TEXT-INDEPENDENT SPEAKER IDENTIFICATION SYSTEM USING DIFFERENT PATTERN MATCHING ALGORITHMS

[1]BALSAM Z. KHOJAH, [2]WADEE S. ALHALABI

[1]Department of computer science, College of Computer and Information Systems Umm Al-Qura University
[2]Department of Computer Science, Faculty of Computing and information technology King Abdulaziz University, Jeddah, Saudi Arabia
Department of Computer Science, College of Engineering Effat University, Jeddah, Saudi Arabia
E-mail: bzkhojah@uqu.edu.sa, wsalhalabi@kau.edu.sa

Abstract- Because words are not the only understandable information gathered from speech, voice biometry is a powerful measurement technology. Listeners can know gender, age, health situation, emotion state, and speaker identity. Automatic Speaker Recognition (ASR) represents the ability of a program or device to identify the speaker from its utterance. ASR is usually divided into Speaker Verification (SV) and Speaker Identification (SI). In this paper we propose a Speaker Identification System (SIS) that identifies effectively all registered speakers based on their speech. SIS is composed of two main modules: feature extraction and feature matching. Mel Frequency Cepstrum Coefficients (MFCCs) is used for extracting features from speech signal. For feature matching, we have applied four common ASR algorithms: Vector Quantization (VQ), Gaussian Mixture Model (GMM), Artificial Neural Networks (ANNs), and Decision Trees (DT). When evaluating performance, fusion method was the best accurate procedure for identifying true speakers and rejecting imposters. Fusion method has an accuracy rate of 99% for 100 speakers and 96% for 25 and 50 speakers. The proposed SIS in this study proved its ability to identify a text-independent closed set of speaker groups effectively.

Index Terms- Speaker Recognition, Speaker Identification, Vector Quantization, Gaussian Mixture Models, Artificial Neural Networks, Decision Trees, Fusion.

I. INTRODUCTION

Communication with other people around the world for different purposes is a daily necessity of our lives. Due to the rich information that can provide, speech is the most powerful form of communication. The information provided by speech is vast and includes gender, attitude, emotion, health situation and identity of a speaker. These aspects can be extracted using speech processing techniques. Speech processing is a diverse field with many applications and Speaker Recognition (SR) area is one of these applications that has a great research attention and rapid advancements with the increase of human-machine interactions. Automatic Speaker Recognition (ASR) is the ability to extract, characterize and recognize the information about speaker identity [1]. SR is usually divided into Speaker Identification (SI) and Speaker Verification (SV). Both SI and SV need a stored speaker database as a reference model for pre known speakers. In SV, there is an identity claim and the SV job is to verify the claimed identity through binary decision for acceptance or rejection [2, 3, and 4].

The main applications of SV are in the security area. In case of SI, there is no identity claim, and the SI job is to identify the unknown speaker from groups of known speakers [2, 3, and 4]. The Speaker Identification System (SIS) is divided into closed set and open-set system [2]. Closed-set identification means the unknown speaker must have pre-recorded data in the speaker database and the result should provide the best match between the unknown speaker and the database. Open-set identification means that unknown speaker does not find a match from enrolled speakers and a possible result it is an impostor and we have to add the category "unregistered" to the system [4]. According to speech modalities, SV and SI can be text-dependent or text-independent. Text-dependent implies that, in order to perform the identification, a fixed text must be spoken. In text-independent mode, the speaker can speak freely. Although text-independent adds more flexibility to the system, it reduces the accuracy and is more vulnerable to mistakes.

Any SIS has three steps: feature extraction, training, and testing. There are several algorithms can be used for identification procedure. This paper proposes a Closed-set SIS based on a text-independent approach.

Many Saudi Arabia universities have a separated section for female students. The teacher is present in another location and the only way for communication is audio. When the student wants to participate to class, her voice is the only biometric available to the teacher. Due to the great development of educational technologies that replaces face-to-face learning such as e-learning, several techniques are needed to achieve a successful learning process. In our university for example, having a program or a device that acts as a SIS can assist the teacher in identify who is participating or attending the class today. This paper proposes a SIS that identifies a closed-set of speakers. The main goal is to use well-known SI algorithms for training and identification purposes and upgrade the identification results so that the proposed system can be efficiently used for SI applications such as lab assistant. To achieve this goal, the research depended on training and testing data on three different speaker corpora that can be

used in speech researches. To extract features, Mel Frequency Cepstrum Coefficients (MFCCs) are applied. Then, four different SR algorithms are used to perform the feature matching part. These are Vector Quantization (VQ), Gaussian Mixture Models (GMMs), Artificial Neural Networks (ANNs), and Decision Trees (DT) algorithms. After that, fusion method is used to evaluate the SIS proposed here.

The paper is organized into six sections. Section II focuses on ASR concepts in general and the basic structure of SISs and their modules. Section III discusses in details the methodology followed in this study. In section IV, results findings are illustrated along with a detailed analysis of results. Section V concludes the work, with a special emphasis on results and limitations. Also, some directions for future work are suggested.

II. OVERVIEW OF SPEAKER IDENTIFICATION SYSTEM

Due to the rich information it contains, speech is a powerful source for communicating because it is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic [5]. Speech processing is the extraction of the needed information from a speech signal to be digitalized and processed by the computer [6]. As speech signal is a carrying message of diverse information, speech processing field has different applications depending on the kind of information we are interested in. Analysis, recognition, and coding are the main fields of research in speech processing [7]. Recognition field in particular is the area that has been most studied for several decades. As mentioned in [8], recognition system is divided into three subsystems: Speech, Speaker and Language recognition systems.

A. Speaker Recognition Principles: SR is defined as the identification of a speaker from his/her speech features [9]. To build a SR system we have to automate the speech waveform by using some measurements capable of identifying who is speaking. The main goal of the ASR is to extract, characterize and recognize the information about speaker identity [10]. A SR system is usually divided into SI and SV systems. The data needed in these systems is categorized into text-dependent and text-independent methods. The following sections will explain some of these principles in details.

1) Speaker Verification: The most popular branch in SR is SV because it has a useful contribution in security and access control. SV is the process of verifying whether the person who claims to be has the same speaker identity with the one already enrolled in the system [7]. It can be considered as true/false binary decision because it is a one-to-one comparison between the claimed speaker and the preregistered speaker as explained in [5]. The decision of accepting or rejecting the claimed identity should be chosen carefully to avoid type-I and type-II errors [11]. Type-I error or False Rejection (FR) occurs when the verification system rejects a speaker who is registered in the system. Type-II error or False Acceptance (FA) occurs when the verification system accepts an imposter as a true speaker. Most of commercial SR applications rely on verification for security purposes like bank services and communication control access.

2) Speaker Identification: SI is more complicated than SV due to the need of comparing the entire group of preregistered speakers' voices with the input speaker. So, SI is the process of finding the identity of an unknown speaker by comparing his voice with voices of a group of registered speakers in the database [5, 7]. This system can be divided into closed set and open-set system. Closed-set means that the unknown input speaker must be one of the enrolled speakers that are already stored in the database of the system. Open-set means that the unknown input speaker may or may not be from the preregistered speakers in the system and this will add the option unknown to the results [7, 11, and 4].

3) Text-Dependent vs. Text-Independent Systems: Depending on the way of training/testing the spoken text of speakers, SV speech modalities are divided into text-dependent and text-independent [11, 4, 12]. In text-dependent systems, the spoken text is previously known in both training and testing phases, where in text-independent systems the speaker is allowed to speak freely. The system relies on text-independent model and should figure out the distinguishing speech characteristics of vocal sounds that belong to speakers participating in that system [11]. On the other hand, the text-dependent system needs specific phrases to be spoken (like card number, passwords etc.) [5] Which leads to better speaker recognition results. Even if the text-independent system does not know the spoken words or phrases, it provides flexibility in speaking and there is no need for co-operation from speakers.

B. Basic Structure of Speaker Identification System Any SIS has two basic modules: feature extraction and feature matching (pattern matching) [13, 11, 7, and 4]. Feature extraction is the process of extracting set of features from a speech signal that represent a specific speaker. Feature matching is the actual procedure of recognition, which finds the best match between the extracted features of unknown input voice signal and the set of predefined speakers stored in database of the system. The speaker identification process has two phases: enrollment phase (training) and identification phase (testing) [5, 14, and 7]. In the training phase, speakers enroll into the system by collecting their voices to build a reference model for each speaker, then store the models in a speaker database. In the testing phase, to make a decision on speaker identity, a test voice of an unknown speaker is entered into the system and compared with all the reference models registered

during the training phase. The comparison can be done by probability density estimation like in GMM model or by measuring the distance like in VQ model [14, 15, and 13]. Both training and testing phases should start with feature extraction module to extract the speaker dependent features from the speech signal and that's why in some works, the processes are divided into three processes: feature extraction, training, and testing [11, 13]. In SIS, the main components block diagram of training and testing phases is shown in
Figure 1 [11].

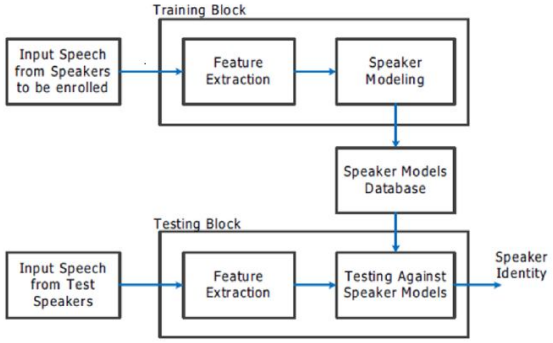


Fig. 1. Block Diagram of an SIS.

C. Feature Extraction: Feature extraction or front-end processing aims to convert the speech waveform to some type of parametric representation for further analysis and processing, in order to produce the speaker discriminative features. The speech signal characteristics are stationary for short periods (between 5 and 100 mSec). However, these characteristics start to change to reflect speaker-specific information for longer periods (1/5 sec or more) [16]. Therefore, the speech signal is a slowly time varying signal and, to characterize the speech signal, we use the most common technique: short-term spectral analysis [17]. This technique carries the speech in frequency domain with short segments of speech through a sequence of analysis, as explained as follows [7, 16]:

1) Signal Pre-Processing: Before extracting feature vectors from the speech signal, a pre-processing procedure should be done. When the speaker speaks into the microphone, the continuous speech signal is converted into the discrete domain. Then, segmentation is done on the speech signal to get overlapped frames, in order to obtain quasi-stationary units of speech. A pre-emphasis filter is applied to each frame to enhance the speech by lifting the high frequency spectral components of speech. Now, the speech signal is ready for the feature extraction subsystem.

2) Mel Frequency Cepstrum Coefficients: MFCC is the most common feature extraction technique used for SI. The technique of computing MFCC is based on short-term analysis, and from each segmented frame a MFCC vector is computed [5]. The identification process is mainly affected by the feature extraction step.

The MFCC processor performs the following processes as illustrated in the block diagram in Figure 2 from [5]:

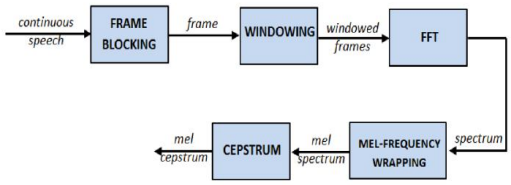


Fig. 2. Block Diagram of MFCCs Processor.

1) Framing: Framing means partitioning a continuous speech signal into adjacent number of segments or frames. Overlapping frames with 30% to 50% are done to avoid losing any information [18].

2) Windowing: To minimize the discontinuities in the signal due to framing, windowing is applied at the beginning and end of each frame [13]. This is done by multiplying each frame with a window function w (n) of length N, where N is the length of the frame [19]. Several window functions can be used. Typically, hamming window is used because it has accurate spectral estimation for its frequency response [16]. The hamming window function the edges of the window [20, 21]:

$$w(n) = 0.54 - 0.46\cos\frac{2\pi n}{N-1}, \quad 0 \le n \le N-1$$

If we define the signal before windowing as x (n), the result of windowing will be the signal y (n) where:

$$y(n) = x(n)w(n), \quad 0 \le n \le N-1$$

3) Fast Fourier Transform: The aim of this step is to convert each frame from the time domain to the frequency domain to obtain the magnitude of the frequency response from each frame [13]. Fast Fourier Transform (FFT) is a name given to fast algorithms to compute the Discrete Fourier Transform (DFT) [22]. The FFT algorithm reduces the time complexity when calculating DFT from $O(N^2)$ to O(nlgn) [23]. FFT is defined on the set of N samples xn, as follow [5]:

$$X_n = \sum_{k=0}^{N-1} x_k e^{\frac{-2\pi jkn}{N}}, \quad n = 0, 1, \ldots, N-1$$

The resulting sequence xn is interpreted as follows:

$$\begin{aligned} f &= 0, && \text{when } n = 0 \\ 0 < f &< F_s, && \text{when } 1 \le n \le \tfrac{N}{2} - 1 \\ \tfrac{-F_s}{2} < f &< 0, && \text{when } \tfrac{N}{2} + 1 \le n < N \end{aligned}$$

Fs denotes the sampling frequency. The result after this step is a spectrum of the signal.

4) Mel-Frequency Wrapping: The goal of this important step is to convert the frequency spectrum to Mel spectrum. The signal now is in actual frequency, f, measured in Hz. Because the human ear frequency perception is non-linear, we need to convert it into Mel-frequency by "Mel-scale" [24]. To compute the "Mel" from the given frequency "f", the following formula is used:

$$mel(f) = 2595 \log_{10}(1 + \frac{f}{700}),\ f: \text{frequency in Hz}$$

To simulate the human perception, a filter bank is built with the bandwidth given by the Mel scale and then, the spectral signal is passed through the bank filters to produce the Mel-frequency spectrum [25]. The filter banks are a set of triangular windows spaced uniformly on the Mel scale. This Mel-frequency wrapping step keeps only the part of useful information [13].

5) Cepstrum: To get the result that is called Mel Frequency Cepstrum Coefficients, the final step is converting the log spectrum back into the time domain. To obtain the Cepstrum, first a logarithmic power spectrum is calculated to be the new analysis window, and then, an inverse FFT is performed to get a signal in the time domain [5]. Therefore, cepstrum is defined as the inverse Fourier transformation of the logarithm of the magnitude of the Fourier transformation [26] and it is denoted mathematically as:

C (n) = ifft log | fft(s (n)) |

Discrete Cosine Transform (DCT) is used to convert the Mel spectrum and their logarithm into the time domain. The MFCCs is calculated using this equation [5]:

$$\tilde{c}_n = \sum_{k=1}^{K} \log \tilde{s}_n \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right],\ n = 1, 2, \ldots, K$$

III. PROPOSED METHODOLOGY

The goal of the work proposed here is to identify a closed set of speakers from different speech corpora and to prove the identification accuracy with different inputs. The speakers recorded their utterances under different conditions and have various dialects of English language. The proposed SIS is composed of two main modules: feature extraction module and pattern matching module. The feature extraction module is done using MFCCs. Then, pattern matching module is applied with VQ, GMM, ANN, and DT algorithms to compare the identification results. Finally, the system tries to improve the identification rate by fusion the results outcome from the four classifiers by applying the majority rule.

Majority Rule Fusion: In SR area, relying on a single feature extraction method or a single pattern classifier does not always come out with the desired high identification rate. To compensate for this issue, a new direction in SR research (called fusion) has been proposed [31, 32]. The idea of fusion is to get the final decision depending on multiple feature extraction methods [33] or multiple classifiers [34, 35] to improve the identification. Data fusion combines scores from these different models trained for a speaker. However, if all models agree on the same type of error, no improvement occurs from fusion [36]. So, some degree of un-correlation between models is needed to improve the identification rate. There are different data fusion techniques that depend on the type of information to be combined. Linear or log opinion pools can be used if we are dealing with probabilities as results [36, 37]. If the results are class labels, voting can be used [38]. In our study, we will follow the voting procedure by majority rule because the outputs are speaker labels. The majority test rule is based on the voting procedure, which means that the score for success rate will be based on the majority of votes received by a testing feature vector during the recognition process. This is done as following: i) if three or all classifiers voted for a certain speaker identification number, then the speaker is identified by the majority voting, or ii) if two classifiers voted for a certain speaker identification number and the other two classifiers disagree on the speaker identification number, then the speaker is identified by the majority voting. Majority voting fail to produce correct result in two cases: i) all four classifiers disagree on the same speaker identification number, or ii) two classifiers agree on certain identification number where the other two classifiers agree on another identification number. Figure 3 illustrates the general structure of the text-independent speaker identification system that is proposed here.

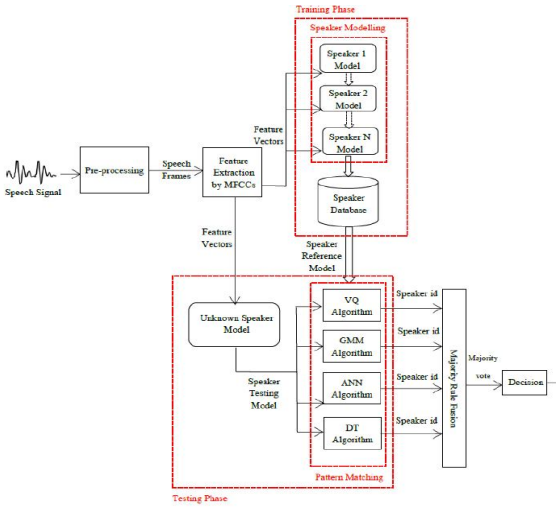


Fig. 3. Architecture of Proposed Speaker Identification System

IV. RESULTS AND DISCUSSION

The experiments in this study are performed using MATLAB R2012a (win64) with operating system version:

Microsoft Windows 7 Home Premium Service Pack1, with the processor: Intel(R) Core(TM) i5- 2450M CPU @ 2.50 GHz., and 4GB system memory. To perform the simulations, three methods are employed. The first one is applied on 10 to 120 speakers to test the accuracy percentage of identification.

The procedure starts with MMFCs feature extraction followed by feature matching algorithm. As mentioned earlier we are applying VQ, GMM, ANN, and DT algorithms. Then, an additional step is added to these results by fusion the four algorithms according to majority rule. In the second approach, we are randomly choosing six groups of 25 speakers and re-apply the same procedure of the first experiments method. These groups are labeled with A, B, C, D, E, and F. The goal of these two methods is to compute the accuracy of identification (%) applied to various speaker groups using different Speaker Identification algorithms according to the following formula:
Identification Rate of SIS=

$$\frac{\text{No.of Correctly Identified Samples}}{\text{Total No.of Tested Samples}} * 100$$

The third method is applied on a group of speakers containing a mixture of true speakers and imposter speakers. This is done to check the performance of the proposed classifiers by using some evaluation measures. To find these parameters we need to compute the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

1) Identification Rate: The main goal of this study is to reach to a system that effectively identifies a person from his speech. To do that, for the training and testing chosen speakers, we depend on common and well known speaker identification algorithms. These algorithms are evaluated based on number of correctly identified test samples divided by the total number of test samples and then multiplied by 100. To measure the accuracy of our experiments, the identification rate is computed. Figure 4 shows the different identification rate for 10 to 120 speakers when applying VQ, GMM, ANN, DT, and majority decision fusion method. Figure 4 illustrates that the best identification rate we got is from GMM then fusion method because both of them maintain identification rate above 90% for 50 speakers and above 60% for 120 speakers.

VQ results are near to the fusion results until 80 speakers then from 90 to 120 speakers, the results are worse than fusion. ANN has stable results of identification rate from 30 to 60 speakers and from 80 to 110 speakers. Although DT scores better than ANN from 10 to 50 speakers, it has the worst identification rate results from 60 to 120 speakers.

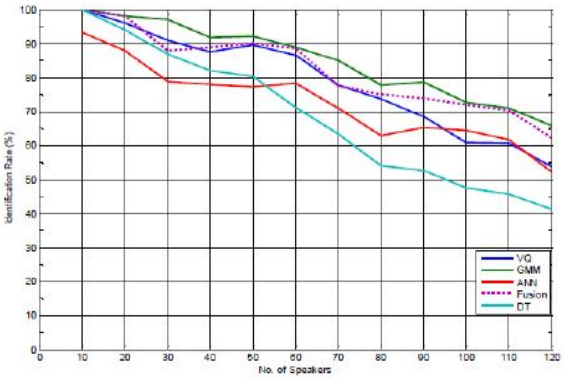


Fig. 4. Identification Rate for 10 to 120 Speakers by Proposed SIS

It is clear from the previous observations that identification rate scores a high rate (between 85% to 100%) when the number of trained speakers is small. Identification rate scores become worse with large number of speakers. The second experiments method was to take groups of 25 speakers randomly from the same speaker database of first method experiments. We labeled six groups with A, B, C, D, E, and F and applied the same four algorithms plus the fusion method to analyze the difference in identification rates. Figure 5 shows the identification rate results of the six groups. From results of groups of speakers when applying VQ, GMM, ANN, DT, and majority decision fusion method, we observed the following: In general, the best identification rate results come from GMM then fusion method. All the four algorithms plus fusion method give the best identification rate results with group A and B of speakers (between 90% and 100%). ANN gives the worst identification rate results with group C. DT gives the worst identification rate results with group F. Since the number of speakers is the same for all groups, the identification rate may differ from a group to another according to the speakers environment, noise, and training time. By taking the average identification rate from all groups A, B, C, D, E, and F for four algorithms plus fusion method, GMM gives the best identification rate, then fusion, then VQ, after that ANN, finally DT as depicted in Figure 6.

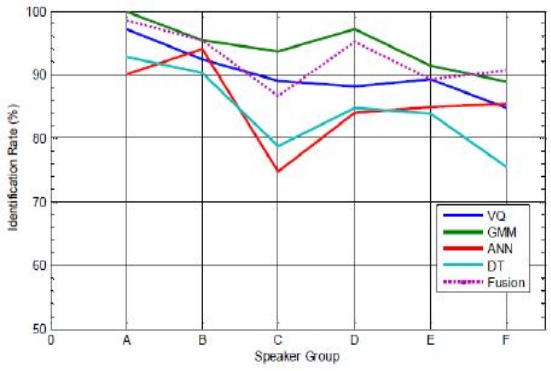


Fig. 5. Identification Rate for Speakers Groups by Proposed SIS

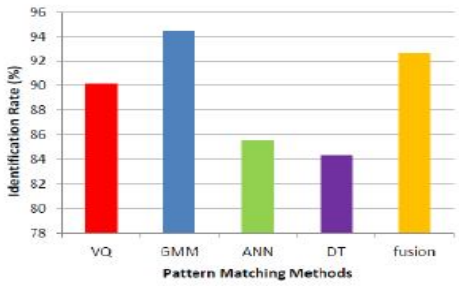


Fig. 6. Average Identification Rate for Proposed SIS over 25 Speakers

2) Speaker Misidentification: Different identification rates indicate that there are some speakers that had not been identified. As shown earlier, as long as the speaker database size increases the number of misidentified speakers increases and this is varies from one algorithm method to another. Figure 7 illustrates the following observations: GMM has the best identification number of speakers because the worst case is 2 misidentification. VQ has almost the same of GMM results but when speaker database becomes 120 speakers, we got 5 misidentification of speakers. Fusion results give us the best identification number of results with 10 to 60 speakers, then with 70 or more speakers: 1 to 4 misidentification starts to appear. When the size of speaker database is 10, all the four algorithms with the fusion procedure identify all the 10. We start to see 1 misidentified speaker by ANN and DT algorithms when the speaker database is between 20 and 30 speakers. Rapid increase in misidentification appears clearly by ANN algorithm with 70 to 120 speakers. Generally, misidentification results start to appear with 80 or more speakers for all the four algorithms plus fusion procedure. The worst result is 26 misidentification with 120 speakers by ANN algorithm. As results start to be not effective when number of speakers is large, we tested another method of experiments by choosing randomly 25 speakers and put them in groups. As shown in Figure 8, misidentification number of six groups labeled with A, B, C, D, E, and F and applied the same previous algorithms to these groups. The worst case is one speaker has not been identified by ANN, DT in group C. Beside group C, groups A, D, and F has one misidentification when ANN algorithm applied. Fusion method has only one misidentification in group

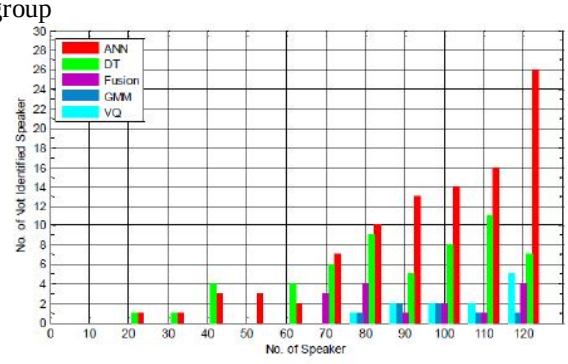


Fig. 7. No. of Misidentification for 10 to 120 Speakers by Proposed SIS

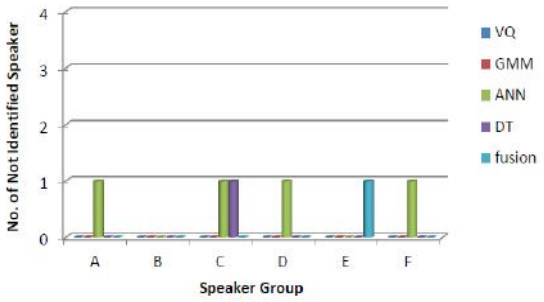


Fig. 8. No. of Misidentification for Speakers Groups by Proposed SIS

CONCLUSION AND FUTURE WORK

In this study, we proposed a SIS that uses different techniques to identify any speaker effectively. The goal of this study is to propose a speaker identification system for an effective identification of all registered persons based on their speech. For extracting features from the speech signal MFCCs was used. After that, VQ, GMM, ANN, and DT algorithms were applied for feature matching. Then, in order to take advantage of the various speaker identification algorithms techniques, the results were fused through majority decision method. Experiments were done on different organizations of speaker databases: the increasing size of speakers from 10 to 120 speakers, and six groups having the same size of 25 speakers chosen randomly. We noticed that identification rate results are affected by increasing the size of speakers databases used. In addition, it is affected by the environment of recorded speech as shown from different identification results from groups A, B, C, D, E, and F of 25 speakers. Overall training and testing speakers in groups with small size achieve better identification rate and less misidentification of speakers. The identification of good approaches for the improvement of speech recognition system is a difficult task, the results obtained showing that by adding majority decision to fuse VQ, GMM, ANN, and DT algorithms the accuracy rate level is improved up to 96% for 25 and 50 speakers and up to 99% for 100 speakers. Identification rate results show that, compared to the other approaches, GMM has the best performance. Also dividing a large number of registered speakers into groups for training gives more stable and higher identification results. Accuracy rate results show that fusion method is the most accurate procedure for identifying true speakers and rejecting imposters. VQ algorithm is the second best in accuracy rate results.

a) Limitations: As shown from simulations, when we depend on a single speaker identification algorithm such as VQ or ANN or DT, we are facing lower identification rate which results in a higher misidentification number of speakers than fusing the results. To compensate this issue we used four different speaker identification techniques to create un-correlations between results and get the most

benefit from each algorithm by majority rule decision. We suggest that the main reason for the decreasing identification results is the increasing number of similar features from different speakers. This is lead to make the identification process more difficult and identifies speakers mistakenly. One choice is to increase the training time for each speaker to get more unique features. Another way is to use more than one feature extraction technique to produce diverse features. As we are aiming to build a real time system as a lab assistant, it is better to parallelize running the four speaker identification algorithms to produce real time results for fusion. This will be more efficient than applying each algorithm separately. However, in this study we did not use the parallelism method. This can be done in future work.

Recommendations for Future Work: As this study relied on an English language speaker database, the performance of the proposed system can be tested on another language such as Arabic. In this phase of work, the SIS goal was to identify speakers by their identification number. We can add the gender, age, emotional state and other factors for identification. Speaker identification is a part of pattern recognition so we can upgrade the performance by adding another biometric for identification such as face, or fingerprint.

Also using another feature extraction method such as Linear Predictive Coding (LPC), Inverted Mel Frequency Cepstrum Coefficients (IMFCCs), etc., can produce different results. Fusion method can be done depending on various theories and can be applied to two or more feature extraction techniques or pattern matching techniques. Speaker corpora can be changed or add some factors as clean and noise environment. In addition, another improvement can be represented by the identification of a speaker when multiple speakers speak simultaneously. Parallelism technique can be applied on several pattern matching models to upgrade the performance in real time systems.

REFERENCES

[1] D. A. Reynolds, "An overview of automatic speaker recognition," in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075), Conference Proceedings.

[2] C.-T. Hsieh, E. Lai, and Y.-C. Wang, "Robust speaker identification system based on wavelet transform and Gaussian mixture model," J. Inf. Sci. Eng., vol. 19, no. 2, pp. 267–282, 2003.

[3] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A real-time text-independent speaker identification system," in Image Analysis and Processing, 2003. Proceedings. 12th International Conference on. IEEE, Conference Proceedings, pp. 632–637.

[4] Q. Jin, "Robust speaker recognition," Ph.D. Thesis, Carnegie Mellon University, , Language Technologies Institute, Pittsburgh, PA, 2007.

[5] A. Rajsekhar. G, "Real time speaker recognition using mfcc and vq," Master's Thesis, National Institute of Technology, 2008.

[6] J. P. Campbell Jr, "Speaker recognition: a tutorial," Proceedings of the IEEE, vol. 85, no. 9, pp. 1437–1462, 1997.

[7] L. Feng, "Speaker recognition," Master's Thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2004.

[8] D. A. Reynolds and L. P. Heck, "Automatic speaker recognition," in AAAS 2000 Meeting, Humans, Computers and Speech Symposium, vol. 19, Conference Proceedings, pp. 101–104.

[9] B. Squires and C. Sammut, "Automatic speaker recognition: An application of machine learning," in ICML, Conference Proceedings, pp. 515–521.

[10] D. A. Reynolds, "An overview of automatic speaker recognition technology," in Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, vol. 4, Conference Proceedings, pp. IV–4072–IV–4075.

[11] Z. Haider, "Robust speaker identification against computer aided voice impersonation," Ph.D. Thesis, University of Surrey, United Kingdom, 2011.

[12] S. Furui, "Recent advances in speaker recognition," in audio- and Video-based Biometric Person Authentication. Springer, Conference Proceedings, pp. 235–252.

[13] T. Mahboob, M. Khanum, M. S. H. Khiyal, and R. Bibi, "Speaker identification using gmm with mfcc," International Journal of Computer Science Issues (IJCSI), vol. 12, no. 2, p. 126, 2015.

[14] D. H. B. Kekre and V. Kulkarni, "Performance comparison of speaker recognition using vector quantization by lbg and kfcg," International Journal of Computer Applications, vol. 3, p. 32, 2010.

[15] A. Abushariah, T. Gunawan, J. Chebil, and M. Abushariah, "Voice based automatic person identification system using vector quantization," in Computer and Communication Engineering (ICCCE),2012 International Conference on. IEEE, Conference Proceedings, pp. 549–554.

[16] D. S. Rodr´ıguez, "Text-independent speaker identification," Master's Thesis, AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY KRAKOW, 2008.

[17] M. R. Hasan, M. Jamil, M. Rabbani, and M. Rahman, "Speaker identification using mel frequency cepstral coefficients," variations, vol. 1, p. 4, 2004.

[18] M. K. Gill, R. Kaur, and J. Kaur, "Vector quantization based speaker identification," International Journal of Computer Applications, vol. 4, no. 2, pp. 1–4, 2010.

[19] J. W. Picone, "Signal modeling techniques in speech recognition," IEEE, vol. 81, no. 9, pp. 1215–1247, 1993.

[20] A. Zulfiqar, A. Muhammad, and A. M. Enriquez, "A speaker identification system using mfcc features with vq technique," in Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on, vol. 3. IEEE, Conference Proceedings, pp. 115–118.

[21] F. Zhuo, S. Choi San, C. Chen, Y. He, and Z. Wang, "Use hamming window for detection the harmonic current based on instantaneous reactive power theory," in Power Electronics and Motion Control Conference, 2004. IPEMC 2004. The 4th International, vol. 2, Conference Proceedings, pp. 456–461 Vol.2.

[22] R. Mukherjee, "Speaker recognition using shifted mfcc," Thesis, 2012.

[23] C. Becchetti and L. P. Ricotti, Speech recognition: theory and C++ implementation. John Wiley & Sons, Inc., 1999.

[24] S. J. Abdallah, I. M. Osman, and M. E. Mustafa, "Textindependent speaker identification using hidden markov model," World Comput. Sci. Inf. Technol. J, vol. 2, no. 6, pp. 1–6, 2012.

[25] H. Beigi, Fundamentals of Speaker Recognition, 1st ed. Springer US, 2011.

[26] L. Rabiner and R. Schafer, Theory and Applications of Digital Speech Processing. Pearson, 2011. [Online]. Available: https://books.google.com.sa/books?id=ME67RAAACAAJ

[27] (2004) Elsdsr english language speech database for speaker recognition. [Online]. Available: http://www2.imm.dtu.dk/_1fen/elsdsr/index.php?page=index

[28] (2006-2015) Voxforge speech corpus. [Online]. Available: http://www.voxforge.org/home/downloads

[29] D. R. McCloy, P. E. Souza, R. A. Wright, J. Haywood, N. Gehani, and S. Rudolph. (2013) The pn/nc corpus. [Online]. Available: http://depts.washington.edu/phonlab/resources/pnnc/

[30] N. Praveen and T. Thomas, "Text dependent speaker recognition using mfcc features and bpann," International Journal of Computer Applications, vol. 74, no. 5,pp. 31–39, 2013.

[31] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. Wiley, 2004. [Online]. Available: https://books.google.com.sa/books?id=B4TZtp7X82oC

[32] H. A. Patil and T. K. Basu, "A new data fusion technique and performance measure for identification of twins in marathi," 2006.

[33] L. Mary, K. S. R. Murty, S. M. Prasanna, and B. Yegnanarayana, "Features for speaker and language identification," in ODYSSEY04-The Speaker and Language Recognition Workshop, Conference Proceedings.

[34] K. Chen, L. Wang, and H. Chi, "Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification," International Journal of Pattern Recognition and Artificial Intelligence, vol. 11, no. 03, pp. 417–445, 1997.

[35] S. Z. Boujelbene, D. Ben Ayed Mezghani, and N. Ellouze, "Application of combining classifiers for textindependent speaker identification," in Electronics, Circuits, and Systems, 2009. ICECS 2009. 16th IEEE International Conference on, Conference Proceedings, pp. 723–726.

[36] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammone, "Speaker recognition—general classifier approaches and data fusion methods," vol. 35, pp. 2801–2821, 2002.

[37] J. A. Benediktsson and P. H. Swain, "Consensus theoretic classification methods," Systems, Man and Cybernetics, IEEE Transactions on, vol. 22, no. 4, pp. 688–704, 1992.

[38] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," Systems, Man and Cybernetics, IEEE Transactions on, vol. 22, no. 3, pp. 418–435, 1992.

★★★