2528-9705

Örgütsel Davranış Araştırmaları Dergisi
Journal Of Organizational Behavior Research
Cilt / Vol.: 7, Sayı / Is.: S, Yıl/Year: 2022, Kod/ID: 22S0-859

Speech Identity Recognition using Artificial Neural Network and Support Vector Machine Algorithms

Maryam Zakavi

Master of Electrical Engineering, Electronic Integrated Circuit, Payame Noor University, North Tehran, Tehran ,Iran

ABSTRACT

This study sought to recognize the identity of individuals from their sounds using artificial neural network (ANN) and support vector machine (SVM) algorithms. The speaker recognition was text-independent, which is an exceedingly complex task. Typically, attempts at speech recognition are narrower than other areas of speech processing. As the research innovation, speech recognition was accomplished using multiple features extracted from the acoustic signals. The classification was based on the ANN algorithm and the SVM classifier. Notably, the whole process was evaluated using grid data. In this study, a new text-independent speech recognizing system was proposed. The features selected included Mel-Frequency Cepstral Coefficients (MFCC), energy, and first- and second-order derivative features. The proposed system worked based on the Perceptron neural network and the SVM. At the best state, the system's recognition accuracy for 15 speakers (10 males and five females) was 99.2%.

Keywords: *Speaker recognition, MFCC, Artificial neural network (ANN), Speech processing, Support vector machine (SVM)*

INTRODUCTION

Humans are differentiable by the discrepancies in their voices. The majority of distinction between the speeches of individuals arises from the shape and size of the sound-producing organs, and the remaining portion is due to diverse diction. Besides anatomical features, properties such as dialect and accent, vocabulary, speech speed, and other personal habits (e.g., emotions) contribute to speech variety. Speech voice can be quickly recorded by various devices or transmitted over telecommunication channels. Other technologies and capabilities such as Skype, Google Talk, and Google voice search have further attracted researchers' attention to speech processing.

Speaker recognition is among the most popular and crucial speech processing applications. It helps isolate sound-producing sources, recognizes speakers in meetings, and is helpful in employee attendance systems. These applications are usually multi-user. In multi-user places such as the Internet and cases of an extensive archive of speeches of many individuals, a powerful search engine to find certain speeches is crucial for speech recognition systems.

Speaker recognition is an identity recognition technique based on biometric properties, recognizing and monitoring humans from their speech signals. These methods suffer from problems, including channel and microphone variety and unstable speaker's health, speech, and physical conditions. Therefore, compensating for these problems leads to a mismatch between training and testing results (Kenny et al., 2007a; Vogt & Sridharan, 2008). Several recently

proposed techniques are available to solve these problems when designing speaker recognition systems, which can largely contribute to fixing issues in this area (Mokgonyane et al., 2019; Singh et al., 2020). Speaker recognition covers two main domains, text-dependent, where the text is similarly voiced in both training and testing phases, and text-independent, in which the speaker can speak any word or sentence. Interestingly, the restrictions on text-dependent systems have promoted their accuracy. In addition, text-independent speaker recognition is more challenging to implement.

Speech signal identity recognition has many advantages. There is no need for the individual in person. For instance, the system can recognize the person's voice in monitoring applications even when an obstacle exists between the person and the sensor (microphone). Or, for online banking affairs using smartphones, the user's identity is verified only from the received voice. This task does not require special equipment and needs standard microphones mounted on various devices such as mobile phones and laptops. Models utilized to show the speakers are available in pattern-based and probability-based categories. Pattern-based models are more straightforward, such as the dynamic time-wrapping (DTW) method (Kim et al., 2018) and vector quantization (VQ) (Todkar et al., 2018). Statistics-based models include the Gaussian mixture model (Chakroun & Frikha, 2020) and the hidden Markov model (HMM) (Li et al., 2019)

Speaker recognition encounters multiple challenges, including environmental noise, channel noise, emotions during speaking, and sound recording systems with different qualities, diminishing the recognition systems' accuracy.

Noise is among the factors that significantly reduce the efficiency of speaker recognition systems. It reduces system performance at both aligned and non-aligned modes. Therefore, speech recognition has explicitly addressed the noise problem (Ming et al., 2007). Thus, achieving a proper solution to fix these obstacles and provide measurable results even in the presence of these problems has persistently been a subject of interest.

This study deals with text-independent speaker recognition with associated critical conditions and complexity. Typically, works on speech recognition are more limited than other areas of speech processing. As the research innovation, speech recognition was accomplished using multiple features extracted from the acoustic signals. The classification was based on the ANN algorithm and the SVM classifier. Notably, the whole process was evaluated using grid data.

Evaluation of the proposed algorithm

1. Database

The database utilized in this study has been designed by scholars at the University of Sheffield in the UK. This database is freely released, known as Grid, and is among the limited accessible databases available to researchers. Grid is a multi-purpose database designed to collect and provide data to develop speech recognition systems through individual speechreading. The database contains 1000 short one-second sentences from 34 speakers (18 males and 16 females). Grid enables implementing other scenarios such as gender recognition or speaker recognition. This study employed 15 speakers to evaluate the system (five females and ten males). The speaker recognition systems were designed using 6 to 12 seconds speeches, while speaker recognition was carried out using one-second sentences. The designed database was a high-

quality signal, having a 24-bit rate and a sampling frequency of 25 kHz. In creating the speaker recognition system, 700 sentences were used to teach the system, and 300 one-second sentences were utilized to evaluate the system.

2. Evaluation criteria

Word recognition rate (WRR) is a criterion utilized in many speech recognition systems. In a multi-class categorizing system, each output denotes one of the categories $i \in \{1,\ldots,K\}$. The i^{th} output rate indicates the probability of $P(i \mid o)$, implying that the test vector O belongs to the i^{th} category. The accuracy and efficiency of speaker recognition systems are based on the WRR criterion, indicating the ratio of the system's correct diagnoses to all diagnoses, as follows:

$$WRR = \frac{\text{Accurately diagnosed expressions}}{All\ the\ test\ expressions}$$

3. Feature extraction

Mel-Frequency Cepstral Coefficients (MFCC) feature extraction is a popular method in speech recognition systems. MFCC enables extracting signal push information, which is critical in speech processing. The signal, which refers to the human speech, is first split into 25 ms frames. These frames are extracted in 10 ms steps to avoid omitting any information from the audio signal by overlapping. Besides these features, using the energy feature and the transient signal information obtained by the first and second derivatives is helpful.

The other feature utilized in this study is the energy of each frame, which is obtained based on the amount of each sample in the frame. Accordingly, besides 12 MFCC features, one energy feature is extracted for each frame, implying the extraction of 13 stationary features from each frame.

4. Proposed MFCC based system

The block diagram of the speaker recognition system is illustrated in Figure (1).

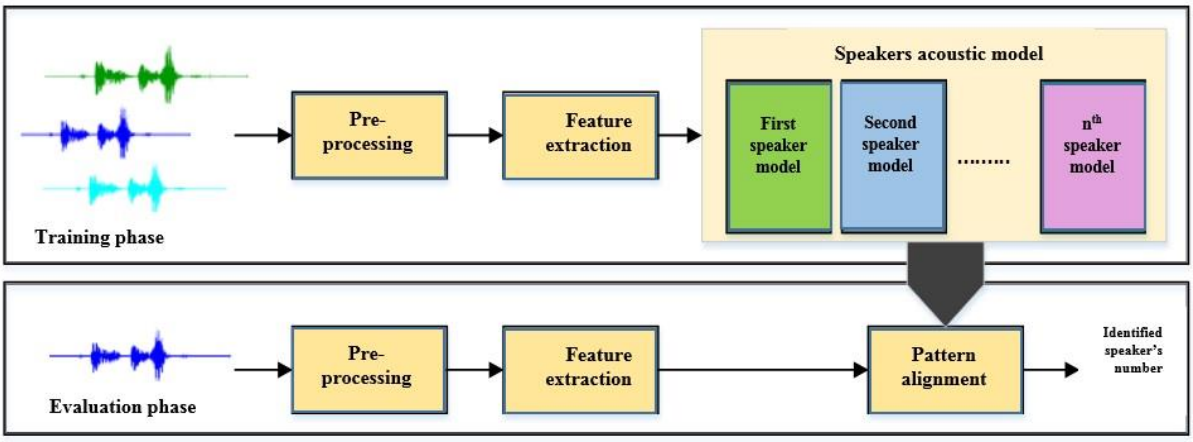

Figure 1. Block diagram of the speaker recognition system

In this method, each speaker is trained by a separate model using the ANN in the training phase. Following constructing the acoustic model of each speaker, the desired features are extracted for each acoustic file in the evaluation stage and are aligned to the acoustic models of each speaker. The model having the highest output score is identified and selected as the recognized speaker. The ANN employed contains neurons in its output layer equal to the number of speakers, where displaying each neuron indicates recognition by the system. According to studies, selecting three consecutive frames in the ANN input layer achieves the best performance. Thereby, the number of input neurons is three times the number of features extracted from each frame. The middle or hidden layer is the system's central processing unit, where each category is learned and modeled. According to investigations, 1000 hidden neurons best compromise performance, accuracy, and computational complexity.

The first proposed system was designed using MFCC features and the ANN algorithm. This feature was extracted in two forms, i.e., extracting 39 and 12 features from each frame separately. The Block diagram of the proposed system is shown in Figure (2).

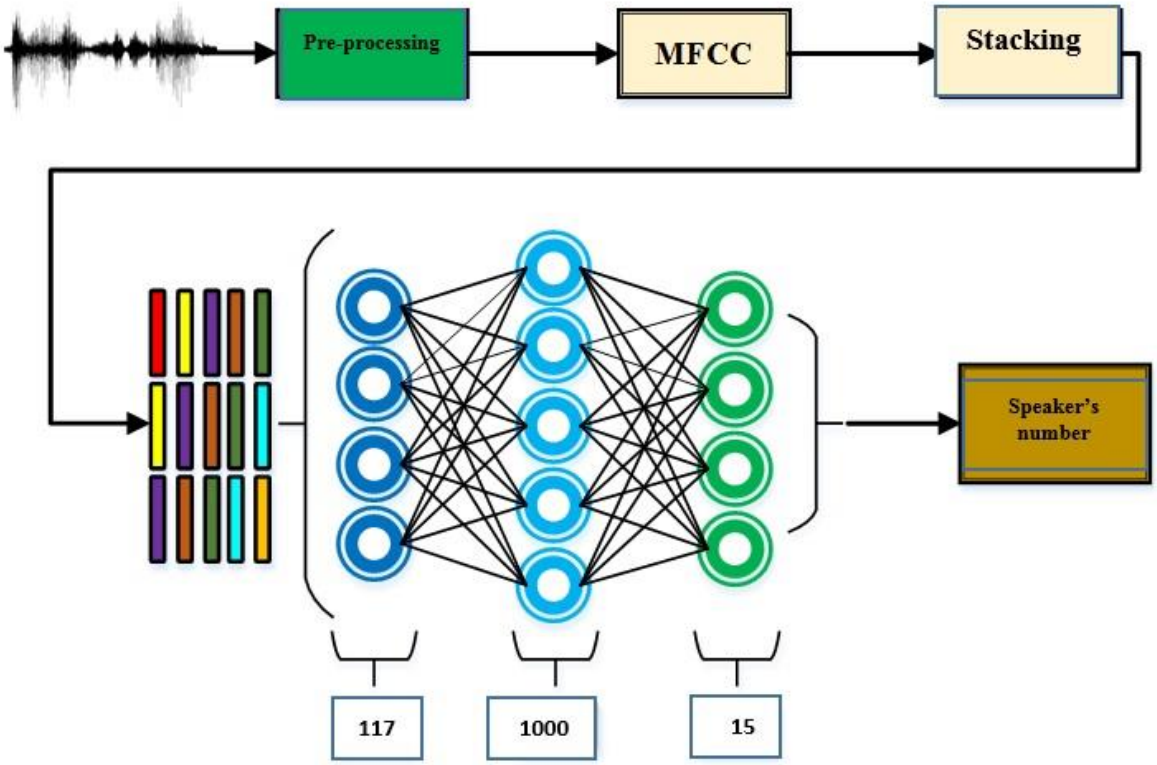

Figure 2. Block diagram of the MFCC-based system

Table 1 gives the results relevant to the MFCC-based system.

Table 1. Recognition rate (%) for each speaker in the MFCC-based system

Speaker's no.	MFCC 12	MFCC 39
1	100	100
2	89.3	100
3	74	98.8
4	100	100
5	95	98.6
6	100	100
7	100	100
8	95	97.3
9	92.6	100
10	100	100
11	93.3	99.6
12	95.6	100
13	94.6	98.6
14	90.6	97
15	91	97.6
Mean	94.1	99.2

According to the results, the proposed system had many errors in some categories, reducing system efficiency. Thus, the system's overall recognition accuracy was 99.2%. A decrease in system accuracy is due to its non-optimal features. Table 2 summarizes the system results.

5. The proposed SVM-based system

The SVM-based system parameters are the same as the previous, but the system is delivered to an SVM for categorizing after extracting the MFCC features and stacking. This study employed Gaussian and binary functions as the backup vector kernels. The proposed system is shown in Figure (3).

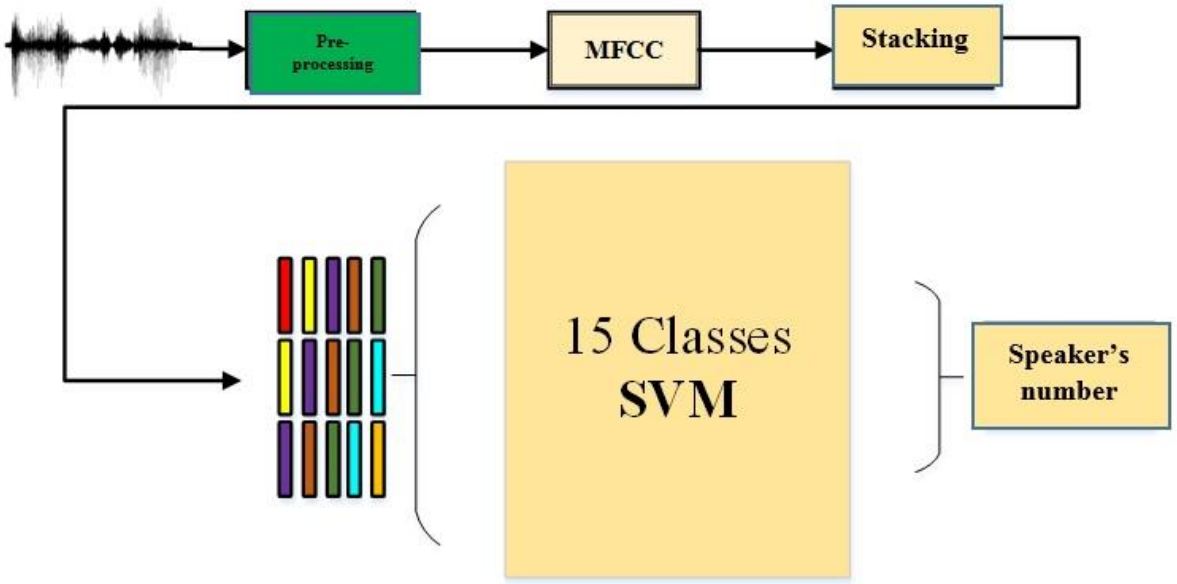

Figure 3. The use of SVM in the proposed system

Table 2 presents the results of the proposed systems. As given in Table 2, the proposed system, having a Gaussian kernel function, possesses a better performance with its recognition accuracy of 95.73%, indicating that using the Gaussian function can better model the behavior of features. In the first scenario, 39 features had the best performance. The SVM-based system only employed 39 features. The system performance was improved significantly when using the ANN algorithm with 39 features, with an accuracy of 99.2%, which is acceptable among speaker recognition systems.

Table 2. Speaker recognition (%) in the SVM-based recognition system

Speaker's no.	Gaussian kernel	Binary kernel
1	100	100
2	96.9	91.6
3	92.3	90.4
4	100	96.6
5	90.9	86.6
6	92.4	89.3
7	93.7	87.3
8	96.4	91.6
9	100	99.3
10	99.3	95.3
11	88.3	86.9
12	100	100

13	89.6	86.4
14	96.6	92.6
15	99.6	96.3
Mean	95.73	92.8

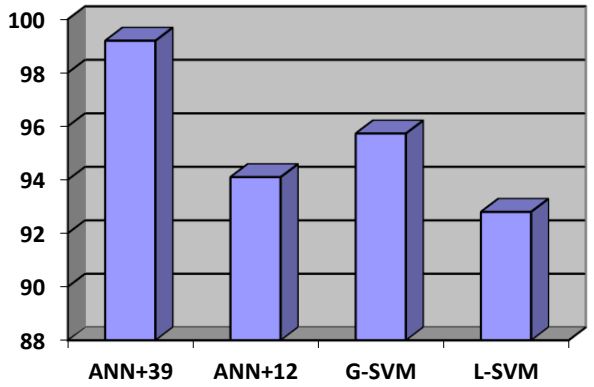

Figure 4. Comparison of the proposed systems

Table 3. Computation duration of the speaker recognition system for 15 speakers				
	MFCC ANN 12	MFCC ANN 39	Gaussian SVM	Logical SVM
Processing duration (s)	10.62	12.17	6.2	6.7

6. Computation duration

The simulations were carried out in a system powered by a core i7 processor and 8GB of RAM. The computational duration for each system was obtained by averaging the processing durations in each case. Table 3 gives the computational period in seconds, i.e., the course takes for the system to process all 300 files. When using the ANN algorithm, the system needs more time to process.

7. Comparison

Table 4 compares the results of the proposed method with those reported in other studies. The proposed system has achieved better results using more appropriate and efficient features and more robust categorization.

Table 2. Comparison of the proposed method with those reported in other studies			
Methods	Features	Number of speakers	Recognition (%)
GMM-based method (Barker et al., 2010)	MFCC	12	98.2

Decision-making threshold (DMT) based method (Hollien et al., 2016)	Based frequency Transient feature vowel letters spectrum	10	98.5
Deep learning based method (Damirchi et al., 2021)	MFCC	15	99.8
The proposed method	MFCC + ANN	15	99.2

Conclusion

Concerning previous research, this study proposed a system to recognize the speaker where the number of speakers has increased to 15. A system was first proposed based on the ANN algorithm and MFCC features. It was restricted to deciding on the speaker's speech signal within 1 s, which is a short evaluation period. Designed for 15 speakers, the system was trained using a fragment of the Grid database. First, the developed system was evaluated using MFCC features and the ANN algorithm, achieving a recognition accuracy of 99.2%. It was then assessed using MFCC features and an SVM as a classifier, achieving a recognition accuracy of 95.73% at best. By comparison, the best computational duration was achieved when using the SVM-based recognition system.

References

1. Barker, J., Ma, N., Coy, A., & Cooke, M. (2010). Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. *Computer Speech & Language, 24*(1), 94-111.
2. Chakroun, R., & Frikha, M. (2020). Robust Text-independent Speaker recognition with Short Utterances using Gaussian Mixture Models. 2020 International Wireless Communications and Mobile Computing (IWCMC)
3. Damirchi, H., Seyedin, S., & Ahadi, S. M. (2021). Improving the loss function efficiency for speaker extraction using psychoacoustic effects. *Applied Acoustics, 183*, 108301.
4. Hollien, H., Didla, G., Harnsberger, J. D., & Hollien, K. A. (2016). The case for aural perceptual speaker identification. *Forensic Science International, 269*, 8-20.
5. Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007a). Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(4), 1448-1460.
6. Kim, H., Ahn, C. R., Engelhaupt, D., & Lee, S. (2018). Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement. *Automation in Construction, 87*, 225-234.
7. Li, T., Choi, M., Fu, K., & Lin, L. (2019). Music sequence prediction with mixture hidden markov models. 2019 IEEE International Conference on Big Data (Big Data)

8. Ming, J., Hazen, T. J., Glass, J. R., & Reynolds, D. A. (2007). Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(5), 1711-1723.
9. Mokgonyane, T. B., Sefara, T. J., Modipa, T. I., Mogale, M. M., Manamela, M. J., & Manamela, P. J. (2019). Automatic speaker recognition system based on machine learning algorithms. 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)
10. Singh, R., Bhattacharjee, U., Singh, A. K., & Mishra, M. (2020). Impact of Noise Levels on SVM-GMM Based Speaker Recognition System. In *Electronic Systems and Intelligent Computing* (pp. 99-107). Springer.
11. Todkar, S. P., Babar, S. S., Ambike, R. U., Suryakar, P. B., & Prasad, J. (2018). Speaker Recognition Techniques: A Review. 2018 3rd International Conference for Convergence in Technology (I2CT)
12. Vogt, R., & Sridharan, S. (2008). Explicit modelling of session variability for speaker verification. *Computer Speech & Language, 22*(1), 17-38.