Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/120609

Deep Neural Network based Speaker Verification System using Features from Glottal Activity Regions

P. Shanmugapriya[1], *V. Mohan*[1,*], *T. Jayasankar*[2] *and Y. Venkataramani*[3]

[1] Department of ECE, Saranathan College of Engineering, Tiruchirappalli–620012, Tamil Nadu, India
[2] Department of ECE, BIT campus, Anna University, Tiruchirappalli, Tamil Nadu, India
[3] Dean (R & D), Saranathan College of Engineering, Tiruhirappalli–620012, Tamil Nadu, India

Received: 2 Jul. 2018, Revised: 2 Sep. 2018, Accepted: 9 Sep. 2018
Published online: 1 Nov. 2018

Abstract: We propose a Deep Neural Network (DNN)-based Speaker Verification (SV) system using features derived from Glottal Activity (GA) regions. Glottal activity regions are detected through Glottal Closure Instant (GCI), Normalized Autocorrelation Peak Strength (NAPS) and Higher Order Statistics (HOS) from speech signal. For the detection of GA regions, the speech signal is represented in terms of Zero Frequency Filtered Signal (ZFFS) and Integrated Linear Predicted Residual (ILPR). Mel Frequency Cepstral Coefficient (MFCC) and Wavelet Transformed Residual Coefficients extracted from the detected GA regions are used for analysing the performance of speaker verification system based on DNN and *i*-vector DNN. The results are reported on TIMIT database, NIST 2001 database and LibriSpeech database which proves that the features extracted from GA regions with *i*-vector DNN performs better than the conventional features based systems.

Keywords: Glottal Activity regions, Non-Glottal activity regions, Gaussian mixture model (GMM), Universal background model (UBM), Zero frequency filtered signal, Integrated Linear Prediction Residual and Speaker verification (SV) system.

1 Introduction

Speech Recognition has become inevitable in cases where it is required to analyze the characteristics of the speaker. To do this analysis, it is required to separate the excitation signal from the composite vocal tract output. The excitation signal which is the input of the vocal tract system is extracted from its output speech signal.

Glottal source processing refers to the process of identifying the glottal activity regions, modeling and characterization of the glottal source. The glottal flow is the passage of air via the vocal folds at the glottis [1]. During the speech production, the airflow from the lungs is modulated by vibration of vocal folds. Speech is a result of convolution between glottal flow signal and vocal tract system function.

Acoustic features such as MFCC and PLP are useful in representing the vocal tract characteristics. On the other hand, the features extracted from the excitation source signal provide useful information about the speaker characteristics.

The objective of this paper is to use glottal source information for determining the portions of the signal in which glottal activity has happened, are called Glottal Activity regions. The acoustic features extracted from the GA regions provide specific information about speaker characteristics [2]. The use of MFCC features extracted from the GA regions in speaker recognition reduces the computational complexity of the speaker recognition system.

The baseline speaker recognition systems use MFCC features as the standard features for recognizing speaker. But the information exists in the excitation source signal is not utilized in MFCC. However, the use of the excitation information can be useful, when speakers having different manners of speaking are considered. Feature extraction methods which use the information contained in the glottal excitation have also been proposed by many researchers in the literature.

Three distinct features namely GCI, NAPS and HOS which are the indicators of energy, periodicity and asymmetrical nature of glottal source signal such as ZFFS and ILPR are extracted and used to detect the glottal

* Corresponding author e-mail: mohansec@outlook.com

significant portions [3]. Detection of GA regions is carried out based on the glottal activity. Further MFCC features from the detected GA regions are used for modeling the speaker through DNN for speaker verification.

Recently, the glottal activity in glottal regions is gaining significance for extracting speaker characteristics [4] in text-dependent speaker verification system. Further, the significance of speaker information exists in the glottal regions is established by [5] through feature extraction from the GA regions and used for *i*-vector speaker verification system.

All aforesaid approaches convey that although the information extracted from the residual signal does not execute better than MFCC features, features extracted from the detected GA regions yields an improvement.

The paper is organized as follows: Section 2 introduces the method of extraction of three parameters from different representations of speech signal used for the detection of GA regions. Section 3 discusses the development of GMM-UBM, *i*-vector and DNN for modeling the speaker. Section 4 includes the results and performance comparison between the baseline systems, namely GMM-UBM system and proposed DNN and *i*-vector DNN-based system for various features.

2 Detection of Glottal Activity and Feature Extraction

In this work, GA regions are detected based on the energy of zero frequency filtered signal and residual signal. Glottal Activity is the excitation of the vocal tract during generation of sounds. It can be detected by the features present in the excitation source signal. Initially, GA is detected from the energy-based representation of excitation [6]. But, the energy is not sufficient for detecting the glottal activity region. Because the significant amount of energy is not present during the entire region of glottal activity. Hence, in addition to GCI, periodic nature during GA and asymmetric nature i.e. having high energy during glottal closure than glottal opening are included for GA detection [7]. These characteristics are predominantly available in the ZFFS and ILPR.

The effect of vocal tract is reduced by applying zero frequency filters to the speech signal. The effect of impulse like excitation reflected at dc component is not affected by the time varying nature of the vocal tract system [8]. Hence when the speech signal is passed through cascade of two ideal zero frequency digital resonators, the output contains the characteristics of discontinuities. The signal obtained at the output of the resonator is called ZFFS. The magnitude of the filtered signal is high during glottal activity. It has periodicity as well as asymmetry nature in each cycle of glottal activity.

LPC is one of the most influential speech analysis techniques and it has gained popularity as an extractor of excitation signal from speech [9]. When the speech signal is passed through the speech analysis filter, the redundancy in the signal is removed and the residual error is generated as output. Speech can be expressed as the response of LTI system excited by either quasi-periodic pulses, or random noise. The characteristics of LTI system which resembles the vocal tract can be extracted by the analysis of linear prediction on the speech signal. This method provides a robust, reliable, and accurate estimation of the parameters. The residual signal has high frequency components due to pre-emphasis and it represents the characteristics of the source. The residual signal can be obtained through LP-based inverse filtering of the speech signal. The Integrated LP Residual (ILPR) is derived from the inverse filtering with non pre-emphasized speech signal. ILPR indicates the closing state of glottal by high magnitude and vice versa. This shows that the ILPR has asymmetrical nature and periodicity also.

2.1 Attributes which demonstrates the Glottal Activity

2.1.1 Glottal Closure Instant (GCI)

The excitation of the vocal tract with strong source signal corresponds to rapid closure of the vocal folds. Maximum intensity in the ZFFS reveals the rate at which the glottal is closed. The narrowband nature of DC resonators is used to measure the GCI. The epoch locations correspond to the peak excitation can be obtained from the slope of zero frequency filtered signal.

Algorithm for measuring the GCI:

1. Filter the signal through DC component removal structure
2. Determine the position of significant excitation or epochs
3. Compute the slope of the filtered signal near the excitation.
4. GCI is the absolute slope of ZFFS at epoch location

It is given by

$$S_e(k) = |y(k+1) - y(k)| \tag{1}$$

where k is the epoch location. $S_e(k)$ gives the GCI at the epoch location.

2.1.2 Normalized Autocorrelation Peak Strength (NAPS)

Quasi periodic nature of ZFFS and ILPR can be extracted from the attribute NAPS. Because the position of peak value in the normalized auto correlation of the signal indicates the position of GA. The NAPS is comparatively large in GA regions and small in other regions.

The NAPS is given by

$$AN_p(k) = \frac{\sum_{n=1}^{N} x(n)x(n-k)}{\sum_{n=1}^{N} x^2(n)} \tag{2}$$

where
p is the number of frame
N is the number of samples in a frame
k is the delay which represent the position of largest peak in NAPS.

2.1.3 Higher Order Statistics (HOS)

In statistics, the term higher-order statistics (HOS) refers to third and higher order moments, namely, skewness and kurtosis. To capture the asymmetric nature of the glottal pulse, a suitable value of Skewness-Kurtosis Ratio (SKR) is used. This ratio makes the position of GA depend on the value of moments and not function of energy of the signal. It is determined as:

$$SKR = \frac{\gamma^2}{\beta^{1.5}} \tag{3}$$

where Skewness (γ) and Kurtosis(β) are given by

$$\begin{aligned} \gamma &= \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\bar{x})^3}{\left(\frac{1}{N}\sum_{n=1}^{N}(x(n)-\bar{x})^2\right)^{\frac{3}{2}}}, \\ \beta &= \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\bar{x})^4}{\left(\frac{1}{N}\sum_{n=1}^{N}(x(n)-\bar{x})^2\right)^{2}} - 3, \end{aligned} \tag{4}$$

where $\bar{x}$ is mean of $x(n)$.

Asymmetrical nature of glottal pulse is determined from SKR since it is high in the GA regions.

2.2 Detection of Glottal Activity

The parameters which describe the characteristics of periodicity, asymmetrical nature, and combination of those two with GCI are investigated to identify the region of GA. Periodicity is calculated from the normalized autocorrelation peak strength (NAPS) of speech signal. The higher-order statistics (HOS) which provides the information about the level of asymmetry of the speech signal is also calculated. The source signal is represented by zero-frequency filtered signal (ZFFS) [10] and integrated linear prediction residual (ILPR) and they are used for further processing. The process of detecting glottal activity is illustrated in Fig. 2. The contribution of this work is in representing the speech signal in terms of features extracted from the detected GA regions and utilization of those features for improving the verification accuracy of the proposed system.

Autocorrelation plays a key role in calculation of the normalized auto correlation peak strength. Delay is used in calculation of NAPS which indicates the location of the largest peak. It is the indicator of periodicity. It is relatively high in GA regions and small in other than GA region. This proves that the ZFFS is also an effective parameter in the separation of GA and other than GA regions. Autocorrelation analysis is mainly used in fluorescence correlation spectroscopy.

Higher-order statistics is calculated by skewness to kurtosis ratio. Kurtosis is defined as the cumulate of order 4 and skewness is the cumulate of order 3. ILPR is obtained by passing the speech signal through inverse filter without performing pre-emphasis. When the inverse filtering is performed based on linear prediction, it can separate the excitation signal from the vocal tract behavior. This residual signal has high magnitude during the closing instant of glottal and less magnitude during opening instant of glottal. ILPR also has the periodicity property like ZFFS and has high energy during glottal activity regions [11, 12]. This measure is computed for every frames of the signal and interpolated.

2.3 Feature Extraction from the detected GA regions

In this work, Mel Frequency Cepstral Coefficients are extracted from speech signal. Two standard feature-based channel compensation techniques, Cepstral Mean Subtraction (CMS) [13] and RelAtive SpecTrAl (RASTA) sifting [14] and feature warping method Gaussianization are applied in order to enhance the recognition rate. At that point first order and second order deltas are also appended to the Gaussianized cepstral vector.

2.3.1 Wavelet Transformed Residual Signal(WTRS)

The temporal pitch variation present in the voiced speech is useful for speaker recognition. The magnitude of pitch harmonics has also been established to be significant feature for speaker identification [15]. The detailed vocal source information can be extracted by inverse filtering the speech signal with the vocal tract filter parameters estimated during the glottal closing instant (GCI).

The linear predictive residual signal is obtained using

$$e(n) = s(n) - \sum_{k=1}^{12} a_k s(n-k) \tag{5}$$

and their amplitude is normalized. With the pitch period of the signal, pitch pulses in the signal are located. Then

for each pitch pulse, pitch synchronous wavelet analysis is applied for the period of two-pitch pulses. The wavelet coefficients are calculated using the windowed residual signal.

$$w(a,b) = \frac{1}{\sqrt{|a|}} \sum_n e_h(n) \Psi^*\left(\frac{n-b}{a}\right) \tag{6}$$

where $a = \{2^k | k = 1,2,\ldots,K\}$ and $b = 1,2,\ldots,N$ where N is the order of the window and $\Psi^*(n)$ is the complex conjugate of fourth order Daubechies wavelet basis function. K is set as 4 such that the signal is decomposed into four sub-bands at different sub bands: 8000–4000 Hz, 4000–2000 Hz, 2000–1000 Hz, and 1000–500 Hz. Each group of coefficients is partitioned into M sub-groups, where M determines the amount of temporal data that can be captured by the wavelet coefficients. Then the L2 norm of each sub-group of coefficients is computed. It is set as a feature parameter.

3 Proposed Speaker Verification System

Initially, the parameters which are required for detecting the GA regions are extracted from the speech signal. After GA regions are identified, they are used for feature extraction process through which MFCC and WTRC are extracted and utilized for training the proposed system. The analysis is made for following systems: (a) Baseline GMM-UBM (b) *i*-vector-based speaker verification system and (c) *i*-vector with DNN-based speaker verification system.

3.1 GMM-UBM Approach

GMM is a probabilistic model which is defined in terms of weighted sum of Gaussian component densities. The parameters of GMM are estimated through Expectation-Maximization (EM) algorithm for the training data. The global model is developed from the entire dataset is used for developing speaker model through MAP adaptation. UBM is represented by $U = \{m_c, \Sigma_c, \eta_c\}$, where $c = 1,2,\ldots,C$, Σ_c is Covariance matrix, η_c is the weight associated with mixture and m_c is the mean vector.

3.2 Deep Neural Network Approach

Deep learning is an emerging tool with different set of algorithms. DNN is used to model high-level data by using multiple processing layers. It is composed of multiple non-linear transformations [16]. One of the features of deep learning utilizes the efficient algorithms for unsupervised or semi-supervised feature learning. Fig. 1 shows the Detection of GA from ZFFS and ILPR.

Deep learning algorithms are based on distributed representations. The underlying assumption behind distributed representations is that observed data are generated by the interactions of factors organized in layers. Deep learning adds the assumption that these layers of factors correspond to levels of abstraction or composition. Fig. 2 shows the architecture of DNN used in the proposed system.

The weight updates can be done via stochastic gradient descent using the following equation:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \tag{7}$$

Here, η is the learning rate, and C is the cost function. The choice of the cost function depends on factors such as the learning type (supervised, unsupervised, reinforcement, etc.) and the activation function. When performing supervised learning on a multiclass classification problem, the activation function and cost function are the softmax functions and cross entropy function, respectively.

The softmax function is defined as

$$p_j = \frac{e^{x_j}}{\sum_k e^{x_j}} \tag{8}$$

where p_j represents the class probability (output of the unit j) and x_j and x_k represent the total input to units j and k of the same level respectively.

Cross entropy is defined as

$$C_j = \sum_j d_j \log(p_j) \tag{9}$$

where d_j represents the target probability for output unit j and p_j is the probability output for j after applying the activation function.

There are various deep learning architectures such as DNN, Convolutional Neural Network (CNN), deep belief networks and recurrent neural networks have been used in many applications. Computer vision, automatic speech recognition, audio recognition and bioinformatics are some of the areas in which the deep learning architectures are applied. Deep learning architectures have been shown to produce state-of-the-art results on various tasks.

DNN employs many layers of nonlinear processing units for feature extraction and transformation [17]. The output from each layer is applied as input for the next layer. DNN is based on the unsupervised learning of multiple levels of features. High-level features are resultant from low-level features to structure a hierarchical representation. In current research works, DNN is used to predict the posterior probability of the speech. The observation probability can be determined from the posteriors and priors using Bayes rule, as follows:

$$P(x/q) = \frac{P(q/x)P(x)}{P(q)} \tag{10}$$

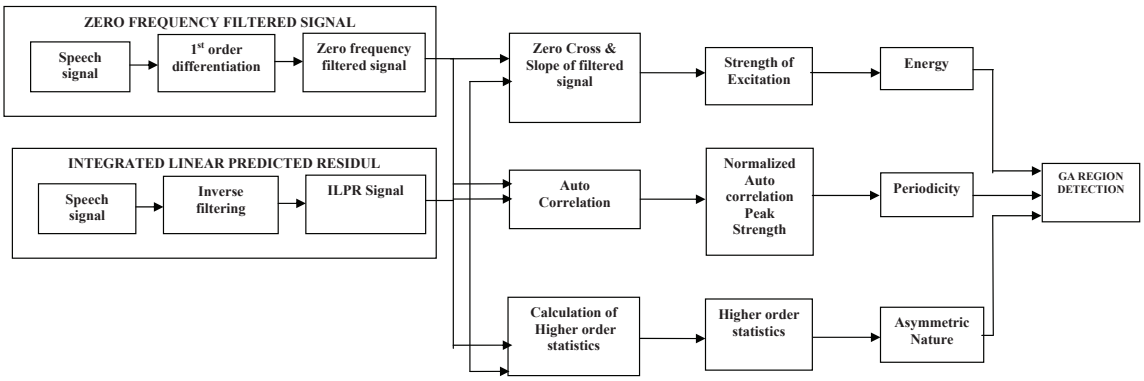

Fig. 1: Detection of GA from ZFFS and ILPR

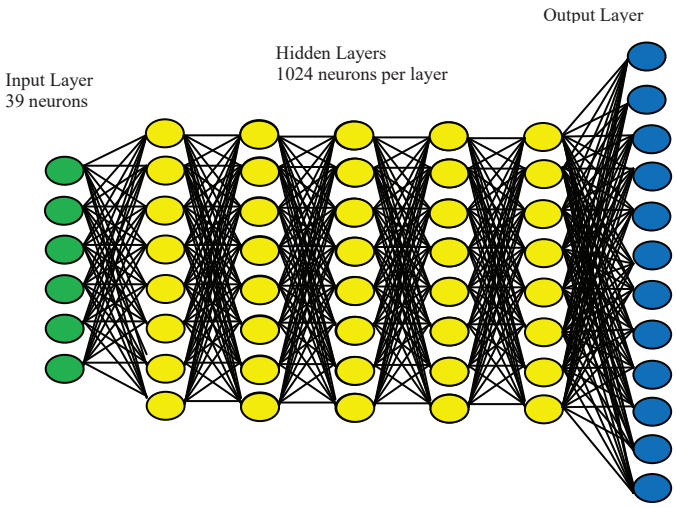

Fig. 2: Architecture of DNN used in the proposed system

where $P(x/q)$ is the observation probability required for decoding. In this method, DNN is used to predict the speaker for a given utterance of speech. Since the entire utterance corresponds to a particular speaker, the frame level DNN posteriors are combined to make a single decision score.

3.3 i-vector Approach

In speaker recognition area, the *i*-vector approach is useful for the transformation of high dimensional Gaussian super vector into a low-dimensional vector. Since it considers the background high-level information, *i*-vector usually improves the recognition rate [18]. Researchers have used *i*-vector for popular speech-related projects, including speaker verification [19], speaker recognition [20], stress recognition [21], and spoofing detection [22].

The *i*-vector modeling is given as:

$$M = m + Tx \tag{11}$$

where M denotes GMM super vector obtained from MAP adaptation—speaker and channel dependent super vector; m denotes speaker and channel independent GMM-UBM mean super vector constructed from the universal background model (UBM); T represents total variability matrix. It is a low-rank projection matrix obtained from all training data by factor analysis training which represents the important variability in the mean super vector space.

x is low-dimensional vector whose MAP estimate is *i*-vector with standard normal distribution.

The speech signal obtained from all the speakers are transformed into short vectors. It is determined from the zero and first order statistics. The procedure to learn the total variability subspace T relies on EM algorithm that maximizes the likelihood over the training set of target speaker and non target speaker.

Given a background model, Σ is initialized by covariance of the UBM. T and x are initialized randomly. They are estimated by a recursive process with the following steps:

Table 1: Datasets, subsets and number of male and female speakers in the database NIST 2001, TIMIT and LibriSpeech

Dataset	Subset	No. of Males	No. of Females
NIST 2001	Background	90	97
	Target	17	30
	Imposter	50	50
TIMIT	Background	48	101
	Target	18	31
	Imposter	25	25
LibriSpeech	Training class	120	100

1.E-step: the parameters of posterior distribution of x for each set of speech signal are calculated using the current estimates of T, Σ and m.
2.M-step: T and Σ are updated by a linear regression.
3.The process is continued till stable expectation of x is reached.

4 Implementation of the Proposed SV System

In this work, experiments are performed using TIMIT database, NIST 2001 database and LibriSpeech database. The experimental study illustrates the effectiveness of proposed approach in providing the speaker discriminating characteristics existing in GA region.

4.1 Data

For performing experimental analysis, we have used two different methods. One is all enrolment and test utterances are within the same dataset.

TIMIT database contains 630 speakers (192 females and 438 males) who come from 8 different English dialect regions, they are used for evaluation of the system. Each speaker provides ten speech samples which are sampled at 16 KHz. All female and male speech samples are used to obtain gender-dependent background models that represent the common characteristics. 384 speakers (192 females and 192 males) are randomly selected and their speech samples are used for developing target and imposter models in GMM-UBM approach.

NIST 2001 database consists of samples from 234 persons (which includes 122 Females and 112 Males). NIST 2001 database is classified into 3 major classes as follows: background, development and evaluation. Background set includes the development set as well. The background class is used for training the UBM and *i*-vector extractor. In all the experiments WCCN and score Normalization are used. A brief review on the databases such as NIST 2001, LibriSpeech and TIMIT databases is illustrated in Table 1.

Table 2: Parameter used for training DNN

Parameter	Value/functions
L2 weight Regularizer	0.001
Sparsity Regularizer	4
Sparsity Proportion	0.05
Loss Function	Mean Square Error (MSE)
Transfer function of both encoder and decoder	sparse Logistic sigmoid function, linear transfer function
Maximum epochs	1000
Techniques used to encode	Hamming, Linear block, Cyclic
Training algorithm	Scaled conjugate gradient descent algorithm

Table 3: Performance comparison of various approaches for two different features on NIST 2001 database

Feature	System	Male		Female	
		EER(%)	DCF	EER(%)	DCF
MFCC	GMM-UBM	0.67	0.031	0.82	0.152
	i-vector	1.52	0.112	1.43	0.141
	DNN	1.91	0.013	1.92	0.015
	i-vector-DNN	0.20	0.001	0.24	0.002
WTRC	GMM	1.51	0.135	1.53	0.142
	i-vector	1.75	0.156	1.67	0.150
	DNN	2.75	0.447	2.68	0.482
	i-vector-DNN	0.62	0.032	0.56	0.028

4.2 Features

39 dimensional MFCCs are used for the development of the proposed system. The feature vector is of 39 dimensions with 13 MFCC, 13Δ MFCC and 13$\Delta\Delta$ MFCC. Initially, the speech signal is processed for detecting the GA regions and the MFCC features are calculated from the detected GA regions. This method reduces the number of frames as well as improves the recognition accuracy. The speech portions corresponding to the GA regions are manipulated in frame duration of 20ms with overlapping of 10ms duration in training and testing. The normalization of feature vector is performed through cepstral mean subtraction and cepstral variance normalization. Besides the cepstral features, wavelet transformed residual signal coefficients are also used in our experiments. The performance of the system is compared for all these features.

4.3 Systems

The aim of the proposed work is to find the performance of the system for the features detected from GAs and non-GAs. The parameters of DNN used for training in the proposed system is listed in Table 2.

Table 4: Performance comparison of features and different models on LibriSpeech database

Features	System	Male		Female	
		EER(%)	DCF	EER(%)	DCF
MFCC	GMM	0.53	0.152	0.48	0.132
	i-vector	0.35	0.121	0.36	0.122
	DNN	0.31	0.082	0.33	0.098
	i-vector-DNN	0.22	0.071	0.20	0.061
WTRC	GMM	2.23	0.661	2.01	0.566
	i-vector	3.44	0.798	3.22	0.698
	DNN	2.76	0.799	2.55	0.687
	i-vector-DNN	1.64	0.325	1.57	0.541

Table 5: Performance comparison of features and models on TIMIT database

Features	System	Male		Female	
		EER(%)	DCF	EER(%)	DCF
MFCC	GMM	0.10	0.011	0.12	0.102
	i-vector	0.22	0.120	0.43	0.121
	DNN	0.09	0.013	0.02	0.015
	i-vector-DNN	0.01	0.001	0.01	0.002
WTRC	GMM	1.16	0.201	1.82	0.289
	i-vector	1.02	0.322	1.03	0.161
	DNN	0.91	0.014	0.92	0.017
	i-vector-DNN	0.02	0.005	0.04	0.007

Table 6: Performance comparison of the proposed system and system without GA detection (MFCC features and, NIST 2001 dataset)

	System	Male	Female
		EER(%)	EER(%)
Features from GA Region	GMM	0.67	0.82
	i-vector	1.52	1.43
	DNN	1.91	1.92
	i-vector-DNN	0.20	0.24
Features without GA Detection	GMM	2.67	2.84
	i-vector	1.83	1.96
	DNN	2.01	2.43
	i-vector-DNN	0.90	0.86

4.4 Result and Discussion

From the analysis elucidated in Table 3 for NIST 2001 dataset, Table 4 for LibriSpeech dataset and Table 5 for TIMIT dataset, it is understood that the performance of the *i*-vector-DNN approach is superior than the other existing techniques in terms of EER in % and DCF. Also it is clear that the performance metrics of the SV system has been considerably improved for the features extracted only from the GA regions and illustrated in Table 6.

5 Conclusion

Experimental results of the proposed *i*-vector DNN-based speaker verification show that the features extracted such as MFCC and Wavelet-transformed residual coefficients provide improved EER in % and DCF for the proposed system. Efficiency of the proposed system have been analyzed with TIMIT database, NIST 2001 database and LibriSpeech database which illustrates that the features extracted from GA regions with *i*-vector DNN perform better than the conventional features-based systems.

References

[1] Yu-Ren Chien, Daryush D. Mehta, Jón Guðnason, Matí as Zañartu and Thomas F. Quatieri, Evaluation of Glottal Inverse Filtering Algorithms Using a Physiologically Based Articulatory Speech Synthesizer, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, No. 8, pp. 1718–1730 (2017).

[2] N. Dhanajaya and B. Yegnanarayana, Voiced/nonvoiced detection based on robustness of voiced epochs, IEEE Signal Processing Letters, Vol. 17, No. 3, pp. 273–276 (2010).

[3] Nagaraj Adiga, S.R. Mahadeva Prasanna, Detection of Glottal Activity using different Parameters of source Information, IEEE Signal Processing Letters, Vol. 22, No. 11, pp. 2107–2111 (2015).

[4] K. Ramesh, S.R. Mahadeva Prasanna, Rohan Kumar Das, Significance of glottal activity detection and glottal signature for text-dependent speaker verification, In Proc. of International conference on Signal Processing and Communication, July 2014.

[5] Ashutosh Pandey, Rohan Kumar Das, Nagaraj Adiga, S.R. Mahadeva Prasanna, Significance of Glottal activity detection for Speaker Verification in degraded and Limited Data Condition, In Proc. of TENCON 2015-IEEE Region 10 Conference, Nov 2015.

[6] K. Murty and B. Yegnanarayana, Epoch extraction from speech signals, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 8, pp. 1602–1613, (2008).

[7] Nagaraj Adiga and S.R. Mahadeva Prasanna, Detection of Glottal Activity Using Different Attributes of Source Information, IEEE Signal Processing Letters, Vol. 22, No. 11, pp. 2107–2111 (2015).

[8] B. Yegnanarayana and V. Suryakanth Gangashetty, Sadhana, Epoch based analysis of speech signals, Indian Academy of Sciences, Vol. 36, No. 5, pp. 651–697 (2011).

[9] K.S.R. Murthy, B. Yegnanarayana and M.A. Joseph, Characterization of glottal activity from speech signal, IEEE Signal Process. Lett., Vol. 16, No. 6, pp. 469–472, (2009).

[10] S.R. Mahadeva Prasanna and G. Pradhan, Significance of Vowel-Like Regions For Speaker Verification Under Degraded Condition, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 8, pp. 2552–2565 (2011).

[11] S.R.M. Prasanna and B. Yegnanarayana, Detection of Vowel Onset Point Events Using Excitation Source Information, In Proc. of Interspeech, pp. 1133–1136 (2005).

[12] S.R.M. Prasanna, B.V.S. Reddy, and P. Krishnamoorthy, Vowel onset Point Detection Using Source, Spectral Peaks, and Modulation Spectrum Energies, IEEE Trans. Audio, Speech, Lang. Process., Vol. 17, No. 4, pp. 556–565 (2009).

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, Front-end factor analysis for speaker verification, IEEE trans. Audio, speech, Langauge Process., Vol. 19, No. 4, pp. 788–798 (2011).
[14] T. Kinnunen and H. Li, An overview of text-independent speaker recognition: From features to supervectors, Speech Communication, Vol. 52, pp. 12–40 (2010).
[15] J. Li, D. Yu, J. Huang and Y. Gong, Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM, In: Proceedings of the IEEE Spoken Language Technology Workshop, pp. 131–136 (2012).
[16] V. Ehsan, L. Xin, M. Erik, L.M. Ignacio, and G.-D. Javier, Deep neural networks for small footprint text-dependent speaker verification, IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Vol. 28, No. 4, pp. 357–366 (2014).
[17] Dong Wang, Lantian Li, Zhiyuan Tang and Thomas Fang Zheng, Deep speaker verification: Do we need end to end?, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 177–181 (2017).
[18] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani and J.H.L. Hansen, Robust *i*-vector extraction for neural network adaptation in noisy environment, In Proc. of Interspeech 2015, pp. 2854–2857 (2015).
[19] D. Reynolds, T. Quatieri and R. Dunn, Speaker verification using adapted gaussian mixture models, Digital Signal Processing, Vol. 10, No. 1, pp. 19–41 (2000).
[20] T. Kinnunen and H. Li, An overview of text-independent speaker recognition: From features to supervectors, Speech communication, Vol. 52, No. 1, pp. 12–40 (2010).
[21] C. Zhang, G. Liu, C. Yu and J. H.L. Hansen, *i*-vector based physical task stress detection with different fusion strategies, In Proc. of Interspeech 2015, pp. 2689–2693 (2015).
[22] C. Zhang, S. Ranjan, M.K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly and J. H.L Hansen, Joint information from nonlinear and linear features for spoofing detection: an *i*-vector/DNN-based approach, In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2016), pp. 5035–5039 (2016).

P. Shanmugapriya is currently working as Associate Professor in the Department of Electronics and Communication Engineering, Saranathan College of Engineering, Tamil Nadu, India. She obtained M.Tech from NIT, Trichy, in the discipline of Master of Communication systems in the year of 2005. She has completed her PhD from the faculty of Information and Communication, Anna University, Chennai, Tamil Nadu, India in 2015. Her fields of interests include Speech Processing, Soft computing and Pattern recognition.

V. Mohan is currently working as Associate Professor in the Department of Electronics and Communication Engineering, Saranathan College of Engineering, Tamil Nadu, India. He obtained M.E., from Mepco Schlenk College of Engineering, Sivakasi, in the discipline of Master of Communication systems in the year of 2001. He obtained his PhD in the area of Image Compression in the faculty of Information and Communication from Anna University, Chennai, Tamil Nadu, India in 2015. He has nearly two decades of teaching experience at UG and PG levels. He delivered many guest lecture at various colleges in the field of Digital signal processing and Image processing and Image Compression. His fields of interests include Image Compression, Soft computing, Pattern recognition, Antenna Design and Digital Signal Processing.

T. Jayasankar received the B.E. degree in Electronics and Communication Engineering from Bharathiyar University, Coimbatore in 2001 and M.E. degree at Madurai Kamaraj University,Madurai in 2003 and Ph.D.in Speech Processing at Anna University Chennai 2017. At present, he is an Assistant Professor in the Electronics and Communication Engineering department, University College of Engineering, Anna University, Bharathidasan Institute of Technology Campus, Tiruchirappalli, Tamilnadu, India. He is a member of IEI, ISTE. He has been a lecturer at graduate and post-graduate level and has participated in a number of International and National level conferences and workshops. He has published around 25 papers in the reputed international journals and more than 15 papers in the international and national conferences. His main interest is currently speech synthesis, speech and image processing and wireless networks.

Y. Venkataramani is a Professor, Dean (R&D), Saranathan College of Engineering, Tamil Nadu, India. He has got more than 40 years of experience in the field of teaching. He has guided several PhDs in the area of Electronics and Communication engineering. His fields of interests include Computer Networking, soft computing and Pattern classification.