Speaker Recognition Using Mel Frequency Cepstral Coefficient and Locality Sensitive Hashing

Ahmed Awais
School of Information and Software Engineering
University of Electronic Science and Technology of
Chengdu China
e-mail: engr.awais86@yahoo.com

She Kun
School of Information and Software Engineering
University of Electronic Science and Technology
Chengdu China
e-mail: kunshe@126.com

Yue Yu
School of Information and Software Engineering
University of Electronic Science and Technology of
Chengdu China

Shaukat Hayat
School of Information and Software Engineering
University of Electronic Science and Technology of
Chengdu China
e-mail: hayat.uestc@yahoo.com

Aftab Ahmed
School of Information and Software Engineering
University of Electronic Science and Technology of
Chengdu China
e-mail: aftabahmed@ibacc.edu.pk

Tianyi Tu
School of Information and Software Engineering
University of Electronic Science and Technology
Chengdu China
e-mail: 578716887@qq.com

***Abstract*—The Mel-Frequency Cepstral Coefficients (MFCC) feature can be cast-off in speaker recognition. The process of feature extraction of the speech signal using Mel-Frequency Cepstral Coefficients (MFCC) feature vectors will generate an acoustic speech signal. Locality Sensitive Hashing (LSH) is frequently used as a classifier for Big Data related problems. In this research, we proposed a new model based on MFCC and LSH to integrate into speaker recognition model. The main returns of our newly proposed model are to get robustness, effective and accurate results in comparison with MFCC+GMM, LPCC+GMM and MFCC+PNN models. This model also contributes to the literature of Big Data. In this model, first, we extract the MFCC features from the wave file then we applied LSH classifier on extracted feature to transform into hash-table. Finally, the hash-tables of train and test wave files are compared and obtained 92.66% speaker recognition accuracy. We compared the accuracy ratio of proposed model with other traditional models namely MFCC+GMM, MFCC+PNN, and LPCC+GMM. Experimental results show that proposed model is more accurate and robust than traditional models and good for speaker recognition.**

Keywords-big data; speaker recognition; MFCC; locality sensitive hashing (LSH)

I. INTRODUCTION

Speaker Recognition is a technology which introduces all the recognizing of human from their voiceprint. Speaker recognition is mostly used in security requirements on behalf of authentication systems, framework and accordingly proceeding. The individual supplies would be in light of verification services, the execution of usual.

Speaker recognition model is not particularly excessive for those speech communicated. Speaker recognition model can be separated under two important section i.e. feature extraction and secondly, speaker training in assessment of those determined speech features.

In a process of feature extraction, Mel Frequency Cepstral Coefficient (MFCC) [1] and Linear Predictive Cepstral Coefficients (LPCC) [2] speech features are frequently determined. Consecutively, Gaussian mixture model (GMM) [3] is commonly used as a classifier for speaker recognition. (MFCC) is a well-known speech characteristic. To its abstraction algorithm, Speech is declined under 13 diverse sub-bands developed by Mel filtering following Discrete Fourier Transform (DFT), as well as at that time 13 Cepstral coefficients are extracted out from the sub-bands to prepare a feature vector. Mel filtering can simulate the human sound-related features it can improve individual speech understandability. Although, (DFT) cannot efficiently examine the non-stationary besides non-linear speech signal [6]. Speech is to break down under 9 separate sub-bands through discrete wavelet transform (DWT), then to transform into energy transmission having 9 energy probabilities, is to derive for making feature vector extracted through sub-bands. In order to solve the problem (DFT), DWT can examine the speech successfully. Additionally, it can react like Mel filtering to consider a human sound-related feature. Nevertheless, the sudden changes occur in

978-1-5386-6987-7/18/$31.00 ©2018 IEEE

human sound-related features, cannot be recognized by energy distribution. It represents one of the main descriptive forms of the signal.

The enhancement of this speech feature is not just proceeded improvement from claiming DWT with considering the speech efficiently; while, further on, speaker arrangement remains different significant extent on behalf of speaker recognition model. Probabilistic neural network (PNN) will be sort feed-forward neural network, which recognizes mutually associated as well as challenging facts in the training process are traditional classifiers. Consistently it simply allows challenging for information in alongside its training procedure. However, the associated information is correspondingly significant. Locality Sensitive Hashing (LSH) controls individual's dimensionality almost high-dimensional information. (LSH) support information properties regular the technique that comparative things guide of the similar "buckets" on behalf of high probability. This identifies individually associated and challenging in the process of training data. Along with improved performance constraint better-quality performance over (GMM). So enhanced the performance (GMM) and (PNN) in different circumstances. In this speaker recognition model, we employ (LSH) such as a classifier. Toward calculation, the maximum circulation of this paper lies formerly utilizing (MFCC) features. When we integration the (LSH) as a classifier and propose to form an innovative speaker recognition model. It has the ability to get appreciated performance on the State that the excellence for speech.

This paper exists structured such as. In Segment II, we define the traditional speaker identification models.

II. Traditional Speaker Recognition Models

Now in this segment, we define three traditional speaker recognition models for example (MFCC+GMM) (MFCC+PNN) as well as (LPCC+GMM). We define Linear Prediction Cepstral Coefficients (LPCC) in Segment A. then we explained Probabilistic Neural Network in section B. Lastly, Gaussian mixture model (GMM) classifier is described in Segment C.

A. Linear Prediction Cepstral Coefficient (LPCC)

Linear Prediction Cepstral Coefficients (LPCC) has been regularly utilized in some speaker recognition proposed models. Assume that (LPCC) exists on standard individual mankind's spoken region. (LPCC) is useful for identifying passionate constituent of speech. They must assist released from initial linear prediction coding (LPC) coefficient utilizing recursive algorithm [4].

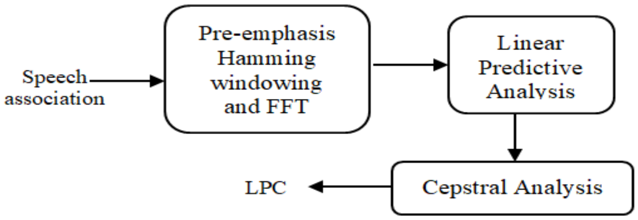
Speech association
Pre-emphasis Hamming windowing and FFT
Linear Predictive Analysis
Cepstral Analysis
LPC

Figure 1. LPCC feature process.

(LPCC) is identical as in Mel-frequency Cepstral coefficient. Pre-emphasis, hamming windowing, then fast Fourier transform are reserved based on (5). LPC (Linear Predictive coding) measures the speech sign to calculate approximately preparation, reducing their impacts start with individuals speech signal, consistently approaching that concentration near additional repetition of the remaining call. The procedure of eliminating individual format can be transposed filtering; besides the remaining indicator will be known as a collection. Now (LPC) system each assessment of the signal can be interconnected respectively a linear signal is called remains. This comparison is known as a linear predictor Cepstral investigation proposes of the transform discovery the cepstrum of a speech association. The idea after Cepstral examination will be deconvolution. In the discourse examine ground; the major consumption of Cepstral assessment remains with collected individual spoken region features start with a speech limit. This split can be arranged by taking the inverse discrete Fourier transform (IDFT) of the linearly joint log ranges for excitation consistently spoken region framework portions.

B. Probabilistic Neural Network

Probabilistic Neural Network (PNN) is a feed-forward neural network and needs to start with individuals Bayesian network. Its arrangement will be shown in figure 2.

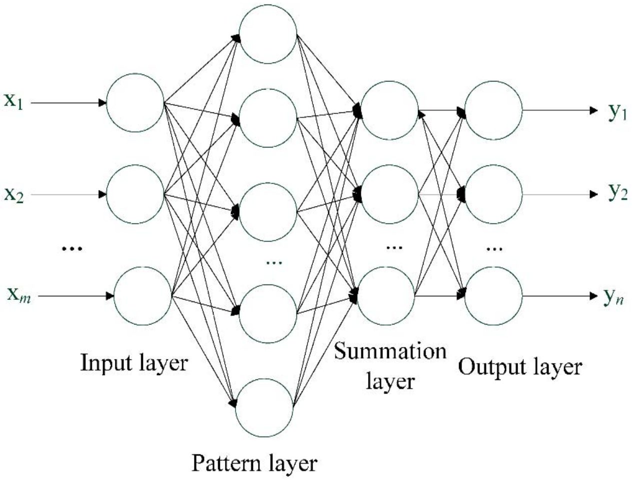

Figure 2. PNN process.

In Figure 2, The network consumes four layers and these are input layer, a pattern layer, summation layer and an output layer. Input layer means the analyzing speech. Pattern layer its function is to figure the Separation concerning the incomprehensible discourse besides the superior part of recognized speeches. The separation can be communicated similarly as.

$$d_{sk} = \exp\left(\frac{\|X_u - X_{sk}\|^2}{\sigma^2}\right) \tag{1}$$

Where x_u are characteristic vectors can be delivered by unidentified speech, so X is the *k-th* identified characteristic vector brought by the sth speaker. σ^2 is the difference of

Gaussian kernel. Inside summation layer, the uncertain probability is considered by utilizing the following equation.

$$p(X_u \mid \lambda_s) = \sum_{k=1}^{k_s} \frac{d_{sk}}{K_s} \qquad (2)$$

The place $p(X_u \mid \lambda_s)$ means the probability that unidentified speech belongs to the speaker signified by $\lambda_s . K_s$ is the full amount of speeches provided by the *s-th* speaker. In output layer, which speaker unidentified speech belongs to remains determined by consuming under below equation.

$$\lambda_u = \arg_s \max \ p(X_u \mid \lambda_s)\, p(\lambda_s) \qquad (3)$$

PNN is a kind of feed-forward neural network, that is not single, recognizes the related information as well as consider finalizing data, although the generative model, such as (GMM) classifier, recently consider the related information [6].

C. Gaussian Mixture Model

Convolutional, Gaussian mixture model (GMM) is broadly consumed for speaker classification in speaker recognition applications. It is a linear permutation of Gaussian probability compactness utility and is stated as below equation [3].

$$P(X \mid \lambda) = \sum_{i=1}^{M} \omega_i G_i(X) \qquad (4)$$

Where G_is Gaussian mixture purpose. X signifies a speech. ω_i $i = 123\ldots, M$, remain the mixture weights, then must fulfill $\sum_{i=1}^{M} \omega_i = 1$, λ represents a speaker and communicated as

$$\lambda = \{\omega_i, \mu_i, \textstyle\sum_i\};\ i = 1,2,3,\ldots M \qquad (5)$$

Where μ_i and $\sum_i$ exist mean and difference of the i-th Gaussian Purpose, correspondingly.

$$\lambda_u = \arg_s \max \ p(X_u \mid \lambda_s)\, p(\lambda_s);\ s = 1,2,3,\ldots,S \qquad (6)$$

In training method, λ can be frequently evaluated toward highest, Probability approximation, which can be enhanced through examining desire algorithm. In speaker classification step, which speaker is unidentified, speech belongs on it will be decided by utilizing the accompanying equation. λ_u in the speaker which is unidentified speech x_u be appropriate for.

The improvement of Gaussian Mixture Model (GMM) it can estimate some sort of density function, thus (GMM) has the ability to signify the conversion distance of the voiceprint information produced initially with a specific speaker's speeches. Though, (GMM) only considers similar information, also disregard challenging information. Consequently, GMM's performance a lot of cases is not perfect [7].

III. Proposed Speaker Recognition Model

This segment, we propose our innovative speaker recognition model called (MFCC+LSH). In Segment A, we define our discourse characteristic named Mel frequency Cepstral coefficients (MFCC). Furthermore, Locality Sensitivity Hashing Function (LSH) is defined in segment B. The speaker recognition model used in this paper is defined in segment C.

A. MFCC Feature Extraction

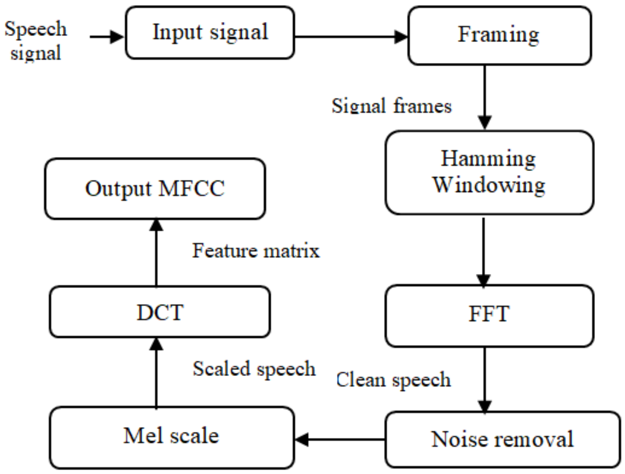

Figure 3. MFCC Feature Extraction Process

MFFC calculation will be correspondingly indicated in Fig.3. Respectively wavelet decomposed signal on separate frequency determination is framed by 20 ms hamming windowing with 10ms correspondence. Fast Fourier transform (FFT) will be accurate to every frame to produce composite spectral principles formerly magnitude of (FFT) is measured. The N-FFT magnitude coefficients are transformed to K-filter bank values. The filter bank values are determined by cross-wise multiplying the N-FFT scale coefficient to the K-triangular filter bank weighting purpose also formerly assembling or keeping the results starting with every filter triangle. The central point purpose of the striangle filter banks is distance set apart according to the Mel scale.

$$mel(f) = 2595 * log_{10}\left(1 + \frac{f}{100}\right) \qquad (7)$$

Once k log channel bank spectral values can be transformed over L Cepstral co-efficient using the Discrete Cosine Transform (DCT).

$$C_n = \sum_{k=1}^{K} log\,(S_k) cos\left[n\left(k - \tfrac{1}{2}\right)\tfrac{n}{K}\right] n = 1,2..L \qquad (8)$$

In this mathematical statement, n=0 i.e. C_n states to the Standard log power of the frame. Therefore over feature extraction phase, MFCCs would achieve the all wavelet decomposed input signal [8], [12].

B. Locality Sensitive Hashing (LSH)

Regardless of the fact that we can apply minhashing to compress huge data in small signatures and reserve the required equality of whatever combine from claiming speaker, it can make comprehensible should discover the pairs for best similarity efficiently. The motive is that the amount of claiming pairs of speakers can have a chance to be excessively large; regardless there would not a really huge numbers speaker.

Jaccard distance

The Jaccard distance for sets A and B will be |A ∩ B |/|A ∪ B |, that is, the amount of the extent union A and B of the extent for their union. We shall denote the Jaccard similarity of A and B by SIM (A, B). An important class of issues that Jaccard distance addresses are for finding speaker comparable complete an extensive quantity. An additional class of requisitions the place similarity for sets will be very important is called combined filtering, A methodology where we suggest to users things that were precious to different users who have exhibited comparable perceptions[9].

Locality Sensitive Hashing (LSH) exists a hashing method directing toward giving a robustly estimated result problem. It procedures family of locality sensitive hashing functions to hash the information into the small hash table is defined in Fig.2. This partition of the wave files. We used here only Mfcc Features to apply hash function. Mfcc feature Matrix size is (20 Rows 55Col) we convert the Mfcc features into the hash table it is also called column vector. The size of a column vector is up to 900 Mfcc features of one wave file, LSH classic algorithm.

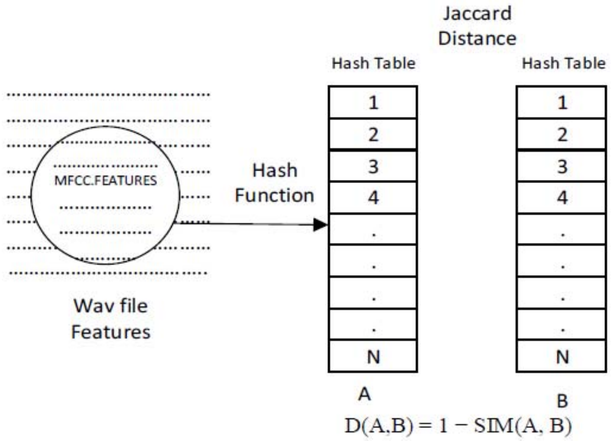

Figure 4. Hash Function and Jaccard Distance.

Above diagram said during the starting of the section, we define the Jaccard distance about sets toward d (a, b) = 1 − SIM (a, b). That is the Jaccard distance is 1 less those proportion of the sizes of the convergence Furthermore Union from claiming sets a and b. We must check that this capacity is of distance measure [9].

D (a, b) will be nonnegative as a result those measure of the crossing point can't surpass those span of the Uni.

D (a, b) = 0 whether a = b, in light of a ∪ a = a ∩ a = a. However, In a 6= b, after that the extent for a ∩ b can be strictly under the measure of a ∪ b, In this way, d (a, b) will be strictly positive.

D (a, b) = d (a, b) a result of both union what's more convergence would symmetric.

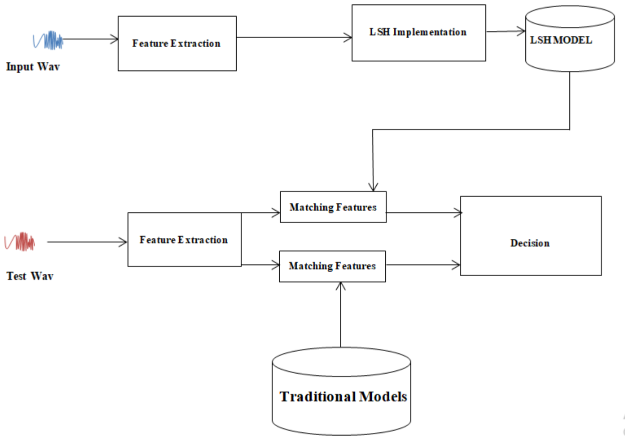

Figure 5. The process of Speaker Recognition Model.

In Figure-3 Represents the process of our proposed speaker identification model. We use input wave and extraction of the features from it and used (LSH) classifier to classify the input wave. Furthermore, we input test wave extracted of features and matching features from input wave features with lsh classifier and traditional models. The whole process made a decision of our model and comparison of traditional model results.

C. Speaker Identification Model

A speaker recognition model contained two leading fragments: characteristics extraction besides speaker classification constructed on the extracted speech feature based. Furthermore, discourse pre-processing stands an essential step earlier the characteristics extraction. In speech pre-processing, the discourse was distributed into lots of small frames. If the frame remained too small, we did not require sufficient trials to achieve the trustworthy data of speech; if it remained also extended, the speech would comprise as well much challenging data all over the frame. In [10], scholars presented that 15 ms frame remained supplementary beneficial than the extended or smaller individual. A 15 ms window at 8 KHz (designed for our data set) converts to 120 pieces. Consequently, in our circumstance, the speech can be partitioned into frames which held 120 trials. Then afterward the pre-processing step, we concentrated those speech characteristic from each frame. Once we extracted the existing (MFCC), we extracted two different categories of speech features such as (LPCC), (MFCC) besides the proposed our model. Separately. For speaker classification, the concentrated speech

Characteristics was utilized similarly as contributions to classifiers. In this paper, we utilized dual categories of classifiers such that (GMM) and (LSH). Reynolds toward el [3] recommended that the amount of Gaussians for (GMM) must be 8, 16 Furthermore 32. We differed those number of Gaussians from 8 to 32 should discover the most appropriate worth Furthermore found that 16 was the ideal worth. Therefore, we utilized the (GMM) for 16 mixtures to accomplish the speaker classification task. Concurrently, The (LSH) is also constructed according to the theory. It can utilize minhashing to compress huge data in small signatures and preserve that required equality of whatever combines from claiming speaker.

IV. Result and Discussion

In this segment, we stated the results of inclusive analysis accomplished proceeding the TIMIT corpus[11]. We primarily defined the investigational technique as well as a dataset in Segment A, and formerly stated the recognition exactness of our model in Segment B. In Segment C, we experiment the false positive error and false negative error of our model.

A. Investigational Dataset and Process

The experiment results were simulated based on python. Those speeches are for our dataset originally from TIMIT speech database. The sampling frequency is 8 KHz, the filter selection was mixed, and the required training wave and test wave. Our dataset contained 30 speakers with 18 male and 12 female speakers. All speakers provided 10 to 5-second-long speeches. All speakers come from 4 changed dialect regions indicated by ddr1, ddr2, ddr3, and ddr4. We presented that the changes among the waveforms of the speeches approaching from the 4 dialect regions were very much, even though the contents of the speeches were same. According to the 4 dialect regions, we separated the data set into 4 groups; the detail of the groups was presented in below Table.

TABLE I. The Details of 4 Speech Groups

Group Name	Different dialect region	No. of Male	No. of Females	No of speeches
g1	Ddr1	4	6	100
g2	Ddr2	5	5	100
g3	Ddr3	6	4	100
g4	Ddr4	5	5	100

In every group, the 10-fold cross-validation examines were used to test the 10 speakers. To define the examination process easily, we supposed that the speakers in a group were signified by s1, s2... s10. If we required testing s1, formerly s1 played the accurate speaker, and additional 9 speakers played the pretenders. In the experiment, we nominated 6 speeches of s1 for training and formerly used the remaining 4 speeches of s1 then 9 speeches approaching from changed pretenders for testing. For all speakers, we run the investigation 10 times with dissimilar training data and testing data to achieve the regular result. The trial results were composed on a computer using 2.0GHz Intel Core i5 CPU and 8 GB of memory.

B. Recognition Accuracy

In this segment, we initially associated the recognition accuracy of the four speaker recognition model like MFCC+GMM, LPCC+GMM, MFCC+PNN finally MFCC+LSH.

TABLE II. The Accuracy of the Four Speech Identification Models

Identification Model	Accuracy (%)			
	g1	g2	g3	g4
MFCC+GMM	80.1	80.2	79.4	79.6
LPCC+GMM	81.9	81.6	81.2	80.7
MFCC+PNN	89.9	89.7	89.5	89.2
MFCC+LSH	93.5	93.3	92.9	92.6

As presented in Table II, we establish the exactness of MFCC+GMM remained actual low once verified through our data set. Indifference, the conclusion in [11] indicated MFCC+GMM achieve an exactness of 87.3%. This modification existed for the reason that [11] consumptions great excellence discourses whose sampling rates remained 16KHz, while we utilized low-quality discourses whose sampling rates remained 8KHz. Those outcomes presented that MFCC+GMM stands not appropriate used for low-quality speech. Comparison, LPCC+GMM achieved great exactness for high-quality speech [2], though achieved low exactness for this dataset. Preceding the furthermore, we can find that MFCC+PNN achieved good exactness in our trial but not enough according to our proposed model. Then we used MFCC+LSH to propose our model which is robust and obtained very high accuracy in our experiment.

Then, we matched the exactness of GMM, PNN, besides LSH once MFCC was utilized the characteristics of the speech. In Table III the result is shown.

TABLE III. THE Accuracy of the GMM, PNN and LSH Consuming MFCC as Speech Feature

Classifiers	Accuracy (%)			
	g1	g2	g3	g4
GMM	80.1	80.2	79.4	79.6
PNN	89.9	89.7	89.5	89.2
LSH	93.5	93.3	92.9	92.6

In Table III, we similarly establish that the speaker recognition model grounded on LSH took enhanced performance than based on GMM and PNN model MFCC can achieve an exactness of greater than 92.9% once LSH was employed such as the classifier, however, it just achieves an exactness of less than 89.5% once PNN was consumed. Furthermore, it just finds correctness of lower than 80.2% when GMM was utilized. The unique motive of these results was that LSH was a type of minhashing big data into a small signature and also considered the associated competing data. It converts whole data into on big signature and then we utilized the similar data and robustness than other models. PNN remained a type of function feed-forward neural network which not only measured the correlated data but similarly measured the competing data, whereas GMM objective measured the associated data. In completely, the accuracy research indicated that our new model, which existed based on MFCC and LSH, found the maximum identification accuracy. Hence, the proposed model is appropriate for speaker identification task.

C. False Positive Error and False Negative Error

If a recognition model consumed capability to discard pretenders, formerly its false positive error (FPE) could be small. Proceeding the additionally, if the model remained competent to admit accurate speaker, formerly its false negative error (FNE) could be small. So, a speaker recognition model could be far from functional, if it's FPE and FNE remained too high. Table IV then Table V presented the FPEs and FNEs of the above speaker identification model.

TABLE IV. THE FPE OF THE THREE SPEECH FEATURES

Recognition Model	**FPE (%)**			
	g1	g2	g3	g4
MFCC+GMM	8.1	7.9	8.4	8.4
LPCC+GMM	5.9	5.8	6.2	6.3
MFCC+PNN	4.9	4.7	5.0	5.1
MFCC+LSH	2.5	3.4	3.0	4.3

TABLE V. THE DETAIL OF LSH USED IN THIS EXPERIMENT

Recognition Model	**FNE(%)**			
	g1	g2	g3	g4
MFCC+GMM	11.8	11.9	12.3	12.1
LPCC+GMM	10.2	10.1	10.6	10.9
MFCC+PNN	4.9	4.7	5.0	5.1
MFCC+LSH	4.0	3.3	4.1	3.4

V. CONCLUSIONS

To increase the achievement of speaker recognition model preceding the complaint that the speech of quality is low, this paper initially proposed a new speaker recognition model titled MFCC+LSH by engaging MFCC and LSH. The important improvement of proposed model can take benefit of MFCC and LSH on similar period to achieve a worthy achievement intended for low-quality discourse communicated. The achievement of this model is predictable consuming TIMIT. Associated with traditional MFCC+GMM, LPCC+GMM then MFCC+PNN models, the investigational outcomes indicate that the proposed model is capable to achieve great exactness.

In future work, we can use the same technique that can use different datasets to know the different speakers with particular name of the speaker. It can be a person with a name or can be animal with a name.

ACKNOWLEDGMENT

The authors thankfully acknowledge the appreciated analysis clarifications then recommendations given by the editors and the anonymous reviewers. The authors thank Professor Kun She for his assistance in the preparation of the document. The authors also thank all of the volunteers for their speech recordings, which are used to carry on the experiments in this paper. This work is supported by Sichuan Science and Technology Support Program under grant No.2016GZ0073

REFERENCES

[1] F. M. Chauhan and N. P. Desai, "Mel Frequency Cepstral Coefficients based on speaker identification in a noisy environment using wiener filter," In proceeding of 2014 International Green Computing Communication and Electrical Engineering, pp.1-5, 6-8 March 2014.

[2] S. Yella, N. Gupta and M. Dougherty, "Comparison of pattern recognition techniques for the classification of impact acoustic emissions", Transportation Research Part C: Emerging Technologies, vol. 15, no. 6, pp. 345-360, 2007.

[3] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," Speech and Audio Processing, vol. 3, pp. 72-83, Jan 1995.

[4] P. Suba B. Bharathi, "Analyzing the Performance of Speaker Identification Task Using Different Short Term and Long Term Features" 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)

[5] Octavian Cheng, Wa1eed Abdulla (2005), 'Performance Evaluation of Front-end Processing for Speech Recognition Systems, Electrical and Computer Engineering Department, School of Engineering, The University of Auckland.

[6] Lei Lei, Kun She," Speaker Identification using Wavelet Shannon Entropy and Probabilistic Neural Network" 978-1-5090-4093-3/16/$31.00 ©2016

[7] E. Manitsas, R. Singh and G. Strbac, "Distribution system state estimation using an artificial neural network approach for pseudo measurement modeling," power system, vol. 27, pp. 1888-1896, Apr 2012.

[8] Vikram.C.M. K.Umarani, Phoneme Independent Pathological Voice Detection Using Wavelet-Based MFCCs, GMM-SVM Hybrid Classifier 978-1-4673-6217-7/13/$31.00_c 2013.

[9] Jure Leskovec Stanford Univ. Anand Rajaraman Milliway Labs Jeffrey D. Ullman Stanford Univ. Mining of Massive Datasets Copyright c 2010, 2011, 2012, 2013, 2014 Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman

[10] S. m. mirhassani and H.-N. Ting, "Fuzz-based discriminative feature representation for children's speech recognition," Digital Signal Processing, vol. 31, pp. 102-114, Aug 2014.

[11] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "Feature extraction technique using ERB like wavelet sub-band periodic and aperiodic decomposition for TIMIT phoneme recognition," International Journal of speech technology, vol. 17, pp. 389-399, Dec 2014.

[12] Jorge, Hector and Enrique, "Speaker recognition using Mel Frequency Cepstral Coefficients and Vector Quantization Techniques," In proceeding of 2012 22nd International Conference on Electrical Communications and Computers, pp.248-251, 27-29 Feb 2009.