2014 Fifth International Conference on Intelligent Systems, Modelling and Simulation

Effect of Spoken Text on Text-independent Speaker Recognition

Mansour Alsulaiman
Computer Engineering Department, Digital Speech Processing Group
College of Computer and Information Sciences, King Saud University
Riyadh 11543, Saudi Arabia
msuliman@ksu.edu.sa

***Abstract*—Automatic speaker recognition (ASR) is a well-investigated area among the researcher of speech processing. Many of the factors that affect the recognition rate of an ASR system are addressed in the literature. Those factors are clean and noisy environments, cross channel, session variability, age and health of a speaker etc. One of the factors that still needs investigation is the effect of type of spoken text on speaker recognition (SR). In this paper, the developed ASR system is trained and tested with different types of text, to observe the effect of the type of text on SR performance. The text includes digits, words, sentences, and paragraphs. All other factors like session, environment, and recording channel are same for this study. The study concluded that careful design of the training and testing text can increase the recognition rate of an ASR system.**

Keywords-Effect of text, speaker recognition, MDLF, GMM, Arabic, KSU speech database

I. INTRODUCTION

Automatic Speaker recognition (ASR) is a process of finding personal identity by analyzing the speech signal [1]. Different speakers do not have identical voices because of their vocal tract shapes, larynx size and other parts of their voice production organs [2]. The speech of a speaker has his or her features that enable us to recognize the speaker. SR is extensively explored by the speech processing community since last two decades. Recently, interest in SR has increased significantly due to the growing use of speech technologies in various areas of daily life. Many applications have been considered for speaker recognition. These include secure access control by voice, information structuring, customizing services to individuals, and forensic investigation [3]. ASR technologies are expected to make our daily live more convenient by creating new services through access control applications, including voice dialing, telephone banking, teleshopping, database access by voice, reservation services, voice mail and remote access to personal computers [4]. ASR could be text-dependent speaker recognition or text-independent speaker recognition. Compared with text-dependent SR, text-independent SR is more flexible, but is more challenging [5].

ASR is facing many challenges to build efficient systems with high accuracy. Different feature extraction techniques have been proposed to develop better ASR systems. Many enhancements have been used in the preprocessing part such as pre-emphasis, noise removal algorithms, etc., to improve accuracy of automatic SR systems. Some state-of-the-art techniques for pattern classification are Gaussian Mixture Model (GMM) [6] and Gaussian Mixture Model-Universal background Model (GMM-UBM) [7]. These techniques have contributed prominently in the area of ASR. Age and health of a person affect the voice of a person, and these factors degrade the recognition rates of an ASR system. Therefore, session variability is an important aspect in ASR [8]. Due to this, the National Institute of Standards and Technology (NIST) speaker recognition competition include session variability in the competition. Variations in one's own voice due to time, health, emotions and age results in intra-speaker variability. The training and testing samples recorded in different environments, and microphones are another factor that decrease the accuracy of the SR. For example, PYKFEC feature [9] outperformed conventional Mel-frequency cepstral coefficients (MFCC) [10], [11], but when Lawson tested it in cross channels and noisy channels, it did not perform well [12]. Jayanna [13] conducted a study on speaker recognition under limited data condition and found that state of the art technique did not perform well. To observe the performance of an ASR system with the samples of different durations, experiments for text-independent SR are performed in [14] with speech samples of few minutes, and then speech samples of 10sec.

A factor, other than those mentioned above, that also affect the recognition rate is the recorded text. In this paper, the effect of type of spoken text on SR system is analyzed. Various types of text such as digits, words, sentences, and paragraphs are considered in this study. Some types of text are recorded by speakers when they read the text from a text-displaying screen. Another type of text, when a speaker read the text (question) from a screen but answer it spontaneously, is also investigated. The developed SR system is trained and tested with different types of text to observe their effect. The lists of text have different time duration and number of alphabets. All other factors: session, recording channel, and environment are same for all types of text. Multi-directional local features with moving average (MDLF-Mavg) [15] is used to extract speaker dependent characteristics from his/her speech signal. To decide about the identity of a speaker, state of the art modeling technique GMM is implemented.

The rest of the paper is organized as follows: section II describes the speaker recognition system, including speech corpus, features extraction and modelling technique. Section III provides experimental setup, results and discussion on effect of text. Section 4 describes the effect of switching the training and testing text. Finally, section 5 draws some conclusions.

2166-0662/14 $31.00 © 2014 IEEE
DOI 10.1109/ISMS.2014.54

II. AUTOMATIC SPEAKER RECOGNITION SYSTEM

An ASR has two important components. The first is the feature extraction technique, and the second is the modeling technique. The KSU (King Saud University) speech database is used for the analysis of effect of text.

A. KSU Speech Database

The text for the experiments is taken form KSU speech database [16], [17]. KSU database is recorded in three different sessions and contains more than three hundred speakers. The database included male and female speakers of nineteen different Arab and non-Arab nationalities. In the first session every speaker is recorded at three different locations, office, cafeteria and soundproof room, representing different environments. At each location, the speaker is recorded by using different recording channels. Various fixed and variable lists of text are recorded for each speaker. The lists contain digits, words, sentences, and paragraphs. These lists are SAAVB sentences, Arabic digits, common words, rich words, paragraphs, distinctive phonetic words, common words, random words, and answer to questions, abbreviation used for the lists are provided in Table I.

TABLE I. ABBREVIATION OF TEXT MATERIAL

Name of List	Abbreviation for List
SAAVB Sentence-1 and -2	S1, S2
Numbers	NO
Common Words-1 and -2	CW1, CW2
Rich Words-1 and -2	RW1, RW2
Paragragh-1 and -2	P1, P2
Distinctive Phonetic Words-1 and -2	DP1, DP2
Common List	CL
Random List	RL
Answers to Questions-1 and -2	AQ1, AQ2

B. Multi-directional Local Features with Moving Average (MDLF-Mavg)

A new type of features names as MDLF-Mavg is proposed in [15]. The MDLF-Mavg extracts local features in four different directions by taking three-point linear regression (LR) along the time-axis (at 0^o), frequency-axis (90^o), and three point moving average along 45^o and 135^o in the time-frequency plane, which correspond to capturing the onset and offset of phonemes, formants contour and voiceprint of a speaker. MDLF-Mavg extracts forty-eight features from each frame of a speech signal. A hamming window is applied on each frame by using the formula given in (1).

$$s_w(n) = s(n) \times w(n), \quad \text{for } n = 0,1,2,...,N-1 \tag{1}$$

Where s (n) and w (n) correspond to the input voice signal and the window function, respectively. N corresponds to the number of samples in each frame. Fourier Transform (FT) is applied to the windowed signal as in (2).

$$X(k) = \sum_{j=0}^{N-1} s_w(j) \times e^{-j\frac{2\pi jk}{N}}, \quad k = 0,1,2,...,K-1 \tag{2}$$

K is the number of FT points (bins of frequency). FT converts the time domain input signal into the frequency domain signal. After passing speech signal through the 29-channel Mel-filter bank, the log compression is applied. Then four types of three-point operations are computed as in Eq. 3, 4, 5 and 6.

$$\text{Along time: } d_{t,f}{}^{t} = \frac{\sum_{i=1}^{3}\left(c_{t+i,f} - c_{t-i,f}\right)}{2*3} \tag{3}$$

$$\text{Along freq.: } d_{t,f}{}^{f} = \frac{\sum_{i=1}^{3}\left(c_{t,f+i} - c_{t,f-i}\right)}{2*3} \tag{4}$$

$$\text{Along time-freq. at } 45°\text{: } c_{t,f}{}^{45} = \frac{\sum_{i=1}^{3}\left(c_{t-i,f-i} + c_{t+i,f+i}\right)}{2*3} \tag{5}$$

$$\text{Along time-freq. at } 135°\text{: } c_{t,f}{}^{135} = \frac{\sum_{i=1}^{3}\left(c_{t+i,f-i} + c_{t-i,f+i}\right)}{2*3} \tag{6}$$

C. Gaussian Mixture Model

GMM [18] is a state of the art modeling technique that copes more with the space of the features, rather than the time sequence of their appearance. All speakers are modeled by GMMs that represent, in a weighted manner, the occurrence of the feature vectors. The well-known method to model the speaker GMM is the Expectation-Maximization algorithm, where model parameters (Mean, variance and mixture coefficients) are adapted and tuned to converge to a model giving a maximum log-likelihood value.

The GMM model is given by the weighted sum of individual Gaussians

$$p(X \mid \lambda) = \sum_{i=1}^{M} w_i g(X \mid \mu_i, \Sigma_i) \tag{7}$$

where X is a D-dimensional continuous-valued data vector (i.e. measurements or features), w_i are the mixture weights, and g are the component Gaussian densities. Each

component density is a D-dimensional Gaussian function of the form,

$$g(X \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu_i)'\sum_i^{-1}(X-\mu_i)\right\} \quad (8)$$

with mean vector μ_i and covariance matrix Σ_i. The mixture weights satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. The model of the GMM is denoted as $\lambda = (w_i, \mu_i, \Sigma_i)$, $i = 1,2,3,..., M$.

III. Effect of Spoken Text

To analyze the effect of spoken text an ASR is developed by using MDLF-Mavg. It calculates the forty-eight speech features that are fed to GMM to create a speaker model for every speaker. During extraction of MDLF-Mavg, a frame of 256 samples is used with overlapping of 50% with the previous frame. Thirty two Gaussian mixtures are used to model each speaker.

An extended investigation was performed to find the effect of the type or the characteristics of the text used in training and testing on the recognition rate of the ASR system. The investigation used the recording from Yamaha mixer in office at session 2. This investigation is based on the fifteen lists of text recorded in KSU speech database. The number of occurrences of each alphabet in each list for a certain speaker is presented in Table II. These alphabets and vowels constitute the set of Arabic language phonemes. The vowels are available with a high occurrence in every list, so they are not counted, and we concentrated on the alphabets only.

The results are given in Tables III and IV. Table III gives the results when the training is done with the long speech utterances, which are the paragraphs, the common list, the random list, SAAVB sentence-1 and SAAVB sentence-2. The length of the long lists varied from 19 to 42 seconds, as can be seen in Table III. Table IV gives the results when training is done with the short speech utterances, which are numbers, common word-1, rich words-1, answer to questions-1, and answer to questions-2. The length of the short lists varied from 10 to 20 seconds, as presented in Table III. In both tables, the testing is done with the rest of the lists. In Table III and IV, T(S), Tol. L and AV represents time in seconds, total letters, and average recognition rate, respectively. .

TABLE II. Distribution Of The Phonemes In The Text Lists For One Speaker.

Letters	IPA	S1	S2	NO	CW1	CW2	RW1	RW2	P1	P2	PD1	PD2	CL	RL	AQ1	AQ2
ء	/□/	61	46	6	12	10	4	1	44	83	0	0	24	20	21	21
ب	/b/	15	10	2	2	1	2	3	4	13	0	0	5	5	5	5
ت	/t/	13	16	2	0	2	1	1	6	17	0	0	5	2	4	3
ث	/θ/	1	1	4	0	1	0	2	3	1	1	0	1	1	1	1
ج	/g/	8	3	0	0	0	0	0	7	11	0	0	1	1	1	1
ح	/ħ/	4	6	1	1	3	2	0	3	12	1	4	1	1	1	1
خ	/x/	2	2	1	1	0	3	0	1	3	0	0	1	1	1	1
د	/d/	2	10	1	1	0	2	1	5	13	0	0	2	2	2	2
ذ	/ð/	3	1	0	0	1	0	0	3	2	0	0	2	2	2	2
ر	/r/	13	13	2	5	1	2	2	15	20	0	0	5	5	5	5
ز	/z/	0	2	0	1	0	1	2	3	1	2	1	1	1	1	1
س	/s/	6	5	4	1	3	2	1	7	11	1	2	3	3	3	3
ش	/š/	3	1	0	1	1	3	0	2	3	2	1	1	1	1	1
ص	/ş/	0	3	1	0	1	2	2	1	4	2	4	1	1	1	1
ض	/ď/	2	4	0	1	1	0	0	2	4	0	0	1	1	1	1
ط	/ţ/	3	4	0	0	0	0	1	1	5	0	0	1	1	1	1
ظ	/đ/	0	3	0	1	0	0	0	3	1	0	0	1	1	1	1
ع	/□/	7	8	3	3	1	1	2	10	11	5	1	3	3	3	3
غ	/□/	1	0	0	0	1	0	1	4	4	3	1	1	1	1	1
ف	/f/	3	8	1	1	2	1	1	10	14	2	7	3	3	3	3
ق	/q/	6	8	0	1	0	2	2	1	10	0	0	3	3	3	3
ك	/k/	4	7	0	3	1	0	3	5	13	0	0	4	4	4	3
ل	/l/	34	24	1	10	6	1	2	27	44	0	0	13	13	13	12
م	/m/	10	17	2	5	3	1	6	22	19	4	4	11	11	11	10
ن	/n/	12	14	3	2	2	1	3	10	15	5	4	9	8	9	8
ه	/h/	11	6	0	1	4	1	0	16	7	2	1	5	5	5	4
و	/w/	7	13	1	0	2	2	5	17	21	0	0	6	6	6	6
ي	/y/	16	16	1	4	3	3	3	13	27	0	0	5	6	6	6
Sum		**247**	**251**	**36**	**57**	**50**	**37**	**44**	**245**	**389**	**30**	**30**	**119**	**112**	**116**	**110**
Missing letters		**3**	**1**	**11**	**8**	**7**	**8**	**8**	**0**	**0**	**16**	**17**	**0**	**0**	**0**	**0**

TABLE III. RECOGNITION RATE (%) WHEN TRAINING WITH P1, P2, CL, RL, S1, AND S2.

Test Text	T (S)	Tot. L	Train Text						AV
			P1	*P2*	*CL*	*RL*	*S1*	*S2*	
S1	32	247	90	89	81	83	--	98	88
S2	32	251	93	91	84	86	99	--	91
NO	10	35	87	91	86	86	91	91	89
CW1	11	57	95	96	86	90	87	92	91
CW2	11	50	92	94	86	91	84	86	89
RW1	10	37	92	88	82	82	80	86	85
RW2	10	44	91	91	83	85	82	86	86
P1	29	245	--	100	95	95	88	95	95
P2	42	389	99	--	96	95	87	93	94
DP1	10	30	84	90	91	87	71	73	83
DP2	10	30	86	92	95	96	75	79	87
CL	19	119	96	97	--	99	82	87	92
RL	20	112	95	97	99	--	86	88	93
AQ1	20	--	85	92	94	96	77	79	87
AQ2	17	--	80	87	91	91	72	72	82
Average			**90**	**92**	**87**	**90**	**83**	**87**	**--**

TABLE IV. RECOGNITION RATE (%) WHEN TRAINING WITH NO, CW1, RW1, AQ1, AND AQ2.

Test Text	T (S)	Tot. L	Train Text					AV
			NO	*CW1*	*RW1*	*AQ1*	*AQ2*	
S1	32	247	72	80	67	69	55	69
S2	32	251	78	86	71	71	54	72
NO	10	35	--	95	81	75	72	81
CW1	11	57	88	--	85	75	71	80
CW2	11	50	88	95	88	73	69	83
RW1	10	37	73	85	--	73	63	73
RW2	10	44	66	81	86	74	64	74
P1	29	245	70	90	75	76	65	75
P2	42	389	70	87	77	81	68	77
DP1	10	30	70	70	64	82	72	72
DP2	10	30	68	76	68	87	68	73
CL	19	119	74	85	70	91	87	81
RL	20	112	73	85	75	92	85	82
AQ1	20	--	65	72	68	--	96	75
AQ2	17	--	60	71	64	99	--	74
Average			**73**	**83**	**75**	**71**	**70**	**--**

From deep study and analysis of Tables III and IV, some important remarks can be obtained that can be used for the design of the training and testing texts for speaker recognition systems. Three very important major remarks worth mentioning first are listed below followed by other important remarks.

- When the training and testing texts are of similar type, the recognition rate is highest. For example when training with P1, CL, S1, and CW1 respectively the best rate is when testing with P2, RL, S2, and CW2 respectively.
- When reversing the training and testing the above point is still true, for example when training with P2 the best result is when testing with P1
- When training with P1, P2, CL, RL, S1 or S2 the result will be high if we test with any of these. This may be because they are long lists with some having excellent distribution of all the phonemes (as in the case of P1, P2, CL and RL) or the majority of the phonemes (as in the case of S1 and S2)

From Tables III and IV, the following important observations can be noted:

- When training with the long lists the best recognition rate was for P2, P1, RL, CL, S2, and S1 in descending order. P1 gave a recognition rate that almost 9% and 4% higher than that of S1 and S2 respectively although P1 length is 29 sec while the length of each of S1 and S2 is 32 sec, so time length. is not the factor causing higher performance for P1. We think that the reason is the style to a large degree and the fact that all the phonemes are available in P1 for every speaker to a lesser degree. CL and RL lengths are 19 and 20 sec respectively, which equal 60% and 63% respectively of the length of S1 and S2, nonetheless CL and RL have higher recognition rates. This may be attributed to the type of text in CL and RL, as will be highlighted below, and the fact that they are phonetically balanced.
- The sentences of CL and RL are not usual sentences (they are sentences that are not usually used in normal dialogues but were crafted or designed so that they will include certain phonemes) and are difficult to pronounce, which need thinking by the speaker before uttering them, so pronouncing them need attention. This makes them different than other lists and may have similarity with AQs. This may explain why the result when training with AQs and testing with CL or RL is better than when training with AQs and testing with P1 or P2.
- When training with P1 or P2 the best recognition rate was when testing with the other paragraph, CL, RL, S2, and S1 in descending order. So again, style of CL and RL and being phonetically balanced allowed them to give recognition rate better than S1 and S2 though their lengths are shorter.

- Training with P1 and testing with P2 gave 99.45% RR while the reverse gave 100%. Both training with CL and testing with RL, and training with RL and testing with CL gave 99.45% RR. This high RR may be attributed to the style of paragraph pronunciation and having all phonemes for the paragraph. Similarly for the CL and RL this may be attributed to the style of CL and RL and the fact that they are phonetically balanced. It is very important to point that CL and RL length is 19 and 20 sec respectively, nonetheless they gave almost 100% recognition rate. Actually, the error is due only to one speaker who was not recognized out of 182 speakers.
- Training with P1 or P2 gave very good result when testing with CW1 then CW2, while testing with other lists of words gave medium results. This means that a good design of the testing text has a significant positive effect on recognition rate.
- Testing with AQ gave excellent results when training was with CL and RL while training with S1 or S2 gave lowest results. Though S1 and S2 have longer time than CL and CR. This may be explained by the fact that CL and RL are not usual sentences (they are sentences that were crafted or designed so that they will include certain phonemes) and are difficult to pronounce, which need thinking by the speaker before uttering them, so pronouncing them need attention similar to the attention needed in AQ. Hence, AQ, CL, and RL can be considered of similar type and their result will be similar to AQ1 with AQ2 or vice versa.
- When training with CL and RL the results of P1, P2, AQ1, and DP2 were the highest. While result of P1, P2 and AQ1 can be explained as we did before but the result of DP2 needs investigation to see among the fricatives and pharyngeal, or other phonemes, what are the phonemes that may cause this high rate (95%). DP2 achieved this result though it is missing the majority of the phonemes (17 phonemes). This confirms our design criteria for DP1 and DP2, which is that fricatives and pharyngeal are excellent for speaker recognition as described in Chapter 5 and will be an important point to investigate in the future.
- CWs gave excellent to very good result and better than RWs. This again emphasizes the need to choose the best text for recognition. Another point to look into is that CW1 and CW2 are easy to pronounce since they are common, so does this have an effect in recognition?
- We expected RWs to give result higher than CWs, because RWs were selected in SAVVB, but they did not. One reason may be that the initial lists did not have all the phonemes and we went more and divided it into two lists, RW1 and RW2. If we compare result of RW with CW1, we see that CW1 had better result although all of three lists are missing eight phonemes. This may attributed to the phonemes present in each list, this is a point to be looked in deeply as we mentioned before. Another explanation may be the one we mentioned in point 7 above.

IV. Effect Of Switching The Order Of The Training And Testing Text

By looking at the results when we switch the order of the training and testing text, we found some interesting observations. These observations are listed below and explained later in the section. The symbol '≈' is used to show that two values are approximately equal. The symbol '<' is used to show that the value on the left side of the symbol is less than the right side value, and the symbol '<<' represent than value on the left hand side is much less.

- Train with S1 and S2 and test with P1 and P2 ≈ Train with P1 and P2 and test with S1 and S2.
- Train with CL and RL and test with P1 and P2 ≈ Train with P1 and P2 and test with CL and RL.
- Train with S1 and S2 and test with CL and RL ≈ Train with S1 and S2 and test with CL and RL
- Train with NO and test with P1 and P2 << Train with P1 and P2 and test with NO.
- Train with CW1 and test with P1 and P2 < Train with P1 and P2 and test with CW1.
- Train with RW1 and test with P1 and P2 << Train with P1 and P2 and test with RW1.
- Train with AQ and test with P1 and P2 << Train with P1 and P2 and test with AQ.
- Train with NO and test with CL and RL << Train with CL and RL and test with NO.
- Train with COM1 and test with CL and RL < Train with CL and RL and test with COM1.
- Train with RW1 and test with CL and RL << Train with CL and RL and test with RW1.
- Train with AQ and test with CL and RL < Train with CL and RL and test with AQ.

These observations can be explained as follows:

In the first three points, the training and testing were similar, i.e. long text, and almost have all the phonemes; hence, the results are the same. For points 4 until 11, in the experiments in the left side of the observation, the test will have phonemes not present in the training. Hence, this leads to lower recognition rate than when reversing the order of training and testing. The recognition rate will be lower or much lower depending on the phonemes missing, which is an important point to investigate.

V. Conclusion

This study shows that careful design of the training and testing text can result in high recognition rate. Examples of such points are having training and testing text of same style, having phonetically balanced training text, using artificially constructed sentences, and using words from daily life.

This investigation is done only to shed light into this topic and draw attention to it. This investigations can be

further extended. For example, we could use parts of P1 or P2 to see if the effect of pronouncing a paragraph will continue even if the time is shortened and some phonemes are missed. We also can make S1 and S2 into one list; hence, it will have all the phonemes as designed, and then investigate the performance of the system. We also can look into the best text to construct SR systems that are robust to noise or session variability.

ACKNOWLEDGMENT

This work is supported by the National Plan for Science and Technology in KSU under grant number 13-INF977-02. The author is grateful for this support.

REFERENCES

[1] V. K. Madisetti, The Digital Signal Processing Hangbook, Second ed., CRC Press, 1999.

[2] T. Kinnunen, and H. Li, “An Overview of Text-Independent Speaker Recognition: From Features to Supervectors,” Speech Communication, 2010, vol. 52, no. 1, pp. 12-40.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille et al., “A Tutorial on Text-Independent Speaker Verification,” EURASIP Journal on Advances in Signal Processing, 2004, vol. 2004, no. 4, pp. 430-451.

[4] N. Singh, R. A. Khan, and R. Shree, “Applications of Speaker Recognition,” Procedia Engineering, 2012, vol. 38, no. 0, pp. 3122-3126.

[5] M. Hébert, "Text-Dependent Speaker Recognition," Springer Handbook of Speech Processing, pp. 743-762: Springer Berlin Heidelberg, 2008.

[6] M. A. Anusuya, and S. K. Katti, “Front End Analysis of Speech Recognition: A Review,” International Journal of Speech Technology, 201,1vol. 14, no. 2, pp. 99-145.

[7] D. A. Reynolds, “Large Population Speaker Identification Using Clean and Telephone Speech,” Signal Processing Letters, IEEE, 1995, vol. 2, no. 3, pp. 46-48.

[8] H. Aronowitz, D. Irony, and D. Burstein, "Modeling Intra-Speaker Variability for Speaker Recognition."Proc. INTERSPEECH’05, 2005, pp. 2177-2180.

[9] S. Pandiaraj, D. S. Vinothini, H. N. R. Keziah et al., "Speaker Identification Using pykfec and AANN." Proc. 3rd International Conference on Electronics Computer Technology (ICECT), 2011, pp. 313-316.

[10] A. Zulfiqar, A. Muhammad, A. Martinez-Enriquez et al., "Text-Independent Speaker Identification Using VQ-HMM Model Based Multiple Classifier System," Proc. Advances in Soft Computing, Lecture Notes in Computer Science,Spinger, 2010, pp. 116-125.

[11] A. Zulfiqar, A. Muhammad, and A. M. M. Enriquez, "A Speaker Identification System Using MFCC Features with VQ Technique.", Proc. 3rd International Symposium on Intelligent Information Technology Application, 2009, pp. 115-118.

[12] A. Lawson, P. Vabishchevich, M. Huggins et al., "Survey and Evaluation of Acoustic Features for Speaker Recognition," Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 5444-5447.

[13] H. S. Jayanna, and S. R. Mahadeva Prasanna, “An Experimental Comparison of Modelling Techniques for Speaker Recognition Under Limited Data Condition,” Sadhana, 2009, vol. 34, no. 5, pp. 717-728.

[14] B. Fauve, N. Evans, N. Pearson et al., “Influence of Task Duration In Text-Independent Speaker Verification,” Proc. INTERSPEECH’07, 2007, pp. 794-797.

[15] A. Mahmood, M. M. Alsulaiman, G., and S. M. Selouani, “MDLF-Mavg A New Speech Feature with a Voice Print,” Proc. 7th IEEE GCC Conference and exhibition, 2013, pp. 602-606, 2013.

[16] M. M. Alsulaiman, G. Muhammad, M. A. Bencherif et al., "Building a Rich Arabic Speech Database," Proc. 5th Asia Modelling Symposium (AMS), 2011, pp. 100-105.

[17] M. M. Alsulaiman, G. Muhammad, M. A. Bencherif et al., “KSU Rich Arabic Speech Database,” Information, 2013, vol. 16, no. 6(B), pp. 4231-4253.

[18] D. Reynolds, "Gaussian Mixture Models," Encyclopedia of Biometrics, pp. 659-663, Springer US, 2009.