Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 3763 – 3778
© TÜBİTAK
doi:10.3906/elk-1812-152

Extracting accent information from Urdu speech for forensic speaker recognition

Falak TAHIR[1], **Sajid SALEEM**[1*], **Ayaz AHMAD**[2]
[1]Faculty of Engineering and Computer Sciences, National University of Modern Languages, Islamabad, Pakistan
[2]Department of Electrical Engineering, COMSATS Institute of Information Technology, Wah Cantt, Pakistan

Received: 21.12.2018 • **Accepted/Published Online:** 19.06.2019 • **Final Version:** 18.09.2019

Abstract: This paper presents a new method for extraction of accent information from Urdu speech signals. Accent is used in speaker recognition system especially in forensic cases and plays a vital role in discriminating people of different groups, communities and origins due to their different speaking styles. The proposed method is based on Gaussian mixture model-universal background model (GMM-UBM), mel-frequency cepstral coefficients (MFCC), and a data augmentation (DA) process. The DA process appends features to base MFCC features and improves the accent extraction and forensic speaker recognition performances of GMM-UBM. Experiments are performed on an Urdu forensic speaker corpus. The experimental results show that the proposed method improves the equal error rate and the accuracy of GMM-UBM by 2.5% and 3.7%, respectively.

Key words: Forensic, classification, speaker recognition, speech features

1. Introduction

Pronunciation varieties of a spoken language are known as accent [1, 2]. Accent generally refers to ways of pronouncing a language within a community. Remarkable attempts have been made to automatically extract accent from the speaker's utterances [3, 4]. Accent provides information about geographical and territorial origin of speakers [5, 6]. Other applications of accent are telephone-based assistant systems, telephone banking, voice mail, voice dialling, and e-learning [7, 8].

This paper deals with extraction of accent from Urdu speech signals for forensic speaker recognition. Speaker profiling, prison call monitoring, and biometric authentication are used by law enforcement agencies to identify the geographical and territorial origin of the criminal suspects [9, 10]. Urdu is the national language of Pakistan. Pakistan is a multilingual country, where people understand, communicate, and speak Urdu along with their native languages. Accents of different native/regional languages of Pakistan such as Punjabi, Sindhi, Balochi, and Pashto influence the Urdu accent [11]. Variations in Urdu accent are used by forensic experts for speaker profiling and identification of territorial origin, geographic background, and ethnicity of the suspects.

Urdu, compared to English and other international languages, is an underresourced language. Limited work on Urdu accent extraction and recognition is available [12, 13]. The work that is available is either text- or speaker-dependent. In contrast, this paper deals with speaker- and text-independent accent recognition using a new method and applies the extracted accent information in improving the forensic speaker recognition accuracy. The main contributions of this paper are:

[*]Correspondence: sajidslm@yahoo.com

- A new speech corpus for Urdu accent recognition,
- A new speech corpus for Urdu forensic speaker recognition,
- A comparison of different speech features for Urdu accent recognition,
- A comparison of different classifiers for Urdu accent recognition,
- A new method for extraction of accent information from Urdu speech signals.

The rest of this paper is organized as follows: Section 2 presents an overview of accent and forensic speaker recognition techniques. Section 3 presents the proposed method. Section 4 presents the experimental setup and results. Finally, the paper is concluded in Section 5.

2. Related work

Accent provides information about the speaker's culture and territorial background [14]. Accent along with acoustic features such as pitch, density, and amplitude are used by forensic experts to make suspect identification decisions [15]. The traditional approaches of forensic speaker recognition are based on manually examining the recordings, which is time-consuming and laborious work. The need for easy-to-apply, reliable, and automatic methods for forensic speaker recognition is rapidly growing [16]. Accent recognition, in this regard, plays a vital role to automate the process of speaker profiling and produces additional information to support either defense or prosecution in forensic cases. With accent recognition system, the search space for identifying a suspect can be limited to an ethnic or regional suspects group [5]. Listening to a speaker, when he/she is speaking a foreign language, it is sometimes difficult to make a decision about his/her accent. With an automatic accent recognition system, this process can be made easy and efficient [17].

In forensic cases it is examined whether a questioned voice belongs to a suspect or not [18]. With accent recognition it becomes relatively simple because the speaking style and way of pronunciation vary from person to person and region to region and results in distinct speech features, which can be used for accent recognition [19]. In [20], different speech features are compared in order to identify appropriate and robust ones for forensic accent recognition. The mel-frequency cepstral coefficients (MFCCs) features are widely used [21]. The performance of different classifiers like Bayesian theory, K-nearest neighbour (KNN), neural network etc. are also compared using MFCC features for accent recognition. It is shown that MFCC compared to traditional feature extraction methods like linear predictive coding (LPC) and linear predictive cepstral coefficients (LPCC) are effective and robust. KNN with MFCC is found suitable for accent recognition in [22] whereas Gaussian mixture models (GMM) and support vector machines (SVM) are used with MFCC in [23, 24]. The selection of features and classifiers vary with respect to speech corpuses and applications. Table 1 summarizes features and classifiers that have been used for accent recognition.

A Panjabi-English in Bradford and Leicester corpus is investigated in [20] to estimate accent similarities across different linguistics in UK. The investigation of accent is then used in forensic casework. Similarly, an Urdu speech corpus consisting of 139 different district names of Pakistan is constructed in different regional accents like Punjabi, Pashto, Sindhi, Balochi, Seraiki, and Urdu [11]. The corpus is then used for automatic speaker recognition for which the accent information is first extracted from the speech samples to improve the speaker recognition rates [11].

Table 1. Summary of features and classifiers for accent recognition.

Method	Classifier	Features	Applications
Huang et al. [2]	GMM, independent component analysis (ICA), principal component analysis (PCA), and hidden Markov model (HMM)	MFCC	Accent adaptation and speech recognition
Sinha et al. [3]	Neural network	MFCC and perceptual linear prediction coefficients (PLP)	Accent identification with different acoustic-phonetic features
Lyn [5]	GMM and K-nearest neighbour (KNN)	MFCC and LPC	Gender and accent identification from Malaysian english
Abbas et al. [25]	Linear discriminant analysis (LDA)	MFCC	Pashto isolated digits database development and recognition for Yousufzai dialect
Huang et al. [26]	GMM and hidden Markov model (HMM)	MFCC	Accent and gender recognition

This paper focuses on Urdu accent recognition and its applications in forensic speaker recognition. For this purpose two different Urdu speech corpora are constructed. The first corpus is used for accent recognition only and consists of text- and speaker-independent samples whereas the second corpus consists of speech samples for accent-based forensic speaker recognition.

3. Proposed method

Figure 1 shows a block diagram for the proposed method. The block diagram consists of an Urdu speech corpus, training & test data, data augmentation, GMM-UBM training, classification, and performance measures.

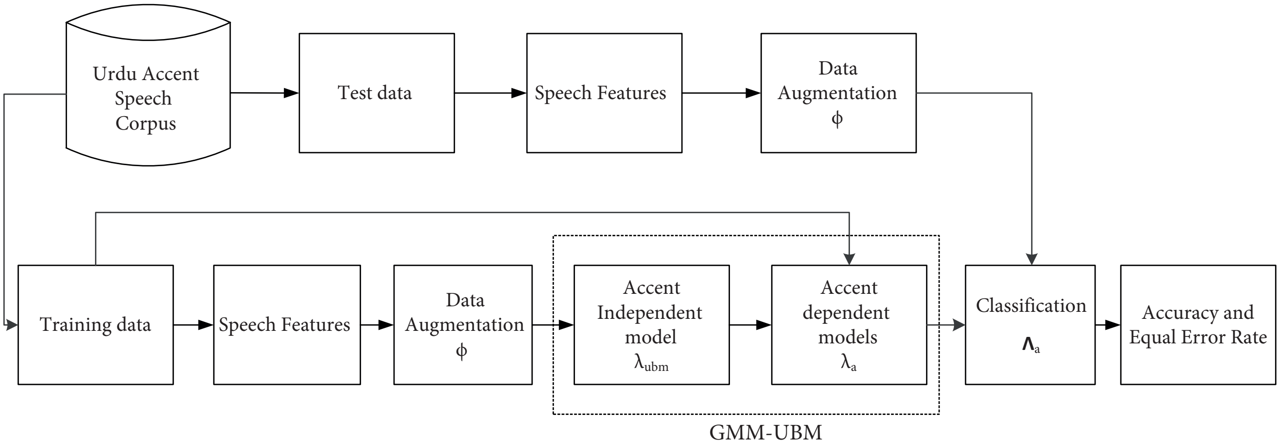

Urdu Accent Speech Corpus
Test data
Speech Features
Data Augmentation ϕ
Training data
Speech Features
Data Augmentation ϕ
Accent Independent model λ_{ubm}
Accent dependent models λ_a
Classification Λ_a
Accuracy and Equal Error Rate
GMM-UBM

Figure 1. Block diagram for the proposed method for recognition of accent from Urdu speech.

3.1. Urdu speech corpus

The Urdu speech corpus consists of speech samples. These samples were collected from different Internet sources. The corpus consists of four different Urdu accents, namely Punjabi, Pashto, Sindhi, and Balochi.

Table 2 summarizes the Urdu speech corpus. The corpus is used for Urdu accent recognition only. Each accent category of the corpus consists of 70 different speech samples. Each sample belongs to a different speaker, which means that there are 70 different speakers per accent category. Each sample is 15-s long, in WAV format, monochannel, and is sampled at 16 kHz. Each sample is also different from all the other samples in the corpus. The reason for this is to implement text- and speaker-independent Urdu accent recognition compared to [11]. Total number of samples in the corpus are 70 speakers × 4 accents = 280 samples.

Table 2. Summary of Urdu Speech corpus for Urdu accent recognition.

Accents	Number of samples	Number of female speakers	Number of male speakers	Duration per sample	Nature of samples
Punjabi	70	30	40	15 seconds	Speaker and text independent
Sindhi	70	32	38	15 seconds	
Pashto	70	35	35	15 seconds	
Balochi	70	31	39	15 seconds	

3.2. Training and test data

Samples of each accent category are randomly divided into two disjoint sets. One set for training and the other one for testing. The training set consists of 50 samples whereas the test set comprises 20 samples.

3.3. Speech features

The MFCC algorithm is used for extraction of speech features from the speech samples. The experimental results of Section 4.3 show that MFCC outperforms other speech features.

3.4. Data augmentation

Data augmentation (DA) is one of the main contributions of this paper. DA appends features to base MFCC features to improve accent and forensic speaker recognition accuracies. It is different from feature mapping used in support vector machines (SVM) [27]. In contrast, this paper uses DA to enhance the Urdu accent and forensic speaker recognition accuracies of GMM-UBM classifier [28]. To understand the proposed DA process, let $X_j = [x_1, x_2, x_3, ..., x_n]$ be an n-dimensional feature vector and $X_j \in \Re^n$. A transformation φ is applied on X_j to obtain a feature vector X_j' as follows:

$$X_j' = \varphi(X_j) = \begin{bmatrix} x_1 & x_2 & x_3 & ..... & x_n \\ x_1^2 & x_2^2 & x_3^2 & ..... & x_n^2 \end{bmatrix} \tag{1}$$

The DA process transforms $1 \times n$ sized feature vector X_j into X_j' i.e a feature matrix of size $2 \times n$, where x_i^2 is obtained by squaring the i^{th} element of X_j. Our experimental results show that such a DA process improves the accuracy of GMM-UBM classifier and outperforms base MFCC features and the features obtained by appending delta (Δ, first derivative of MFCC) and delta-delta (Δ^2, second derivative of MFCC) to base MFCC features in the same fashion as x_i^2. The experimental results also show that it also outperforms linear SVM and SVM based on polynomial and RBF kernels.

In case of sequence of training feature vectors i.e $X = [X_1, X_2, X_3, ..., X_m]$, where the size of each feature vector X_j is $1 \times n$ and the size of X is $m \times n$, where m represents number of training feature vectors. Through the proposed DA process, a new sequence of training vectors X' are obtained i.e $X' = \varphi(X)$. The size of X' is $2m \times n$ because each $1 \times n$ vector is mapped to $2 \times n$ sized feature vectors using (1).

3.5. Accent-independent model (λ_{ubm})

The training speech features after passing through the proposed data augmentation process is provided to GMM-UBM classifier for training purpose. GMM-UBM is trained using different Gaussian mixture components starting from 2 up to 256 components. GMM-UBM provides an accent-independent model known as the background model (λ_{ubm}). To obtain λ_{ubm}, GMM-UBM combines the training speech features of all the accent categories and computes M-different Gaussian mixture components as illustrated in Figure 2a. λ_{ubm} model is parameterized by M Gaussian mixture components having mixture weights as ω_i, mean vectors as μ_i of size $n \times 1$, and $n \times n$ sized covariance matrices Σ_i and $\sum_i^M \omega_i = 1$. The mixture density for feature vector $X_j \in \Re^n$ is then obtained as follows:

$$p(X_j|\lambda_{ubm}) = \sum_i^M \omega_i p_i(X_j) \tag{2}$$

The density is a linear combination of M Gaussian densities:

$$p_i(X_j) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(X_j - \mu_i)'(\Sigma_i)^{-1}(X_j - \mu_i)\right\} \tag{3}$$

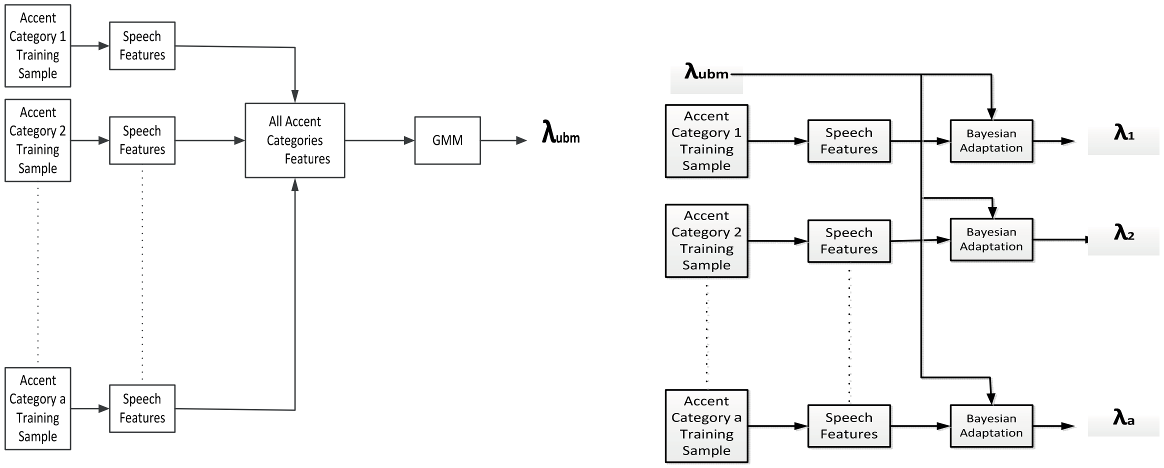
In case of a sequence of feature vectors, where $X = [X_1, X_2, X_3, ..., X_m]^T$ the log likelihood is computed as follows:

$$log\ p(X|\lambda_{ubm}) = \sum_{t=1}^m log\ p(X_t|\lambda_{ubm}) \tag{4}$$

3.6. Accent-dependent model (λ_a)

Accent-dependent models (λ_a) are computed from λ_{ubm} with Bayesian adaptation process [29]. The adaption process adapts the parameters of λ_{ubm} i.e. mean, covariance, and mixture weights for each accent category by using its training speech features as illustrated in Figure 2b. In our case there are four different accent categories i.e. $a = \{1, 2, 3, 4\}$. Let $X_a = [X_{a_1}, X_{a_2}, X_{a_3}, ..., X_{a_k}]^T$ be a set of training feature vectors of a^{th} accent category, and let i be the i^{th} Gaussian mixture of λ_{ubm}, then:

$$Pr(i|X_{a_t}) = \frac{\omega_i p_i(X_{a_t})}{\sum_{j=1}^M \omega_j p_j(X_{a_t})} \tag{5}$$


Accent Category 1 Training Sample
Speech Features
Accent Category 2 Training Sample
Speech Features
All Accent Categories Features
GMM
λubm
Accent Category a Training Sample
Speech Features
λubm
Accent Category 1 Training Sample
Speech Features
Bayesian Adaptation
λ1
Accent Category 2 Training Sample
Speech Features
Bayesian Adaptation
λ2
Accent Category a Training Sample
Speech Features
Bayesian Adaptation
λa

(a)

(b)

Figure 2. Illustration of GMM-UBM method for estimation of (a) accent-independent model λ_{ubm} and (b) accent-dependent model λ_a.

$Pr(i|X_{a_t})$ is used in the computation of sufficient statistics for parameter adaptation as follows:

$$s_i = \sum_{t=1}^{k} Pr(i|X_{a_t}) \tag{6}$$

$$E_i(x) = \frac{1}{s_i}\sum_{t=1}^{k} Pr(i|X_{a_t})X_{a_t} \tag{7}$$

$$E_i(x^2) = \frac{1}{s_i}\sum_{t=1}^{k} Pr(i|X_{a_t})X_{a_t}^2 \tag{8}$$

The sufficient statistics create the adapted parameters for i^{th} mixture of accent model λ_a from the i^{th} mixture of λ_{ubm} as follows:

$$\hat{\omega_i} = [\alpha_i^{\omega} s_i/k + (1-\alpha_i^{\omega})\omega_i]\gamma \tag{9}$$

$$\hat{\mu_i} = \alpha_i^{m} E_i(x) + (1-\alpha_i^{m})\mu_i \tag{10}$$

$$\hat{\sigma_i^2} = \alpha_i^{v} E_i(x^2) + (1-\alpha_i^{v})(\sigma_i^2+\mu_i^2) - \hat{\mu_i^2} \tag{11}$$

where α_i^{ω}, α_i^{m}, α_i^{v} are the adaptation coefficients, γ is a scale vector computed over all the adapted mixture weights to ensure they sum to unity and α_i^{p}, $p \in \{weight(\omega),\ mean(m),\ variance(v)\}$ is defined as:

$$\alpha_i^{p} = \frac{s_i}{s_i + r^{p}} \tag{12}$$

where r^p is a fixed relevance factor for parameter p and $r = 16$ is used.

3.7. Classification

Each accent-dependent model λ_a is parameterized by $\hat{\omega_i}$, $\hat{\mu_i}$, $\hat{\sigma_i^2}$. Test samples are provided to λ_{ubm} and λ_a for accent recognition/classification. Let there be a test sample, which consists of a sequence of test feature vectors $Y = [Y_1, Y_2, Y_3, ..., Y_k]^T$, where $Y_k \in \Re^n$. Then the difference between log-likelihood is computed as:

$$\wedge_a(Y) = log\ p(Y|\lambda_a) - log\ p(Y|\lambda_{ubm}) \tag{13}$$

$$\hat{a} = \arg \max_{a=1}^{4} \wedge_a \tag{14}$$

Accent is predicted for the test sample. The predicted accent ($\hat{a}$) belongs to an accent category of the corpus which maximizes $\wedge_a$. After that, accuracy and equal error rate (EER) are computed.

4. Experimental setup and results

This section presents experimental setup and results. The experimental results are divided into two parts. In the first part a comparison of different features and classifiers is presented for Urdu accent recognition and the performance of the proposed method is compared with state-of-the-art methods. In the second part, experimental results for forensic speaker recognition are presented with and without using the accent information for speaker recognition.

4.1. Performance measures

Two different performance evaluation metrics are used, i.e accuracy and EER. Accuracy is defined as a ratio of correctly classified samples to total number of test samples. For instance, if a classifier correctly classifies 20 samples out of 30 test samples then accuracy would be $(20 \times 100)/30 = 66.7\%$. EER is defined as a point where the false acceptance rate becomes equal to the false rejection rates.

4.2. I-vector

For I-vector, first M Gaussian mixture components are extracted from combined training samples of all the accent categories, as explained in Section 3.5 for accent-independent model (λ_{ubm}). In fact these components represent the universal background model (UBM). Each mixture component c is parameterized by a weight (ω_c), mean (μ_c), and covariance matrix (Σ_c). Let us say that each feature vector is n dimensional. To capture interutterance (training speech samples) variability, an $R \times 1$ vector y (i.e. I-vector) is associated with each utterance. Similarly with each mixture component c an $n \times R$ matrix V_c, is associated. In this paper we use different values for R (I-vector dimension) i.e. 10 to 200 with step size of 10. The dimension that gives the best results for a given number of Gaussian mixture components is presented in the experimental results section.

The feature vectors of each utterance associated with the mixture components are supposed to be distributed with mean m_c and Σ_c as:

$$m_c = \mu_c + V_c y \tag{15}$$

The likelihood for the utterance having X_t as a sequence of feature vectors is computed as

$$\sum_c \left(N_c \ln \frac{1}{(2\pi)^{n/2}|\Sigma_c|^{1/2}} - \frac{1}{2}\sum_t (X_t - V_c y - m_c)^* \Sigma_c^{-1} (X_t - V_c y - m_c) \right) \tag{16}$$

where $*$ is conjugate transpose, the sum over c is over all mixture components, the sum over t is over all features aligned with c, N_c, and F_c are defined as follows using the Baum-Welch statistics:

$$N_c = \sum_t \gamma_t(c) \tag{17}$$

$$F_c = \sum_t \gamma_t(c) X_t \tag{18}$$

where $\gamma_t(c)$ is the posterior probability that X_t is generated by the mixture component c and it is calculated using the UBM (i.e. accent-independent model). In fact, in I-vector method, the parameters of each mixture component i.e. μ_c and Σ_c are copied from the UBM and only matrix V_c is estimated from the speech samples of the training set. Moreover, the posterior distribution of y for a given utterance is calculated with an assumption that prior is standard normal. To estimate V_c, the first and second order moments $\langle y(s) \rangle$ and $\langle y(s)y^*(s) \rangle$ are computed for each training utterance (s) and the maximum likelihood update formula for V_c is obtained as:

$$V_c = \left(\sum_s F_c(s) \langle y^*(s) \rangle \right) \left(\sum_s N_c(s) \langle y(s) y^*(s) \rangle \right)^{-1} \tag{19}$$

where the sum over s is over all utterances in the training set. For each utterance s, $N_c(s)$, and $F_c(s)$ are the Baum–Welch statistics of order 0 and 1 for mixture component c. Let LL^* be Cholesky decomposition of the following matrix:

$$\frac{1}{S} \sum_s \langle y(s) y^*(s) \rangle \tag{20}$$

where S is the number of training utterances and the sum is over all the utterances of the training set. Then the following transformations are applied and I-vector for each utterance is obtained:

$$V_c \leftarrow V_c L \tag{21}$$

$$y(s) \leftarrow L^{-1} y(s) \tag{22}$$

We use Gaussian probabilistic linear discriminant analysis as described in [30] for scoring of I-vectors. Such a scoring is based on calculating the batch likelihood ratio as described below:

$$\ln \frac{P(y_{target}, y_{test} | H_1)}{P(y_{target} | H_0) P(y_{test} | H_0)} \tag{23}$$

where y_{target} and y_{test} are two I-vectors that belong to training and test sets, respectively, H_1 means that accent are same and H_0 means accent are different. The accent category that maximizes (23) is identified and assigned as a predicted accent for the test utterance, y_{test}.

4.3. Accent recognition results

In this section, a comparison of different features and classifiers is presented for Urdu accent recognition. Two different corpora are used: (i) Urdu speech corpus (as explained in Section 3.1) (ii) Kaggle accent corpus[1]. The Kaggle corpus is a text-dependent corpus. It contains recording of an English paragraph by different speakers of

[1] https://www.kaggle.com/rtatman/speech-accent-archive/home

177 different countries in their native accents. We choose five different accent categories from the Kaggle corpus which are Arabic, French, Mandarin, Spanish, and English. The selection of these categories are based on the number of samples, which are more than 70 for each category, which we randomly divide into two disjoint sets: one for training and the other one for testing purposes.

Table 3 shows a comparison between MFCC, shifted delta cepstra (SDC), LPCC, and LPC features for accent recognition. EER(%) is used as a metric for performance comparison. GMM-UBM is used as a classifier. The objective of this comparison is to identify the best features for accent recognition. The GMM-UBM is trained using different Gaussian mixture components i.e. 2, 4, 8, 16, 32, 64, 128, and 256. The comparison shows that MFCC, compared to SDC, LPCC, and LPC, demonstrates better EER on the Urdu corpus. MFCC achieves minimum equal rate of 9.7% with 256 components whereas SDC, LPCC, and LPC demonstrate minimum EER rate of 12.3%, 19.4%, and 26.3%, respectively, with 256 mixture components. Similarly, MFCC outperforms SDC and other features on Kaggle corpus by achieving minimum EER of 29% with 64 components.

Table 3. EER (%)-based comparison between MFCC, SDC, LPCC, and LPC features for accent recognition using GMM-UBM classifier with different Gaussian mixture components.

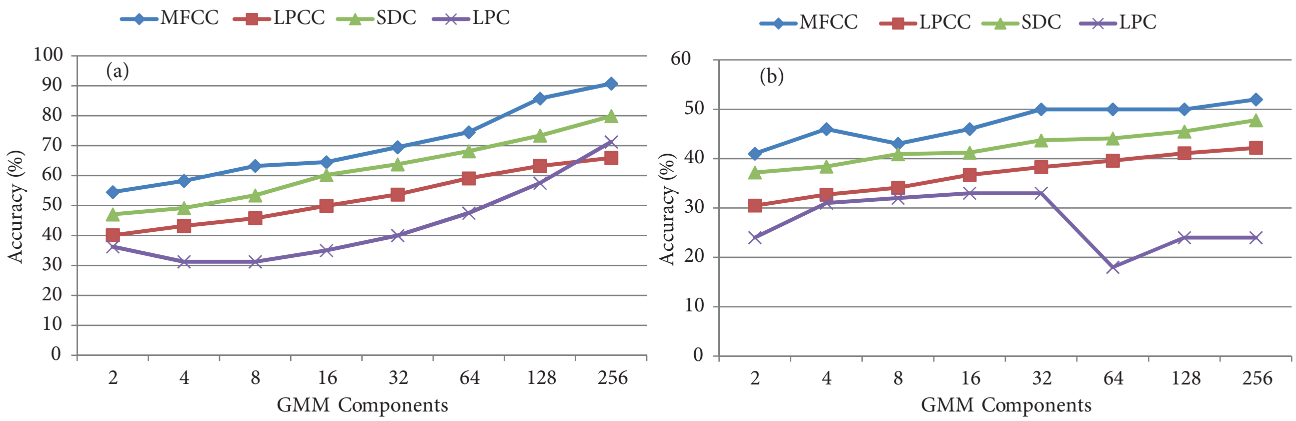
Gaussian components		2	4	8	16	32	64	128	256
Urdu corpus	MFCC	31.8	30.5	29.3	26.8	21.8	18.0	13.0	9.7
	SDC	33.6	32.8	31.2	29.6	25.7	21.4	17.9	12.3
	LPCC	38.2	36.7	34.5	31.6	29.8	26.9	22.1	19.4
	LPC	45.4	45.4	46.3	44.6	40.0	35.0	32.1	26.3
Kaggle corpus	MFCC	34.0	34.0	34.0	31.0	30.0	29.0	29.3	31.0
	SDC	38.2	37.5	36.1	35.3	32.9	31.5	31.2	31.5
	LPCC	43.2	42.2	41.9	39.1	38.6	36.2	34.9	33.9
	LPC	49.0	49.5	48.3	47.0	47.3	45.0	44.0	42.2

Figure 3a shows an accuracy (%)-based comparison between MFCC, SDC, LPCC, and LPC on the Urdu corpus. GMM-UBM is used as a classifier. The results show that accuracy varies with respect to Gaussian mixture components and MFCC outperforms SDC, LPCC, and LPC. Similarly, Figure 3b shows a comparison between features, where MFCC outperforms other features on the Kaggle corpus.

Table 4 shows a comparison between GMM-UBM and I-vector methods for accent recognition. Both are trained using MFCC features and different Gaussian mixture components. The comparison shows that GMM-UBM achieves minimum EER of 9.7% on the Urdu corpus with 256 components whereas I-vector method demonstrates minimum EER of 13.7% with 256 components. Similarly GMM-UBM outperforms I-vector on the Kaggle corpus.

Table 4. EER (%)-based comparison between GMM-UBM and I-vector classifiers for accent recognition using MFCC features and different Gaussian mixture components.

Gaussian components		2	4	8	16	32	64	128	256
Urdu corpus	GMM-UBM	31.8	30.5	29.3	26.8	21.8	18.0	13.0	9.7
	I-vector	52.5	37.9	35.2	31.5	27.6	23.1	17.2	13.7
Kaggle corpus	GMM-UBM	34.0	34.0	34.0	31.0	30.0	29.0	29.3	31.0
	I-vector	38.0	32.5	36.0	36.0	32.3	32.0	34.0	33.5



MFCC and LPCC comparison on Urdu Corpus using GMM-UBM

MFCC and LPCC comparison on Kaggle Corpus using GMM-UBM

Figure 3. Accuracy (%)-based comparison between MFCC, SDC, LPCC, and LPC features for accent recognition using GMM-UBM on the (a) Urdu and (b) Kaggle corpora

Figure 4a shows an accuracy-based comparison between GMM-UBM and I-vector methods on the Urdu corpus. MFCC features are used. The experimental results show that accuracy varies with respect to Gaussian mixture components, and GMM-UBM outperforms I-vector. Similarly, the accuracy-based comparison, as shown in Figure 4b, shows that GMM-UBM also outperforms I-vector on the Kaggle corpus. Based on the above experimental results, it can be said that GMM-UBM and MFCC outperform all other classifiers and features, respectively. Therefore, in the proposed method we use MFCC and GMM-UBM.

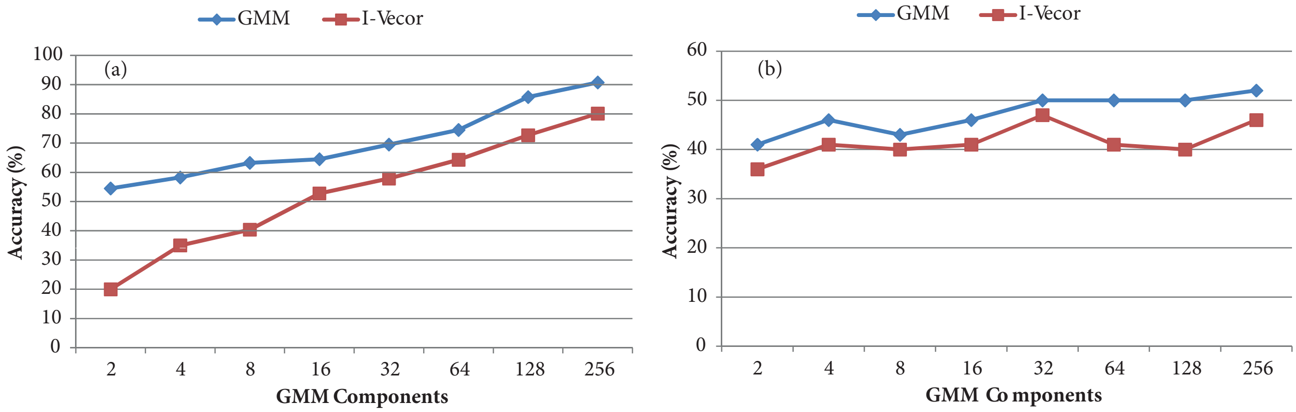


Figure 4. Accuracy (%)-based comparison between GMM-UBM and I-vector for accent recognition using MFCC features on the (a) Urdu and (b) Kaggle corpora.

Table 5 shows an EER (%)-based comparison between GMM-UBM and the proposed method. MFCC features are used. Difference between GMM-UBM and the proposed method is the data augmentation step which is only used in the proposed method. It can be seen that GMM-UBM on the Urdu corpus with 256 components gives EER of 9.7% whereas the proposed method gives EER of 8.4%. Thus, the proposed method improves EER by almost 1.3%. On the Kaggle corpus, GMM-UBM and the proposed method achieve 26% and 31% EER, respectively. Improvement is almost 5%. This shows that data augmentation used in the proposed method efficiently improves the performance of GMM-UBM classifier for accent recognition.

Table 6 shows an accuracy-based comparison between GMM-UBM and the proposed method. MFCC

Table 5. EER (%)-based comparison between GMM-UBM and the proposed method for accent recognition using MFCC features and different Gaussian mixture components.

Gaussian components		2	4	8	16	32	64	128	256
Urdu corpus	GMM-UBM	31.8	30.5	29.3	26.8	21.8	18.0	13.0	9.7
	Proposed	38.8	31.3	25.1	22.6	19.7	15.1	11.3	8.4
Kaggle corpus	GMM-UBM	34.0	34.0	34.0	31.0	30.0	29.0	29.3	31.0
	Proposed	37.8	31.5	30.0	29.0	27.5	26.0	27.0	26.0

features are used. It can be seen that the proposed method demonstrate 1.2% and 3% better accuracy rates than GMM-UBM on Urdu and Kaggle corpora, respectively.

Table 6. Accuracy (%)-based comparison between GMM-UBM and the proposed method for accent recognition using MFCC features and different Gaussian mixture components.

Gaussian components		2	4	8	16	32	64	128	256
Urdu corpus	GMM-UBM	54.5	58.3	63.3	64.5	69.5	74.5	85.8	90.8
	Proposed	57.0	58.3	64.5	70.8	74.5	83.3	89.5	92.0
Kaggle corpus	GMM-UBM	41.0	46.0	43.0	46.0	50.0	50.0	50.0	50.0
	Proposed	32.0	42.0	42.0	45.0	46.0	51.0	53.0	53.0

Tables 7 and 8 show a comparison of the proposed method and the methods based on appending the first (Δ) and second derivatives (Δ^2) of MFCC to base MFCC features. The first and second derivatives are appended in the same fashion as we appended the data (x_i^2) in the proposed method. The results show that the proposed method outperforms augmentation of Δ and Δ^2 features.

Table 7. EER (%)-based comparison between the proposed method and appending Δ and Δ^2 features to base MFCC features for accent recognition with GMM-UBM and different Gaussian mixture components.

Gaussian components		2	4	8	16	32	64	128	256
Urdu corpus	Δ + MFCC	35.5	32.2	27.7	25.3	20.1	16.4	11.9	8.7
	Δ^2 + MFCC	35.9	33.7	28.2	25.9	22.6	17.6	12.8	9.2
	Proposed	38.8	31.3	25.1	22.6	19.7	15.1	11.3	8.4
Kaggle corpus	Δ + MFCC	35.2	33.2	32.9	30.1	28.1	27.0	26.8	26.7
	Δ^2 + MFCC	36.6	34.5	33.7	31.2	28.9	28.3	27.9	27.2
	Proposed	37.8	31.5	30.0	29.0	27.5	26.0	27.0	26.0

Table 9 summarizes the accuracy rates achieved on both corpora using GMM-UBM, I-vector, linear SVM, SVM-RBF, SVM-polynomial, and the proposed method. MFCC features are used. On the Urdu speech corpus, the proposed method demonstrates the best accuracy rate of 92% followed by GMM-UBM (90.8%), I-vector (80.1%), SVM-RBF (61.25%), and linear SVM (55%). The SVM with polynomial kernels with degrees 2 and 3 do not perform well. It can be seen that the accuracy of SVM decreases with increase in the polynomial degree. Similarly on the Kaggle corpus, the proposed method achieves an accuracy of 53% and outperforms all other classifiers. The accuracy achieved on the Kaggle is low compared to the Urdu corpus because samples of

Table 8. Accuracy (%)-based comparison between the proposed method and appending Δ and Δ^2 features to base MFCC features for accent recognition with GMM-UBM and different Gaussian mixture components.

Gaussian components		2	4	8	16	32	64	128	256
Urdu corpus	Δ + MFCC	55.3	58.8	63.9	66.5	73.8	80.9	88.2	91.6
	Δ^2 + MFCC	53.1	57.2	63.4	64.9	72.5	78.6	87.9	91.2
	Proposed	57.0	58.3	64.5	70.8	74.5	83.3	89.5	92.0
Kaggle corpus	Δ + MFCC	36.4	43.5	43.8	44.3	49.4	49.9	51.5	52.6
	Δ^2 + MFCC	35.1	42.6	43.2	44.0	48.9	49.5	51.0	51.5
	Proposed	32.0	42.0	42.0	45.0	46.0	51.0	53.0	53.0

the Kaggle corpus are text-dependent and different speakers record the same English paragraph in their native accents.

Table 9. Accuracy (%)-based comparison between different classifiers for accent recognition using MFCC features.

	GMM-UBM	I-vector	Linear SVM	SVM-Poly Degree-2	SVM-Poly Degree-3	SVM-RBF	Proposed
Urdu corpus	90.8	80.1	55	31.25	30.2	61.25	92
Kaggle corpus	50	47	42	27	20	44.3	53

4.4. Forensic speaker recognition

This section presents experimental results for forensic speaker recognition. Accent classification (AC) prior to speaker recognition is investigated in this section. The experimental results for forensic speaker recognition with and without AC are presented. The experiments are performed on an Urdu forensic speech corpus, which is summarized in Table 10.

Table 10. Summary of Urdu forensic speaker recognition. Corpus.

Accent categories	Number of speakers per category	Number of samples per speaker	Training samples per speaker	Test samples per speaker	Nature of speech samples
Balochi	4	60	40	20	Text-independent
Punjabi	4	60	40	20	
Pashto	4	60	40	20	
Sindhi	4	60	40	20	

The Urdu forensic speech corpus consists of four different accent categories i.e. Balochi, Punjabi, Pashto, and Sindhi. Each accent category consists of four different speakers. The speech samples of each speaker are text-independent and are 60 in number. Each sample is 15-s long, in WAV format, monochannel, and is sampled at 16 kHz. We randomly divide the samples of each speaker into two disjoint sets. One set for training (40 samples) and other one for testing (20 samples). Figure 5 shows the block diagram used for GMM-UBM–based forensic speaker recognition without AC. The same block diagram is also used for the proposed method. the only difference is the data augmentation step which is used only in the proposed method.

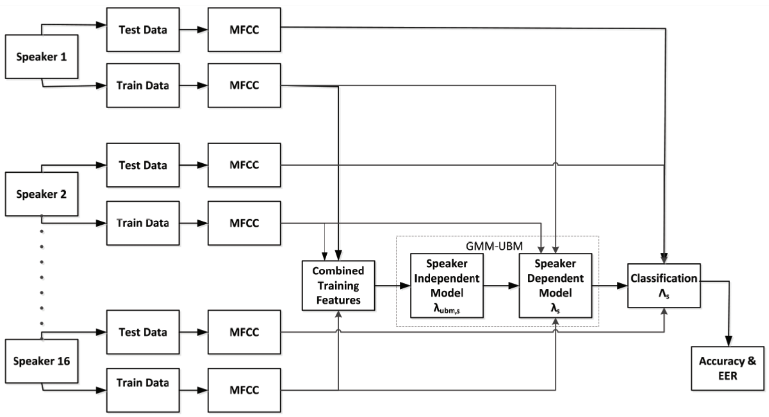

Speaker 1
Test Data
MFCC
Train Data
MFCC
Speaker 2
Test Data
MFCC
Train Data
MFCC
Speaker 16
Test Data
MFCC
Train Data
MFCC
Combined Training Features
GMM-UBM
Speaker Independent Model $\lambda_{ubm,s}$
Speaker Dependent Model λ_s
Classification Λ_s
Accuracy & EER

Figure 5. Block diagram for GMM-UBM–based Urdu forensic speaker recognition system without using accent classification as a preprocessing step.

For Urdu forensic speaker recognition, first the background model (i.e. speaker-independent model) is obtained ($\lambda_{ubm,s}$) for GMM-UBM and the proposed method by combining the MFCC features of training samples of all the speaker categories of the corpus. Then the speaker-dependent model (λ_s), one for each speaker category is adapted from ($\lambda_{ubm,s}$) using the Bayesian adaptation. Since there are four speakers per accent, the total number of the adapted models are 4 × 4 (accents) = 16. Having computed $\lambda_{ubm,s}$ and λ_s, the speaker in a test sample is recognized. Let Y represent a set of MFCC feature vectors obtained from the test speech sample. The difference between log-likelihood for Y is computed as:

$$\wedge_s = log\ p(Y|\lambda_s) - log\ p(Y|\lambda_{ubm,s}) \tag{24}$$

$$\hat{s} = \arg \max_{s=1}^{16} \wedge_s \tag{25}$$

The predicted speaker ($\hat{s}$) for the test sample belongs to that speaker category of the corpus which maximizes $\wedge_s$. After that, accuracy and EER are computed.

For forensic speaker recognition with AC, the accent information form the test speech sample is first extracted. For this purpose the background model (λ_{ubm}) and the accent-independent models (λ_a) computed in Sections 3.5 and 3.6 are used. Then, for a given test sample, first accent is recognized using (14) and then the test sample is processed for speaker recognition within the speakers of the predicted accent category as:

$$\wedge_{s,\hat{a}}(Y) = log\ p(Y|\lambda_{s,\hat{a}}) - log\ p(Y|\lambda_{ubm,\hat{a}}) \tag{26}$$

$$\hat{s_a} = \arg \max_{s=1}^{4} \wedge_{s,\hat{a}} \tag{27}$$

where $\lambda_{ubm,a}$ and $\lambda_{s,a}$ are the speaker-dependent and -independent models of a^{th} accent category, respectively and $a = \{1,2,3,4\}$. The predicted speaker ($\hat{s_a}$) for the test sample belongs to ($\hat{a}$) accent category and maximizes $\wedge_{s,\hat{a}}$. To compute $\lambda_{ubm,a}$, the MFCC features of training samples of all the speakers of the a^{th} accent category are combined. The combined features belong to different speakers, but the accent of all the speakers are the same. For instance, the Balochi accent category consists of four different speakers. All the speakers have the same Balochi accent. Thus, $\lambda_{ubm,a}$ is speaker-independent model within accent category a.

The speaker-dependent models $\lambda_{s,a}$ one for each speaker of the accent category are then adapted from $\lambda_{ubm,a}$ using the Bayesian adaptation process. To recognize a speaker from a test sample, firstly, the accent is identified using (14) and then the speaker is recognized using (27).

Table 11 shows EER rate achieved with and without AC for speaker recognition using different Gaussian mixture components. GMM-UBM and the proposed method with 256 mixtures components achieve 11.4% and 10.4% EER without AC, whereas they achieve 9.6% and 7.1% EER with AC, respectively. Thus, using AC as a preprocessing step improves the EER by almost 1.8% and 3.3% for GMM-UBM and the proposed method, respectively. Similarly, the accuracy rates given in Table 12 shows that speaker recognition rates obtained with AC are better than those without AC. However, the proposed method in both cases outperforms GMM-UBM by achieving better accuracy rates.

Table 11. EER (%)-based comparison between GMM-UBM, I-vector, and the proposed method for forensic speaker recognition

Gaussian components		2	4	8	16	32	64	128	256
Without AC	GMM-UBM	28.8	21.3	18.1	16.8	15.0	15.2	12.5	11.4
	I-vector	30.2	24.7	19.8	17.2	16.6	16.0	13.9	12.7
	Proposed	20.3	18.8	16.3	15.0	13.8	12.5	11.8	10.4
With AC	GMM-UBM	25.0	20.0	17.1	14.6	12.9	10.4	10.0	9.6
	I-vector	27.8	23.4	19.4	16.8	14.2	12.6	11.3	10.1
	Proposed	20.0	12.9	11.3	9.2	8.8	7.1	6.7	7.1

Table 12. Accuracy (%)-based comparison between GMM-UBM, I-vector, and the proposed method for forensic speaker recognition.

Gaussian components		2	4	8	16	32	64	128	256
Without AC	GMM-UBM	42.5	50.0	56.3	68.8	68.8	71.3	71.3	68.8
	I-vector	38.3	49.2	53.5	60.1	65.4	66.9	68.4	69.3
	Proposed	53.8	62.5	66.3	66.3	70.0	68.8	72.5	70.0
With AC	GMM-UBM	61.3	71.3	73.8	82.5	83.8	84.0	82.5	83.8
	I-vector	58.9	70.1	72.8	80.2	82.7	83.1	83.9	84.6
	Proposed	77.5	78.8	86.3	85.0	85.0	86.3	87.5	87.5

5. Conclusion

This paper presents a new method for extraction of accent information from Urdu speech signals. Accent information is used in forensic speaker recognition. The experimental results show that MFCC features, compared to SDC, LPCC, and LPC features, demonstrate better Urdu accent recognition performances. The GMM-UBM classifier compared to I-vector method achieves better Urdu accent recognition results. The proposed method, which is based on GMM-UBM, MFCC, and a data augmentation process, improves the GMM-UBM performance by 1.3% (EER) and 1.2% (Accuracy). Compared to RBF-SVM, Linear-SVM, and Polynomial-SVM, it achieves 30.7%, 37%, and 60% better accuracy rates, respectively. The experimental results for forensic speaker recognition show that GMM-UBM and the proposed method with accent classification give better forensic speaker recognition rates compared to those without using the accent classification as a preprocessing step. However, the proposed method demonstrates 2.5% and 3.7% better EER and accuracy rates compared to GMM-UBM, in forensic speaker recognition experiments, respectively.

References

[1] Bartkova K, Jouvet D. On using units trained on foreign data for improved multiple accent speech recognition. Speech Communication 2007; 49 (10): 836-846. doi: 10.1016/j.specom.2006.12.009

[2] Huang C, Chen T, Chang E. Accent issues in large vocabulary continuous speech recognition. International Journal of Speech Technology 2004; 7 (2): 141-153. doi: 10.1023/B:IJST.0000017014.52972.1d

[3] Sinha S, Jain A, Agrawal SS. Acoustic-phonetic feature based dialect in hindi speech. International Journal on Smart Sensing & Intelligent Systems 2015; 8 (1): 235-254. doi: 10.21307/ijssis-2017-757

[4] Tjalve M, Huckvale M. Pronunciation variation modelling using accent features. In: 2005 European Conference on Speech Communication and Technology; Lisbon, Portugal; 2005. pp. 1341-1344.

[5] Lyn GE. Gender and accent identification for Malaysian English using MFCC and Gaussian mixture model. MSc, Universiti Teknologi Malaysia, Johor, Malaysia, 2013.

[6] Benzeghiba M, Demori R, Deroo O, Dupont S, Erbes T et al. Automatic speech recognition and speech variability: a review. Speech Communication 2007; 49 (11): 763-786. doi: 10.1016/j.specom.2007.02.006

[7] Behravan H. Dialect and accent recognition. MSc, University of Eastern Finland, Joensuu, Finland, 2012.

[8] Kumar VR, Vydana HK, Vuppala AK. Significance of GMM-UBM–based modelling for Indian language identification. Procedia Computer Science 2015; 54: 231-236. doi: 10.1016/j.procs.2015.06.027

[9] Alavijeh AHP. Speaker profiling for forensic applications. MSc, Katholieke Universiteit Leuven, Leuven, Belgium, 2014.

[10] Algabri M, Mathkour H, Bencherif MA, Alsulaiman M, Mekhtiche MA. Automatic speaker recognition for mobile forensic applications. Mobile Information Systems 2017; 1-6. doi: 10.1155/2017/6986391

[11] Rafaqat A, Irtza S, Farooq M, Hussain S. Accent classification among Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu language. In: 2014 The Conference on Language and Technology; Lahore, Pakistan; 2014. pp. 1-7.

[12] Rauf S, Hameed A, Habib T, Hussain S. District names speech corpus for Pakistani languages. In: IEEE 2015 International Conference on Asian Spoken Language Research and Evaluation; Shanghai, China; 2015. pp. 207-211.

[13] Qasim M, Nawaz S, Hussain S, Habib T. Urdu speech recognition system for district names of Pakistan: development, challenges and solutions. In: IEEE 2016 Conference on Coordination and Standardization of Speech Databases and Assessment Techniques; Bali, Indonesia; 2016. pp. 28-32.

[14] Lazaridis A, Khoury E. Swiss French regional accent identification. In: 2014 The Speaker and Language Recognition Workshop; Joensuu, Finland; 2014. pp. 106-111

[15] Ilina O, Koval S, Khitrov M. Phonetic analysis in forensic speaker identification: an example of routine expert actions. In: 1999 Congress of Phonetic sciences; San Francisco, USA; 1999. pp. 157-160.

[16] Brown G. Exploring forensic accent recognition using the y-accdist system. In: 2016 Annual Conference of the International Speech Communication Association; Dresden, Germany; 2016. pp. 305-308.

[17] Maher RC. Audio forensic examination. IEEE Signal Processing Magazine 2009; 26 (2): 84-94. doi: 10.1109/MSP.2008.931080

[18] Drygajlo A. Automatic speaker recognition for forensic case assessment and interpretation. In: Neustein A (editor). Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism. New York, NY, USA: Springer 2012. pp. 21-39.

[19] Stefanus I, Sarwono RJ, Mandasari MI. GMM-based automatic speaker verification system development for forensics in Bahasa Indonesia. In: IEEE 2017 International Conference on Instrumentation, Control, and Automation; Yogyakarta, Indonesia; 2017. pp. 56-61.

[20] Brown G, Wormald J. Automatic sociophonetics: exploring corpora with a forensic accent recognition system. The Journal of the Acoustical Society of America 2017; 142 (1): 422-433. doi: 10.1121/1.4991330

[21] Ma Z, Fokoué E. A comparison of classifiers in performing speaker accent recognition using mfcc features, Open Journal of Statistics 2014; 4: 258-266. doi: 10.4236/ojs.2014.44025

[22] Bhatia M, Singh N, Singh A. Speaker accent recognition by MFCC using k-nearest neighbour algorithm: a different approach. International Journal of Advanced Research in Computer and Communication Engineering 2015; 4 (1): 153-155. doi: 10.17148/IJARCCE.2015.4131

[23] Maesa A, Garzia F, Scarpiniti M, Cusani R. Text-independent automatic speaker recognition system using mel-frequency cepstrum coefficient and gaussian mixture models. Journal of Information Security 2012; 3 (4): 335-340. doi: 10.4236/jis.2012.34041.

[24] Hanani A, Russell M, Carey MJ. Speech-based identification of social groups in a single accent of british english by humans and computers. In: IEEE 2011 International Conference on Acoustics, Speech and Signal Processing; Prague, Czech Republic; 2011. pp. 4876-4879.

[25] Abbas AW, Ahmad N, Ali H. Pashto spoken digits database for the automatic speech recognition research; In: IEEE 2012 International Conference on Automation and Computing; Loughborough, UK; 2012. pp. 1-5.

[26] Huang R, Hansen JHL, Angkititrakul P. Dialect/accent classification using unrestricted audio. IEEE Transactions on Audio, Speech, and Language Processing 2007; 15 (2): 453-464. doi: 10.1109/TASL.2006.881695.

[27] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20 (3): 273-297. doi: 10.1007/BF00994018.

[28] Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 2000; 10 (1): 19-41. doi: 10.1006/dspr.1999.0361.

[29] Gauvain JL, Lee CH. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing 1994; 2 (2): 291-298. doi: 10.1109/89.279278.

[30] Kenny P. Bayesian speaker verification with heavy-tailed priors. In: 2010 The Speaker and Language Recognition; Brno, Czech Republic; 2010. pp. 1-10.