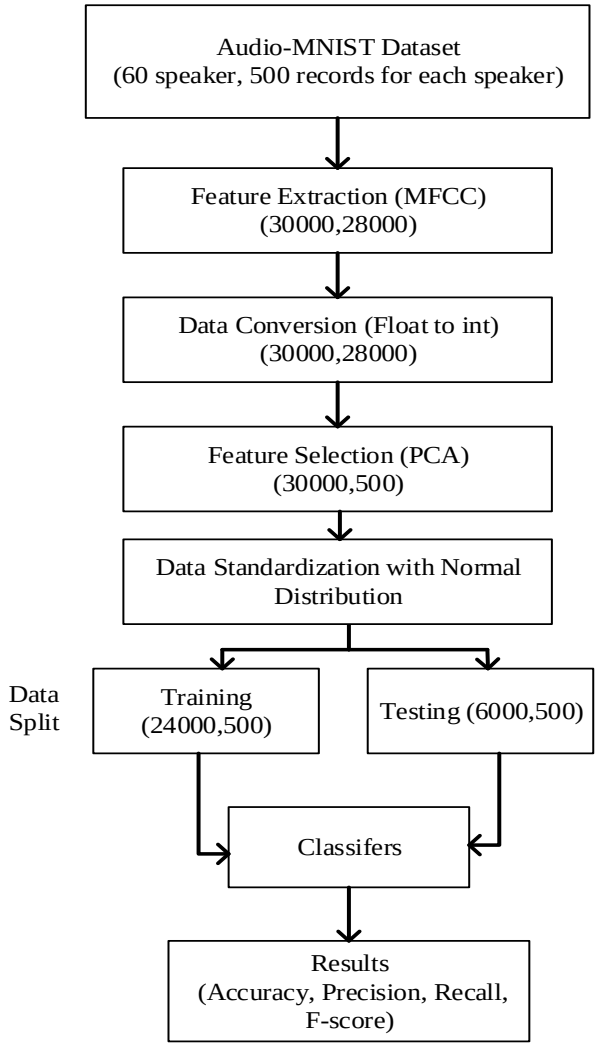

Figure 2. Flowchart of Proposed Model

4. Experimental Results

In this study, a text-independent speaker recognition model was proposed. The tests performed with the proposed model were evaluated comparatively.

In the first stage of the tests, the results obtained without PCA feature selection for 11 different machine learning algorithms were as shown in Table-1.

According to the classification results using 28000 features without PCA, the most successful algorithms were SVM, logistic regression, kneighbors and extra tree, respectively. While the accuracy value was 95.6% with the SVM algorithm and precision, recall and f-score were 97.5%, 96.5% and 96.9%. In general, high classification rate was achieved. It is thought that the decrease in performance, which is at the level of 3.5% is due to testing with text-independent samples. At this point, the biggest problem encountered in the tests with the feature vector containing 28000 features is that the training and testing times take quite a long time on the standard notebook computer. Despite achieving a high level of success, it was the handicap of the model that was proposed to require a long time. As a solution to this problem, the feature size has been reduced with PCA. After reducing the feature size to 500, the results obtained in repeated tests were as shown in Table-2.

A slight improvement in the classification performance was noted after PCA. Accordingly, the highest accuracy value was obtained with SVM with 97.1%. Precision, recall and f-score values were 98.0%, 97.1% and 97.4%, respectively. A very high classification rate was achieved for a multiclass (60 classes) problem. In addition, thanks to feature selection with PCA, significant efficiency has been

Table 1. Classification Results without PCA

Classification Algorithm	Number of Training-Test Samples	Feature Size	Accuracy	Precision	Recall	F-score
Logistic Regression	24000-6000	28000	0.964	0.964	0.964	0.964
Kneighbors	24000-6000	28000	0.966	0.967	0.966	0.966
Decision Tree	24000-6000	28000	0.520	0.522	0.520	0.520
Gaussian NB	24000-6000	28000	0.046	0.135	0.046	0.034
Linear Discriminant Analysis (LDA)	24000-6000	28000	0.920	0.922	0.923	0.920
AdaBoost	24000-6000	28000	0.053	0.072	0.053	0.034
Gradient Boosting	24000-6000	28000	0.879	0.882	0.880	0.870
Extra Tree	24000-6000	28000	0.960	0.960	0.960	0.960
XgBoost	24000-6000	28000	0.912	0.923	0.922	0.922
Random Forest	24000-6000	28000	0.909	0.910	0.908	0.909
SVM	24000-6000	28000	0.965	0.975	0.965	0.969

Table 2. Classification Results with PCA

Classification Algorithm	Number of Training-Test Samples	Feature Size	Accuracy	Precision	Recall	F-score
Logistic Regression	24000-6000	500	0.965	0.975	0.965	0.965
Kneighbors	24000-6000	500	0.967	0.968	0.967	0.967
Decision Tree	24000-6000	500	0.462	0.464	0.462	0.464
Gaussian NB	24000-6000	500	0.568	0.633	0.568	0.564
Linear Discriminant Analysis (LDA)	24000-6000	500	0.950	0.952	0.950	0.950
AdaBoost	24000-6000	500	0.144	0.141	0.144	0.125
Gradient Boosting	24000-6000	500	0.889	0.892	0.890	0.890
Extra Tree	24000-6000	500	0.882	0.883	0.882	0.881
XgBoost	24000-6000	500	0.922	0.923	0.922	0.922
Random Forest	24000-6000	500	0.919	0.920	0.919	0.918
SVM	24000-6000	500	0.971	0.980	0.971	0.974

achieved in training and testing times. A 70% gain was obtained for SVM algorithm. Thus, text-independent speech recognition is performed much faster with high accuracy. The confusion matrix where we can show a certain number of classes in order to observe the results on a class basis for a problem with 60 classes is as in Figure-3.

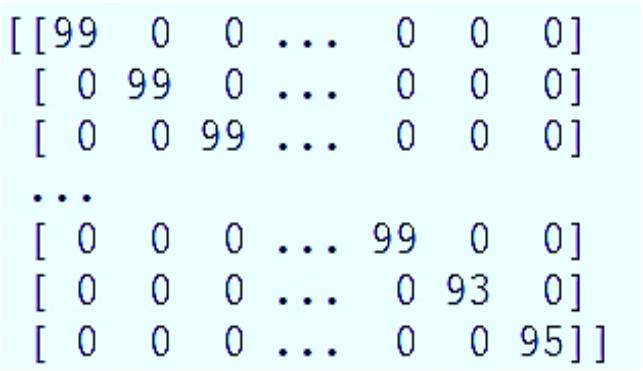

Figure 3. Confusion Matrix of SVM Classification

Since it is a problem with 60 classes, a confusion matrix representing all classes could not be shown. However, as can be seen in Figure-3, the high numbers for all classes are concentrated on the diagonal. This shows that the SVM classifier is successful for all classes.

5. Conclusions

In this study, it is aimed to develop a text independent speaker recognition model. For this purpose, Audio-MNIST dataset was used. After applying a series of pre-processing steps on the dataset, classification was made with 11 different machine-learning algorithms and the results were shown comparatively. Accordingly, successful results were obtained with SVM, logistic regression and kneighbors algorithms in the tests performed on 30000 records for the problem involving 60 different speakers. As a result, 97.1% classification rate was achieved, and the model produced successful results for all classes. In the future deep learning structures will be used to increase the success of the model.

References

[1] Rafizah, H., Khalid, I., and Shamsul., M. (2021). "A review on speaker recognition: Technology and challenges." Computers and Electrical Engineering. doi:10.1016/j.compeleceng.2021.107005

[2] Ertas, F., and Hanilci., C. (2008) "A New Classifier for Speaker Identification System", ELECO 2008, Bursa, Turkey, 28 - 30 November.

[3] Hizlisoy, S., Yildirim, S., and Tufekci, Z., (2020). "Music emotion recognition using convolutional long short term memory deep neural networks." Eng. Sci. Technol. an Int. J. doi: 10.1016/j.jestch.2020.10.009.

[4] Hizlisoy, S., and Tufekci, Z. (2016). "Noise robust speech recognition using parallel model compensation and voice activity detection methods." In Proceedings of the 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), pages 1–4. IEEE.

[5] Karasartova, S. (2011). "Investigation and Implementation of Text-Independent Speaker Identification Systems ". Msc Thesis, Ankara University.

[6] Uzuncarsili, M. (2005). "Investigation of Vector Quantization Based Speaker Identification Algorithms", Msc Thesis, Ankara University.

[7] Kr´al, P. (2010). "Discrete wavelet transform for automatic speaker recognition." In: Image and signal processing (CISP), 3rd international congress on.

[8] Nakagawa S, Wang L and Ohtsuka S. (2012). "Speaker identification and verification by combining MFCC and phase information." IEEE Trans Audio, Speech, Lang Process 20:1085–95.

[9] Hanilci, C. (2017). "Effect of Language Mismatch on Turkish Speaker Verification", Uludag University Journal of Faculty of Engineering, Vol. 22, No.1, doi: 10.17482/uumfd.309477

[10] Celiktas, H. and Hanilci, C. (2017). "A study on Turkish text -

Dependent speaker recognition," 25th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, doi: 10.1109/SIU.2017.7960304.

[11] Soleymanpour, M., and Marvi, H. (2017). "Text-independent speaker identification based on selection of the most similar feature vectors." Int J Speech Technol 2017;20: 99–108.

[12] Jagiasi, R., Ghosalkar, S., Kulal, P., and Bharambe, A. (2019). "CNN based speaker recognition in language and text-independent small scale system." In: Proceedings of 3rd international conference on IoT in social, mobile, analytics and cloud (I-SMAC); p. 176–9.

[13] Chowdhury, A., and Ross, A. (2020). "Fusing MFCC and LPC features using 1D-triplet CNN for speaker recognition in severely degraded audio signals," IEEE Trans. Inf. Forensics Security, vol. 15, pp. 1616-1629.

[14] Jahangir, R., et al. (2020). "Text-independent speaker identification through feature fusion and deep neural network." IEEE Access 2020;8:32187–202.doi: 10.1109/ACCESS.2020.2973541.

[15] Hourri, S., Nikolov, N. S., and Kharroubi, J. (2021). "Convolutional neural network vectors for speaker recognition". Int J Speech Technol 24, 389–400. doi: 10.1007/s10772-021-09795-2

[16] Becker, S., et al. (2018). "Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals", CoRR.

[17] Arslan, R. S., and Barisci, N. (2020). "A detailed survey of Turkish automatic speech recognition". Turkish journal of electrical engineering and computer sciences", 28, 3253-3259.

[18] Xu, Y., Zhang, D., and Yang, J. Y. (2010). "A feature extraction method for use with bimodal biometrics". Pattern Recognition, 43:1106–1115.