ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

Improvement of the Speaker Verification System with Feature Level and Score Level Normalization Techniques

Kshirod Sarmah[1], Utpal Bhattacharjee[2]

Research Scholar, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India[1]

Associate Professor, Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, Arunachal Pradesh, India[2]

ABSTRACT: The performance of a text independent Speaker verification (SV) system has degraded when speaker model training is done in one environment while the testing is done in another, due to mismatching of phonetic contents of speech utterances, recording environment, session variability and sensor variability of training and testing criteria, which are major problems in speaker verification system. The robustness of SV system has been improved by applying different Voice Activity Detection (VAD) techniques like Cepstral Mean Normalization (CMN), Cepstral Variance Normalization (CVN) techniques in features level and score normalization techniques in score level. In this paper we report the experiment carried out on the recently collected speaker recognition database Arunachali Language Speech Database (ALS-DB). The collected database is evaluated with Gaussian mixture model and Universal Background Model (GMM-UBM) and Mel- Frequency Cepstral Coefficients (MFCC) with its first and second order derivatives as well as Prosodic features as a front end feature vectors. The performance of the speaker verification system has been improved by applying CVN at the feature level as well as score normalization technique Test-normalization (T-Norm) in the score level. And also we observe that the performance of SV system vastly improved while applying CVN in feature level and T-Norm in score level at the same time. We observe that combining MFCC with Prosodic features improved the performance of the SV system with **7.08**%, while T-Norm improved the SV system with **3.22**% and CVN has improved with **3.90**%.

Keywords: Speaker Verification; GMM-UBM; Mel-Frequency Cepstral Coefficients; CVN; T-Norm

I. INTRODUCTION

Automatic Speaker Verification (ASV) is one of the most natural and economical methods for solving the problems of unauthorised use of computer and communication systems and multilevel access control [1]. Speaker Verification is one of the biometric tasks to be verified the validity of user's acceptance or rejection for a computer system. Speaker verification (SV) is a technology which is used to verify a person's identity from their speech utterances. In general, the ASV system consists of five phases: Speech data acquisition in digital format, speaker related feature extraction, enrolment to generate speaker models ,pattern matching, and finally making an accept or reject decision. Figure 1 describes the basic concept of speaker verification system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

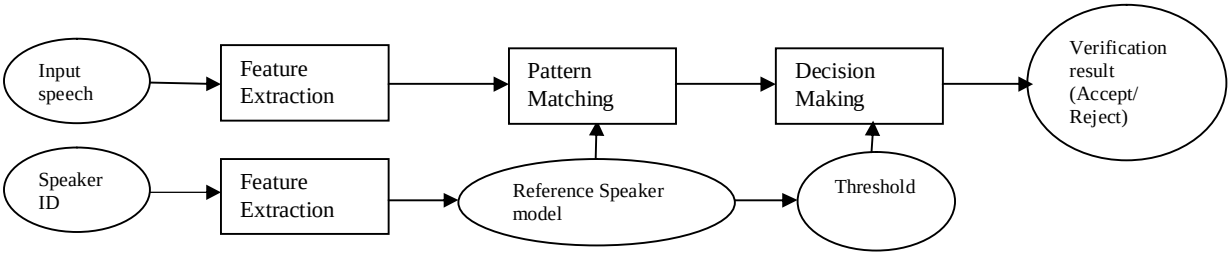

Fig.1 Speaker Verification System.

For speaker modeling technique, the state of the art speaker verification systems uses Gaussian mixture models (GMM) approached with adapted mean from universal background models (UBM) [2] or support vector machines (SVM) over GMM super-vectors [3]. Mel-frequency cepstral coefficients (MFCC) are the most popular features, although, the traditional MFCC is very sensitive to noise interference for which the performance of SV system degrades also because of the mismatches between training and testing, but MFCC combined with supra-segmental information such as prosody and speaking style for improved performance [4]. The main applications of the SV technology are employed in person authentication and in forensic science [5]. With the growth in the wireless telecommunication, many of these applications are now accessed through mobile phones.

II. RELATED WORKS

Speaker Recognition systems have typically used generative models like Vector Quantization (VQ) [12] or Gaussian Mixture Models (GMMs) [2].A single-state hidden Markov model (HMM) that presently known as Gaussian Mixture Model (GMM) was proposed by Rose, as one of the robust parametric modelling technique for the text-independent speaker recognition system [6].The generative model is generally trained using maximum likelihood (ML) principle. The main disadvantage of the ML approach is that it doesn't generalize well to unseen speech data with finite amount of training material. To solve this problem Maximum a Posteriori (MAP) approach of training is sufficient which is also known as universal background model (UBM) [2]. In MAP approach, prior knowledge of the distribution of model parameters is incorporated into modeling process [15].

Besides the GMM-UBM, other speaker modeling techniques are developed recently. The most successful ones include the support vector using GMM super vector (GSV-SVM) [9]. That concatenates the GMM mean vectors as the input for SVM training test. Another important development namely joint factor analysis (JFA) [11], which jointly models the channel subspace and the speaker subspace. Although these innovative methods achieve rapid progress, GMM-UBM is still the basis for their recent developments.

The primary aim of the Voice Activity Detection (VAD) is to identify the speech segments from a given audio signal which is an important sub-component of SV system [17]. In this case VAD computes the energy values of all frames and cancels the too low absolute energy frames based on the threshold. In our case the threshold value is (-55dB).

The principle of feature normalization is to use generic noise suppression techniques to enhance the quality of feature extraction strategy [17]. In this case we used Cepstral Mean Normalization (CMN) and Cepstral Variance Normalization (CVN) which reduces the channel affect.

Score normalization is the transformation of speaker verification output scores to enhance the effectiveness of the detection threshold by aligning the score distribution speaker models. T-Norm is one of the popular score normalization methods. T-Norm speaker models are scored in parallel with the target speaker model [13].As the adapted universal background model (UBM) provides fast scoring so T-norm is efficient in an adapted UBM system [14].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

The rest of the paper is organized as follows: Section–III describes the details of the speaker verification database. Section–IV details the speaker modeling techniques. Features and score normalization techniques have been explained in the Section-V. The experimental setup, data used in the experiments and result obtained are described in Section VI and VII. Finally the paper is concluded in Section–VIII.

III. SPEAKER VERIFICATION CORPUS

In this section we have used the recently collected Arunachali Language Speech Database (ALS-DB) consisting 200 speakers of Arunachal Pradesh [19][20].Arunachal Pradesh of North East India is one of the linguistically richest and most diverse regions in all of Asia, being home to at least thirty and possibly as many as fifty distinct languages in addition to innumerable dialects and subdialects there of [7].

To study the impact of language variability as well as channel variability on speaker verification task, ALS-DB is collected in multilingual and multi-channel environment. Each speaker is recorded for three different languages – English, Hindi and a Local language, which belongs to any one of the four major Arunachali languages - Adi, Nyishi, Galo and Apatani. Each recording is of 4-5 minutes duration. Speech data are recorded in parallel across four recording devices, which are listed in table -1.

TABLE 1: DEVICE TYPE AND RECORDING SPECIFICATIONS

Device Sl. No	Device Type	Sampling Rate	File Format
Device 1	Table mounted microphone	16 kHz	wav
Device 2	Headset microphone	16 kHz	wav
Device 3	Laptop microphone	16 kHz	Wav
Device 4	Portable Voice Recorder	44.1 kHz	mp3

The speech data collection is done in laboratory environment with air conditioner, server and other equipments switched on. The speech data is contributed by 120 male and 80 female informants chosen from the age group 20-50 years. During the recording, the subject was asked to read a story from the school book of duration 4-5 minutes in each language for twice and the second reading was considered for recording. Each informant participates in four recording sessions and there is a gap of at least one week between two sessions.

IV. GMM-UBM AS SPEAKER MODEL

The GMM-UBM approach for speaker verification system can be considered primarily as a four phase process. At the first phase, a gender independent UBM model is generated which is a GMM that built based on the Expectation-Maximization (EM) algorithm and using utterances from a very large population of speakers [2]. The target speaker specific models are then obtained through the adaptation of mean from the UBM using the speaker's training speech and a modified realization of the maximum a posteriori (MAP) approach [2]. In the testing phase, a fast scoring procedure is used in order to reduce the number of computations [2]. This involves determining the top few scoring mixtures in the UBM for each feature vectors and then computing the likelihood of the target speaker model using the score for its corresponding mixtures. The scoring process is then repeated for all the feature vectors in the test utterance to obtain the average log likelihood score for each of the UBM and the target speaker model. Finally, UBM-based

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

normalization is performed by subtracting the log likelihood score of the UBM from that of the target speaker model. This is firstly to minimize the effect of unseen data, and secondly to deal with the data quality mismatch [2].

Normally, SV systems use mel-frequency cepstral coefficients (MFCCs) as a feature vector and the speaker model λ_s is parameterized by the set $\{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ where w_i are the weights, $\boldsymbol{\mu}_i$ are the mean vectors, and $\boldsymbol{\Sigma}_i$ are the covariance matrices. In the testing stage, feature vectors **X** are extracted from a test signal. A log-likelihood ratio $\Lambda(\mathbf{X})$ is computed by scoring the test feature vectors against the claimant model and the UBM.

$$\Lambda(X) = \log P(X|\lambda_S) - \log P(X|\lambda_{UBM}) \quad (1)$$

The claimant speaker is accepted if $\Lambda(\mathbf{X}) \geq \theta$ or else rejected. The important problem in SV is to find a decision threshold θ for the decision making [21][22]. The uncertainty in θ is mainly due to score variability between the trials.

V. FEATURES AND SCORE NORMALIZATION

Normalizations at the stage of feature extraction are implemented to reduce the effect of the noise, speech signal distortion as well as the channel distortion. State-of-the –art speaker recognition system have used several approaches in order to enhance the performance in feature level scores. The cepstral mean substraction (CMS) [8] is a blind deconvolution that comprises the substraction of the utterance mean of the cepstral coefficients from each feature. In the similar way, the variance normalization (CVN) is also applied. Hence, the new features will fit a zero mean and variance one distribution. Another well- known feature normalization is RASTA (Relative Spectras) [23]. While CMS focus on the stationary convolution of the noise due to the channel, RASTA reduces the effect of the varying channel; which removes low and high modulation frequencies [16]. The three of them are the most commonly used feature normalization techniques in the SV system.

In score normalization, the final score of the SV system is normalized relative to a set of other speaker models termed as cohort. Score normalization techniques have been mainly derived from the study of Li and Porter [18]. The main purpose of score normalization is to transform scores from different speakers into a similar range so that a common speaker independent verification threshold can be used [17]. As we know that in SV system the score variability comes from various sources. First, the probable mismatch between enrollment data which is used for training speaker models and the data that is used for testing is one of the main problems in SV system. Secondly, the nature and properties of the enrollment data can vary between the speakers, the phonetic content, the duration, the environmental noises as well as the quality of the speaker model training. Other two main factors intra-speaker and inter-speaker variability also affects in the performance in SV system. On the other hand some environment condition changes in transmission channel, recording devices or acoustical environment may also considered as a potential factor affecting the reliability of decision boundaries. To overcome above problems score normalization techniques have been introduced to cope with score variability and to make speaker-independent decision threshold tuning easier [12]. The basic of the normalization techniques is to center the imposter score distribution by applying on each score generated by the SV system. The All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

The general formula to compute score normalization for speech signal X and speaker model λ is given as follows.

$$\check{s}_\lambda(X) = \frac{S_\lambda(X) - \mu_\lambda}{\sigma_\lambda} \quad (2)$$

Where $\check{s}_\lambda(X)$ is the normalized scores, $S_\lambda(X)$ is final score and μ_λ and σ_λ are normalized parameters known as estimated mean and standard deviation of the imposter score distribution. Imposter distribution represents the largest part of the score distribution variance.

There are different types of normalization techniques which can be seen in speaker recognition system. These are Z-Norm, H-Norm, T-Norm, HT-Norm, C-Norm etc.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

The zero normalization (Z-Norm) had been more used in SV in the middle of nineties. The advantage of Z-norm is that the estimation of the normalized parameters can be performed offline during speaker model training.

The test-normalization (T-Norm) can be performed online during testing. In T-norm, during testing, the incoming speech signal is classically compared with claimed speaker model as well as with a set of imposter models to estimate imposter score distribution and normalized parameters consecutively [12].

The handset T-Norm (HT-Norm) is introduced to deal with handset type information. Here, handset-dependent normalization parameters are estimated by testing each incoming speech signal against handset dependent imposter models [12].

VI. EXPERIMENT SETUP OF THE BASELINE SYSTEM

In these works, the baseline system, a speaker verification system was developed using Gaussian Mixture Model with Universal Background Model (GMM-UBM) based on modeling approach. Following specification is used:

An energy based silence detector (VAD) is used to identify and discard the silence frames prior to feature extraction with threshold (-55dB).

Front end features = Total 42 dimensions (36 MFCCs + 6 Prosodic Features)

MFCC Features = Total 36 dimensions with first and second order derivatives of total 12 no. of MFCC coefficients excluding the 0^{th}cepstral coefficient.

Prosodic Features = Total 6 dimensional prosodic features vector consist of pitch , short time energy and its first and second order derivatives (Δpitch, Δenergy, ΔΔpitch and ΔΔenergy)

The coefficients were extracted from a speech sampled at 16 KHz with 16 bits/sample resolution.

Frame size: 20 ms

Frame rate: 10 ms

Windowing: Hamming

Pre-emphasis filter $H(z) = 1- a.z^{-1}$ with Pre-emphasis factor a = 0.97

No. Of Filterbanks = 24

Cepstral Mean Subtraction (CMS) has been applied on all features to reduce the effect of channel mismatch. In this approach we can also apply Cepstral Variance Normalization (CVN) which forces the feature vectors to follow a zero mean and a unit variance distribution in feature level solution to get more robustness results.

The Gaussian mixture model with 512 Gaussian components has been used for both the UBM and speaker model. The UBM was created by training the speaker model with 50 male and 50 female speaker's data with 256 Gaussian components each male and female model with Expectation Maximization (EM) algorithm. Finally UBM model is created by pooling the both male and female models of total 512 Gaussian components. The speaker models were created by adapting only the mean parameters of the UBM using Maximum a Posteriori (MAP) approach with the speaker specific data.

Finally, we have applied T-Norm technique to improve the performance of SV system. The normalization parameters are estimated using score derived from a set of imposter speaker models from the 2003 NIST SRE database that trained with same baseline system.

The detection error trade-off (DET) curve has been plotted using log likelihood ratio between the claimed model and the UBM and the equal error rate (EER) obtained from the DET curve has been used as a measure for the performance of the speaker verification system. Another measurement MinDCF values has also been evaluated.

VII. EXPERIMENT AND RESULT

All the experiments reported in this paper are carried out using the database ALS-DB described in section II. An energy based silence detector VAD is used to identify and discard the silence frames prior to feature extraction. Only

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

data from the headset microphone (Device 2) has been considered in the present study. All the four available sessions were considered for the experiments. Each speaker model was trained using first two complete sessions. The test sequences were extracted from the next two sessions. The training set consists of speech data of length 120 seconds per speaker. The test set consists of speech data of length 15 seconds, 30 seconds and 45 seconds. The test set contains more than 3500 test segments of varying length and each test segment will be evaluated against 11 hypothesized speakers of the same sex as segment speaker [10]. The overall performance of the SV system has been evaluated in the Fig.2 with DET curves and MinDCF values.

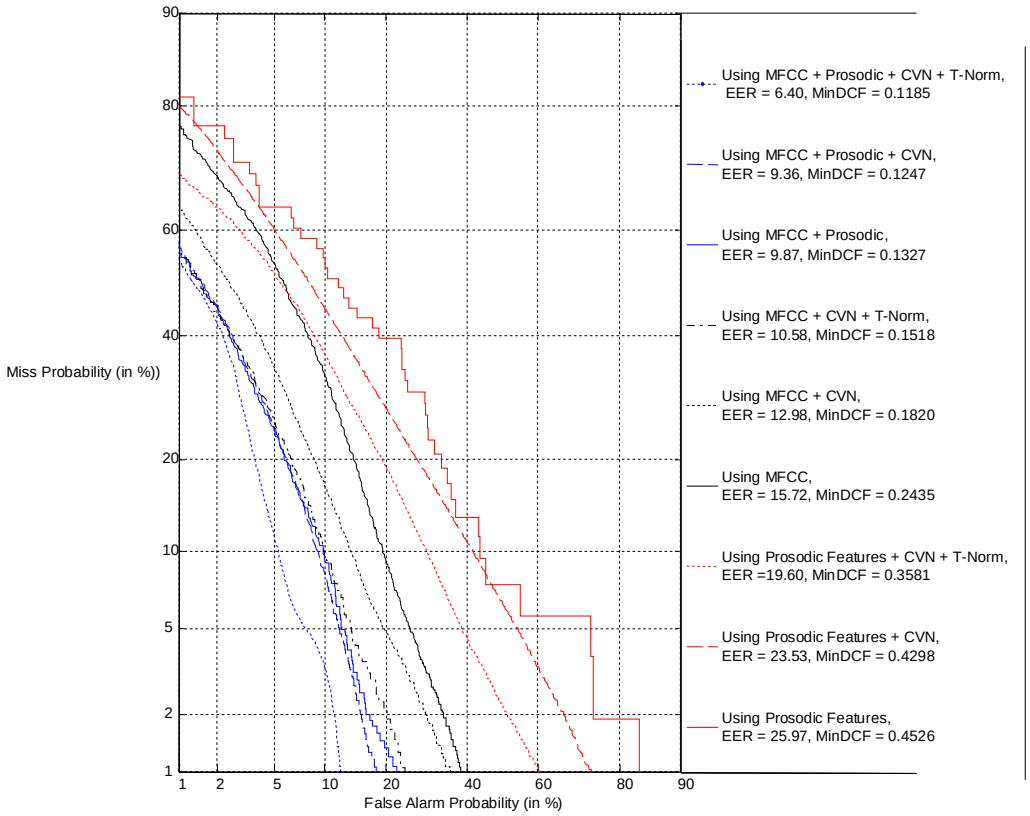

Fig.2 DET curves for the speaker verification system using feature and score normalization techniques for Device-2.

TABLE 1. THE PERFORMANCE OF SPEAKER VERIFICATION SYSTEM IN EER AND MINDCF VALUES

Features	EER%	MinDCF
MFCC + Prosodic + CVN + T-Norm	6.40	0.1085
MFCC + Prosodic + CVN	9.36	0.1247
MFCC + Prosodic	9.87	0.1327
MFCC + CVN + T-Norm	10.58	0.1518
MFCC + CVN	12.98	0.1820
MFCC	15.72	0.2435

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

Prosodic + CVN + T-Norm	19.60	0.3581
Prosodic + CVN	23.53	0.4298
Prosodic	25.97	0.4526

VIII. CONCLUSION

In the present study, experiments have been carried out on a recently collected speech database (ALS-DB) for evaluation of the effectiveness of GMM–UBM for speaker verification with feature and score normalization techniques. From the experimental point of view we can conclude that the performance of the speaker verification system is very poor while using only prosodic features, but it has been improved while combing the both acoustic (MFCCs) and prosodic features. And also we observe that the performance of SV system can be vastly improved while applying CVN in feature level and T-Norm in score level at the same time. Here we found EER of **6.40**% with Minimum DCF value **0.1085**. We observe that combining MFCC with prosodic features improves the performance of the SV system with **7.08%,** while T-Norm improves the SV system with **3.22%** and CVN has improved with **3.90%.**

ACKNOWLEDGEMENT

This work has been supported by the ongoing project grant No. 12(12)/2009-ESD sponsored by the Department of Information Technology, Government of India.

REFERENCES

[1]. J.P. Campbell, (1997), 'Speaker recognition: a tutorial', In Proceedings of the IEEE *85*, 9, September 1997, pp.1437–1462.

[2]. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn (2000), 'Speaker Verification Using Adapted Gaussian Mixture Models', Digital Signal Processing, vol. 10(1–3), pp. 19-41.

[3]. W. M. Campbell, D. E. Sturim, and D. A. Reynolds (2006), 'Support vector machines using GMM supervectors for speaker verification', IEEE Signal Processing Letters, vol. 13, pp. 308–311.

[4]. A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey (2003), 'Modeling prosodic dynamics for speaker recognition', in Proc. ICASSP '03, vol. 4, pp. IV–788–91.

[5]. B.C.Haris, G.Pradhan, A.Misra, S.Shukla, R.Sinha and S.R.M.Prasanna (2011), 'Multi-variability Speech Database for Robust Speaker Recognition', In Proc. NCC, pp. 1-5.

[6]. R. Rose and R.A. Reynolds (1990). 'Text independent speaker identification using automatic acoustic segmentation', Proc. ICASSP, pp. 293-296.

[7]. Arunachal Pradesh, http://en.wikipedia.org/wiki/Arunachal_Pradesh.

[8]. S. Furui (1981), 'Cepstral analysis technique for automatic speaker verification', IEEE Transactions on Acoustic,Speech and Signal Processing 29,2. pp 254-272.

[9]. W. M. Campbell , D. E. Sturim , D. A. Reynolds (2006), 'Support vector machines using GMM supervectors for speaker verification', IEEE Signal Processing Letters 13,5, pp. 308-311.

[10]. NIST 2003 Evaluation plan, http:// www.itl.nist.gov/iad/mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.

[11]. S.C.Yin, R.Rose, P.Kenny and P. Dumouchel (2007), 'A Joint factor analysis approach to progressive model adaptation in text independent speaker verification'. IEEE Transactions on Audio, Speech and Language Processing, 15 (7), pp. 1999-2010.

[12]. F.J. Bimbot (2004), 'A Tutorial on Text-Independent Speaker Verification'. EURASIP Journal on Applied Signal Processing 2004:4,pp. 430–451.

[13]. D. A. Reynolds and D.E. Sturim., " Speaker Adaptive Cohort Selection for Tnorm in text-independent speaker verification." MIT Lincoln Laboratory, Lexington, MA USA.

[14]. D.A Reynolds (1997), ' Comparison of Background Normalization Methods for Text-Independent Speaker Verification', In Proceeding of EUROSPEECH '97, Rhodes, Greece, pp. 963-966.

[15]. Ville Hautamaki, Tomi Kinnunen , Ismo Karkkainen, Juhani Saastamoinen, Marko Tuononen and Pasi Franti (2008), 'Maximum a Posteriori Adaptation of the Centroid Model for Speaker Verification', IEEE signal Processing letters, vol.15.

[16]. Leibny P.G. Perera, Roberto A.Lopez and J.N. Flores (2011), ' Speaker Verification in Different Database Scenarios', Computation y Sistemas Vol.15 No.1, pp 17-26.

[17]. Tomi Kinnunen ,Haizhou Li (2010), 'An Overview of Text-Independent Speaker Recognition: from Features to Supervectors', Speech Communication 52(1):12-40.

[18]. K. P. Li and J. E. Porter (1988), 'Normalizations and selection of speech segments for speaker recognition scoring', in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '88), vol. 1, pp. 595–598.

[19]. Utpal Bhattacharjee and Kshirod Sarmah (2012), ' A Multilingual Speech Database for Speaker Recognition', Proc. IEEE, ISPCC.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

[20]. Utpal Bhattacharjee and Kshirod Sarmah (2013), 'Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment', International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6.

[21]. R. Auckenthaler, M. Carey, and H. Lloyd-Thomas (2000), 'Score normalization for test-independent speaker verification system', Digital Signal Processing, vol. 10, no. 1, pp. 42–54.

[22]. J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez (2004), 'Exploiting general knowledge in userdependent fusion strategies for multimodal biometric verification', Proc. Int. Conf. Acoust. Speech, Signal Process., vol. 5, pp. 617– 620..

[23]. H. Hermansky and N. Morgan (1994), ' RASTA processing of speech', IEEE Trans. On Speech and Audio Processing 2, pp. 578-589.

BIOGRAPHY

Kshirod Sarmah is a Research Scholar in the Computer Science and Engineering, Department, Rajiv Gandhi University, India. He received his Master of Science (M.Sc.) in Computer Science from Gauhat University, India in the year 2004. Currently he is pursuing his Ph.D. in Computer Science and Engineering from Rajiv Gandhi University. His research interest is in the field of Speech Processing and Robust Speaker Recognition.

Utpal Bhattacharjee is a Associate Professor in the Computer Science and Engineering Department, Rajiv Gandhi University, India. He received his Master of Computer Application (MCA) from Dibrugarh University, India and Ph.D. from Gauhati University,India in the year 1999 and 2008 respectively.His research interest is in the field of Speech Processing and Robust Speech/Speaker Recognition.