International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 4, August 2019, pp. 2960~2967
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i4.pp2960-2967

Bayesian distance metric learning and its application in automatic speaker recognition systems

Satyanand Singh
College of Engineering Science and Technology, Fiji National University, Fiji Island, Suva

Article Info

Article history:

Received Mar 4, 2019
Revised Mar 22, 2019
Accepted Apr 4, 2019

Keywords:

Automatic speaker recognition (ASR)
Language recognition evaluation (LRE)
Linear discrimination analysis (LDA)
Neighborhood component analysis (NCA)
Phone recognition and language modelling (PRLM)

ABSTRACT

This paper proposes state-of the-art Automatic Speaker Recognition System (ASR) based on Bayesian Distance Learning Metric as a feature extractor. In this modeling, I explored the constraints of the distance between modified and simplified i-vector pairs by the same speaker and different speakers. An approximation of the distance metric is used as a weighted covariance matrix from the higher eigenvectors of the covariance matrix, which is used to estimate the posterior distribution of the metric distance. Given a speaker tag, I select the data pair of the different speakers with the highest cosine score to form a set of speaker constraints. This collection captures the most discriminating variability between the speakers in the training data. This Bayesian distance learning approach achieves better performance than the most advanced methods. Furthermore, this method is insensitive to normalization compared to cosine scores. This method is very effective in the case of limited training data. The modified supervised i-vector based ASR system is evaluated on the NIST SRE 2008 database. The best performance of the combined cosine score EER 1.767% obtained using LDA200 + NCA200 + LDA200, and the best performance of Bayes_dml EER 1.775% obtained using LDA200 + NCA200 + LDA100. Bayesian_dml overcomes the combined norm of cosine scores and is the best result of the short2-short3 condition report for NIST SRE 2008 data.

Corresponding Author:

Satyanand Singh,
College of Engineering Science and Technology,
Fiji National University,
Fiji Island, Suva.
Email: Satyanand.singh@fnu.ac.fj

1. INTRODUCTION

Distinguish an individual by his or her voice is a God gifted human quality most take assure case in regular human-to-human association/ correspondence. Talking somebody via phone more often than not starts by distinguishing who is talking and, at the minimum in instances of recognizable speakers, a subjective confirmation by the audience that the personality is right and the discussion can continue. ASR frameworks have developed as a vital method for authenticating individuality in several online applications and in addition all in all business collaborations, criminology, what's more, law enforcement. Aside from individual validation for access control, speaker recognition is a vital device in law implementation, national security, and legal sciences [1, 2].

Extraordinarily, people routinely identify people by their voices with striking precision, particularly when the level of commonality with the subject is high. Normally, even a short nonlinguistic line, for example, a laugh quietly, is sufficient for us to recognize a well-known individual [3]. ASR research groups have produced different techniques pretty much autonomously for five decades [4]. On the other hand, native speaker recognition is a regular capacity of people which is, on occasion, extremely exact and

efficient. Latest studies on mind imaging [5, 6] have uncovered numerous subtle elements on how to perform psychological based speaker recognition, which may motivate new headings for ASR approach.

To begin with, considering the general exploration area, it would be helpful to illuminate what is incorporated by the term speaker recognition, which comprises of two option undertakings: speaker identification and verification. In speaker recognition, the assignment is to distinguish an unknown speaker from the database of known speakers. There are two kinds of ASR system open-set and close-set. If all speakers are known in a set then it is called as closed-set on the other hand, if the test speaker could likewise, be from outside the predefined known speaker set, this turns into a the open-set situation, and, along these lines, a world model or universal background model (UBM) [7] is required.

Speaker identification can be based on a voice stream that is content dependent or content free. This is more significant in speaker-verification systems in which a claimed person speaks a predefined text, like a password or personal identification number (PIN), to access system. All through this paper, the emphasis will be on text-independent automatic speaker recognition systems.

A few algorithmic and its computational advances have empowered noteworthy ASR performance in the cutting edge. Approaches utilizing phonotactic data, phoneme recognizer followed by language models Phone Recognition and Language Modeling (PRLM) parallel PRLM, have been appeared to be very effective [8]. In this phonotactic demonstrating structure, an arrangement of tokenizes is utilized to interpret the speech information into token strings or cross sections which are later scored by n-gram dialect models [9] or mapped into a sack of trigrams highlight vector for support vector machine (SVM).

In fact traditional Hidden Markov model (HMM) based speaker verification is generally based on the tokenizer class model, all tokenizations connected here to make a system [10], for example, Gaussian Mixture Model (GMM) tokenization [11], universal phone recognition (UPR) [12], articulator the property based methodology [13], deep neural network based telephone recognizer [14], just to give some examples.

With the presentation of shifted delta-cepstral (SDC) acoustic components [15], promising results utilizing the GMM system with the variable investigation [16, 17], supervector model [18]. Furthermore, maximum mutual information (MMI) based discriminative training [19] have likewise been accounted for LID. In this work, I concentrate on the acoustic level frameworks.

Acoustic-phonetic methodology, which is normally taken by specialists prepared on this, requires quantitative acoustic estimations from speech signal tests, and based on fact examination of the result. By and large, comparable phonetic units are removed from the known and addressed speech signal, and different acoustic parameters measured from these portions are evaluated. The Logistic Regression (LR)can be helpfully utilized as a part of this methodology since it depends on numerical parameters [20].

In spite of the fact that the acoustic-phonetic a methodology is a more target approach, it has some subjective components. For instance, an acoustic–phonetician may distinguish speech signal as being influenced by anxiety and after that perform the objective examination. In any, the case, whether the speaker was really under anxiety at that minute is a subjective amount controlled by the inspector through his or her experience. As on the date, aggregate variability i-vector ASR modeling has achieved critical consideration in both LID and SV areas because of its remarkable efficiency, less system complexity and compact system in size. In i-vector modeling, initial, a solitary element investigation is utilized as a front end to produce a low dimensional aggregate variability space which together models dialect, speaker and channel variability all together. At that point, inside this i-vector space, variability costs techniques, for example, Within-Class Covariance Normalization(WCCN), Linear Discriminative examination (LDA) and Nuisance Attribute Projection (NAP) [18], are performed to diminish the variability for consequent modeling (e.g., utilizing SVM, LR, and neural network and probabilistic linear discriminate analysis (PLDA) for Language Identification (LID).

In this paper, the conventional i-vectors are stretched out to label regularized regulated i-vectors by connecting the marked vector and the straight relapse lattice toward the end of the mean supervector what's more, the i-vector element stacking network, separately [21, 22]. I can let the added name vector be the parameter vector that I need to perform and relapse with age paralinguistic measures to make the proposed system reasonable for regression. The explanation behind utilizing a linear regression matrix W is that numerous back end classification modules in LID and SV is linear. Additionally, if the regression connection is not linear, I can use non-linear mapping as a preprocessing venture before creating the mark vectors. The commitment weight of each supervector measurement and every objective class in the target capacity is consequently ascertained by iterative preparing. The conventional i-vector framework serves as our baseline.

As a final point, motivated by the achievement of strong works for corrupted information based SV jobs, I likewise considered Gammatone frequency Cepstral coefficients (GFCC) features and the spectrotemporal Gabor features for powerful LID assignment on the corrupted information as extra execution enhancer steps. At the point when combined with customary MFCC and SDC speaker specific feature based frameworks, the general framework execution was further upgraded.

2. THE BASELINE i-VECTOR MODELING

Let us consider a model λ of C component GMM UBM with $\lambda_c = \{p_c, \mu_c, \sum_c\}$ where c = 1, . , , ..C and the speech signal Y={y_1, y_2……..,y_L } having L feature sequences. UBM based on BW statistics computed as follows;

$$N_c = \sum_{t=1}^{L} P(c|y_t, \lambda) \quad (1)$$

where GMM component is $c = 1, \dots, \dots, \dots C.$ and y_t on λ_c is occupancy probability component $P(c|y_t,)$.

$$F_c = \sum_{t=1}^{L} P(c|y_t, \lambda)(y_t - \mu_c) \quad (2)$$

I can generate the corresponding central mean supervector $\check{F}$ by concatenating all together $\check{F}_c$ as follows;

$$\check{F}_c = \frac{F_c = \sum_{t=1}^{L} P(c|y_t,\lambda)(y_t-\mu_c)}{\sum_{t=1}^{L} P(c|y_t,\lambda)} \quad (3)$$

$\check{F}_c$ can be projected on rectangular total variability matrix T (low rank factor loading) and i-vector x as follows;

$$\check{F}_c \rightarrow T_x \quad (4)$$

The total variability matrix T with C component GMM and acoustic features in D dimension can be represented as CDXK matrix similar to eigenvoice matrix V. By considering $\check{F}_c$ and the i-vector can be computed as follows;

$$x = \{I + T^t\Sigma^{-1}NT\}^{-1}T^{-1}\Sigma^{-1}N\check{F} \quad (5)$$

where N is CD X CD dimension diagonal matrix. To reduce the variability I applied two channel compensation methods in total variability space (i) LDA and (ii) WCCN. LDA minimizes intra-class variance and WCCN normalizes the cosine kernel. Let us consider two i-vector 1 and 2 then cosine kernel is defined as follows to adapt either PLDA or SVM classifier.

$$k(x_1, x_2) = \frac{\langle x_1, x_2 \rangle}{\|x_1\|_1 \|x_2\|_2} \quad (6)$$

3. IMPLEMENTATION OF MODIFIED AND SUPERVISED I-VECTOR

Using a combination of SVM classifiers and GMM supervectors have been an extremely fruitful methodology for ASR. Observing the way that the channel considers to contain speaker-dependant data, the speaker and channel elements were consolidated into a solitary space termed the total variability space.

3.1. Label-regularized supervised i-vector

Let us assume that the hidden variable i-vector is generation mean supervector. The steps involved in label-regularization as follows:

Step I. Compute multivariate Gaussian distribution for j^{th} utterances of

$P(x_j) = \mathcal{N}(0, I)$, $P(\check{F}_j|x_j) = \mathcal{N}(Tx_j, N_j^{-1}\Sigma)$. Where x_j = i-vector of j^{th} utterances.

Step II. Computation of posterior distribution of hidden variabl i-vector

$$P(x_j|\check{F}_j) = \mathcal{N}\left\{\left(I + T^t\Sigma^{-1}N_jT\right)^{-1}T^t\Sigma^{-1}N_j\check{F}_j, \left(I + T^t\Sigma^{-1}N_jT\right)^{-1}\right\}$$ where N_j = N vector.

Step III. Computation of discriminative i-vector, $P(x_j) = \mathcal{N}(0, I)$.

Step IV. Regularization of label information,

$$P\left[\begin{pmatrix}\check{F}_j \\ L_j\end{pmatrix}|x_j\right] = \mathcal{N}\left[\begin{pmatrix}Tx_j \\ Wx_j\end{pmatrix}\right], \left[\begin{pmatrix}N_j^{-1}\Sigma_1 \\ n_j^{-1}\Sigma_2\end{pmatrix}\right]$$

where $\check{F}_j$= Mean super vector, L_j= Label vector, Σ_1= CD dimension mean super vector, Σ_2= M dimension label vector.

Step V. Designing of two supervised label vectors

$$\text{Type 1: } L_{i,j} = \begin{cases} 1 & \text{for class i} \\ 0 & \text{otherwise'} \end{cases}$$

The class label will be correctly classified by regression matrix W. M is denoted as the dimensionality of label vector L_j. Total number of speakers in database to recognize =H then L_j is $H(H = M)$ dimension binary vector. $H - 1$ numbers of elements will have "0" and one element will have value "1".

Type 2 : $L_j = \overline{X}_{sj}, W = I$. The last iteration $\overline{X}_{sj}$ specify the sample mean vector and compel regression matrix to be identity matrix.

Step VI. Computation of the likelihood of the total training utterances of ASR system is:

$$\sum_{j=1}^{\Gamma} \ln\{P(\check{F}_j, L_j, x_j)\} = \sum_{j=1}^{\Gamma} \left[\ln\left\{ P\left(\begin{pmatrix} \check{F}_j \\ L_j \end{pmatrix} | x_j \right) \right\} + \ln\{P(x_j)\} \right]$$

Step VII. Computation of objective function J_m for Maximum Likelihood (ML):

$$J_m = \sum_{j=1}^{\Gamma} \left[\frac{1}{2} x_j^t x_j + \frac{1}{2} (\check{F}_j - Tx_j)^t \sum_1^{-t} N_j (\check{F}_j - Tx_j) + \frac{1}{2} (L_j - W_{xj})^t \sum_2^{-t} N_j (\check{F}_j - Tx_j) - \frac{1}{2} \ln(|\Sigma_1^{-t}|) - \frac{1}{2} \ln(|\Sigma_2^{-t}|) \right]$$

After simplifying objective function equation.

$$J_m = \sum_{j=1}^{\Gamma} \left[\frac{1}{2} x_j^t x_j + \frac{1}{2} (A)^t \sum_1^{-t} N_j (A) + \frac{1}{2} (B)^t \sum_2^{-t} N_j (B) - \frac{1}{2} \ln(|\Sigma_1^{-t}|) - \frac{1}{2} \ln(|\Sigma_2^{-t}|) \right]$$

where $A=\check{F}_j - Tx_j$ and $B=L_j - W_{xj}$

3.2. Modified and simplified i-vector

The cosine distance score is fast and robust, but additional computation are required to standardize the score. A better generation model will fully simulate the speech data and generate scores without standardization or calibration.

Feature extraction and training of ASR system based on i-vector is computationally very expansive. Let us consider the GMM size C, feature dimension as D and factor loading size as K. A single i-vector generation and its computation cost is $O[(K)^3 + (K)^2.C + (K.C.D)]$. The main objective is to redefine and reweight each and every speech data with mean super vector so that imbalance in intra-super vector can be compensated. The steps involved in simplification of i-vector as follows:

Step I - i-vector with approximated computational cost $O[(K)^3 + (K)^2.C + (K.C.D)]$.
Step II - supervised i-vector with approximated computational cost $O[(K)^3 + (K)^2.C + \{K.(C.D + M)\}]$.
Step III - modified i-vector without ID with approximated computational cost $O[(K)^3 + (K.C.D)]$.
Step IV - modified i-vector with ID with approximated computational cost $O[(K.C.D)]$.
Step V - modified and supervised i-vector without ID with approximated computational cost $O[(K)^3 + \{K.(C.D + M)\}]$.
Step VI - modified and supervised i-vector with ID with approximated computational cost $O[\{K.(C.D + M)\}]$.

For ASR application, I use a simple cosine distance classifier on simplified and modified i-vector of the target speker utterances w_{target} and test atturances w_{test} with decision threshold θ as follows.

$$score(w_{target}, w_{test}) = \frac{(w_{target}^t) w_{test}}{\|w_{target}\|.\|w_{test}\|} \gtrless \theta \quad (7)$$

3.3. Linear discrimamt Analysis for session compensation

Treating a single speaker as a class LDA attempts to define a new axis to minimize intra-class variance caused by session/channel effects and to maximize the differences between classes. In the representation of total variability, there is no clear compensation for variability between intersessions. However, the low dimensional representation allows making technical compensation in the new place,

with the benefit of less computational cost. I use a linear discrimination analysis (LDA) for session compensation. Displaying speakers as the class, LDA attempts to define a new axis to reduce inter-class variation due to session/channel effects and to maximize differences between classes.

I can define the problem of optimizing the LDA to find directions of q that maximize fisher criteria $J(q) = q^t S_b q / q^t S_w q$. Between-class and within-class covariance matricws are represented as $S_b = \sum_{s=1}^{S} (\overline{w^s} - \overline{w})(\overline{w^s} - \overline{w})^t$ and $S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} \sum_{s=1}^{S} (\overline{w^s} - \overline{w})(\overline{w^s} - \overline{w})^t$. Where n_s is the number of utterances of each speaker s, $\overline{w^s} = \left(\frac{1}{n_s}\right) \sum_{i=1}^{n_s} w_i^s$ is the mean of i-vector foe each speaker, speaker population is $\overline{w}$ and total number of speakers is S. Projection matrix is composed by eigenvectors of $S_w^{-1} S_b$ general matrix to get the optimization.

3.4. Distance matric learning technique

Using modified and supervised as a low-dimensional representation of the linguistic expression, cosine distance classifiers is used to measure the distance between classified cosine and the target user's accent and accentuation of the test. To define the distance between the vectors aiming to find a good distance metric in the feature space, there is an important issue in classification. In recent years, extensive research has been conducted on distance learning [23]. I explored two ways to distance learning metrics that are supervised in this paper. From now on, I will use modified and supervised to represent an utterance.

3.5. Component analysis based on neighborhood method

Neighborhood Component Analysis (NCA) near the random selection rule learns distance metrics to reduce the average classification error. Using transformation matrix B, each simplified and supervised modified and supervised w_i chooses another simplified and supervised modified and supervised as its neighbor with some probability $p_{i,j}$, which is defined at Euclidean distance in the transformed space:

$$p_{i,j} = \frac{\exp\left(-\|Bw_i - Bw_j\|^2\right)}{\sum_{k \neq i} \|Bw_i - Bw_k\|^2}, p_{i,i} = 0 \quad (8)$$

The probability of simplified and supervised modified and supervised w_i select a neighbor with the same speaker is $p_i = \sum_{j \in c_i} p_{i,j}$, where C_i is a set of the same speaker modified and superviseds. The projection matrix B should maximize the number of expected simplified and supervised modified and superviseds that select neighbors from the same speaker:

$$B = argmax_B f(B) = \sum_i \sum_{j \in c_i} p_{i,j} = \sum_i p_i \quad (9)$$

3.6. Bayesian distance metric learning approach

Here I used conjugate gradient method to obtain the best seed. The NCA provides the one time estimate of the distance metric and can be unbelievable when the number of training data is less. I use the Bayesian structure to estimate the distribution after the distance metric [8]. Looking at the speaker tag for each utterance, I can create two sets of barriers for the same speaker S and D. The possibility of defining two utterances and related to the same speaker or different speakers under a given matrix:

$$pr\left(y_{i,j} \middle| w_i, A, \mu\right) = \frac{1}{1 + exp\left(y_{i,j} \|w_i - w_j\|_A^2 - \mu\right)} \quad (10)$$

This parameter μ is the threshold for separating expressions for the same speaker parameters with different speakers. Only when value of μ less than distance from matrix A, two expressions are likely to be identified by the same speaker. The $y_{i,j}$ is defined as follows [24, 25]:

$$y_{i,j} = \begin{cases} +1 & (w_i, w_j) \epsilon S \\ -1 & (w_i, w_j) \epsilon D \end{cases} \quad (11)$$

I use the NCA as a preprocessing technique so that the vectors can be projected in such a place where the nearest neighbors of each simplified and supervised modified and supervised can share the same tag of high probability. Bayesian distance learning methods can emulate the distance between the carriers better and more dependable in a new place. Experimental results show the advantages of the Bayesian distance metric learning approach.

4. EXPERIMENTAL RESULTS

Experiments are performed on the 13 different speaker detection tests that are defined by the duration and type of training and test data on the NIST 2008 SRE dataset. I present results on the short2-short3 and 10sec-10sec conditions for the training and test voice data of conversational of five minutes duration. In this research it is also used for LDA and NCA training, and as the impostor set in the score normalization step. A 600- dimension modified and supervised is extracted from each utterance. The Equal Error Rate (EER) and the minimum Detection Cost Function (minDCF) are used as metrics for evaluation. Cosine similarity scoring and Bayesian distance metric learning on the short2-short3 condition of the NIST 2008 SRE dataset. The Bayesian distance metric learning algorithm is referred to as \Bayes dml, cosine score after the combined score normalization for Cosine Score combined norm and PLDA with Gaussian GPLDA. Constraints from all possible modified and supervised pairs from the same speaker S, apply the cosine score to all possible modified and supervised pairs from different speakers and selects those with the highest score to form constraint D since these pairs are the largest distinction that distinguishes the metric distance. Since the number of all possible different speaker pairs is very large, I chose twice the number of similar speaker pairs from all possible speaker pairs to form a D. The experimental experiment showed a large set of offset speaker constraints (similar speaker pairs) does not improve performance but requires more calculations, and a smaller set of speaker constraints (same size) with different speaker constraints will degrade the ASR performance. The comparison of cosine scores, Bayesian_dml of NIST 2008 SRE GPLDA normalization is shown in Table 1.

Table 1. Comparison of cosine scores, Bayesian_dml of NIST 2008 SRE GPLDA normalization

Combination Norms		EER	minDCF
LDA 200	Cosine Score	2.541%	0.01445
	Cosine Score combined norm	1.790%	0.0097
	Bayes dml	2.158%	0.0106
	Bayes dml+znorm	2.159%	0.0107
	Bayes dml+tnorm	2.158%	0.0107
	GPLDA	3.12%	0.0156

As one can see in Table 1, the cosine score combined with the LDA200 standard can achieve the best results, and GPLDA performs the worst. However, Bayesian dml performs better than the cosine score if the score is not normalized. Compared to the state-of-the-art ASR performance of the integrated cosine score standard, the gap with Bayesian_dml is very small. In addition, there is almost no advantage in the normalization of scores in Bayesian_dml.

By understanding the differences between the Cosine Score combined criteria and Bayesian_dml, I compare their performance to the different combinations of technical preprocessing. Pretreatment techniques include LDA and NCA, which are applied before the scoring model. The comparison of cosine scores, Bayesian_dml of NIST 2008 SRE different normalization is shown in Table 2.

Table 2. Comparison of cosine scores, Bayesian_dml of NIST 2008 SRE different normalization

Combination Norms		EER	minDCF
LDA 200		1.790%	0.0097
	NCA150+LDA150	43.345%	0.0987
	NCA200	2.345%	0.0131
	NCA200+LDA100	2.017%	0.0097
	NCA200+LDA200	1.767%	0.0095
	NCA200+LDA600	41.078%	0.0197
Bayes_dml	NCA200	4.567%	0.0278
	LDA 200	2.176%	0.0107
	LDA200+NCA150+LDA150	42.345%	0.1005
	LDA200+NCA200	3.034%	0.179
	LDA200+NCA200+LDA100	1.775%	0.0097
	LDA200+NCA200+LDA200	1.817%	0.1100
	LDA600+NCA200+LDA200	43.786%	0.1001
	NCA200+LDA200	3.564%	0.0178

Table 2 gives us an idea of how NCA and LDA represent hidden structures in the total variation space. The worst performance appeared in the second and sixth lines. In both cases, the size of the NCA is different from the size of the previous LDA. That is to say, the NCA acts to reduce the size and has a serious impact on the results. In the fourth row, the NCA 200 only performs one rotation following the LDA 200,

and the LDA 100 later further reduces the size of the feature space, which has little effect on performance. With the exception of the second and fourth rows, the results in the seventh row are almost at the same level as the other rows, although the size of the NCA 200 is reduced on the 600-dimensional modified and supervised. The reason may be that this size reduction is done in the original total variability space, while the second and fourth line quota reductions are made in the reduced function space after the LDA. Improvements in the fourth row indicate that the LDA projection corrects the functional space indication learned from the NCA. Therefore, I can conclude that NCA did not play a role in reducing the dimension.

The best performance of the combined cosine score is obtained using LDA200 + NCA200 + LDA200, and the best performance of Bayes_dml was obtained using LDA200 + NCA200 + LDA100. Bayesian_dml overcomes the combined norm of cosine scores and is the best result of the short2-short3 condition report for NIST SRE 2008 data.

With a specific end goal to check the performance of LDA200 + NCA200 + LDA200 based ASR, processed the real match scores with LDA200 + NCA200 + LDA100 with the impostor match scores. The Detection Error Trade-off (DET) of LDA200 + NCA200 + LDA200 performance with NIST 2008 SRE dataset an EER value of about 1.767% and with LDA200 + NCA200 + LDA100 the EER of about 1.776% is shown in Figure 1.

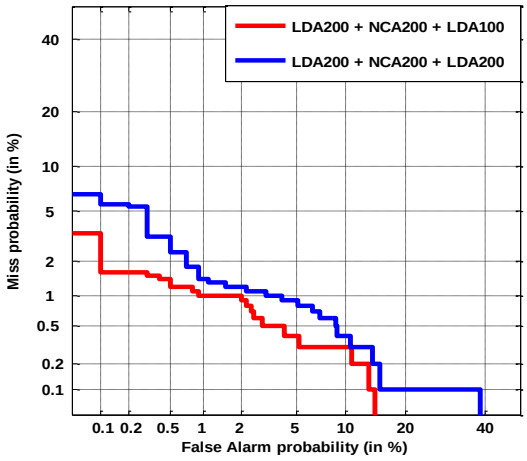

Figure 1. Equal error rate of LDA200 + NCA200 + LDA200 and Bayes_dml using LDA200 + NCA200 + LDA100 based ASR system

5. CONCLUSION

This paper presents the ASR application with Bayesian Distance Learning Metric technique. The traditional i-vector has been modified by catenating the label vector, linear regression matrix, mean super vector and modified and supervised factor loading matrix. The modified and supervised i-vetor in proposed ASR has a very high degree of discriminatively in order to regularize the speaker specific label information to enhance the performance of ASR.

The proposed Bayesian Distance Learning Metric technique in ASR can be utilized as an integrated to enhance the efficiency and outperform the traditional i-vector. The proposed ASR system has achieved the best performance of its EER as 1.767% using LDA200 + NCA200 + LDA200 with combined cosine score normalization and best performance of Bayes_dml was obtained using LDA200 + NCA200 + LDA100 of its EER at 1.776%.

REFERENCES

[1] S. Singh, "Forensic and Automatic Speaker Recognition System," *International Journal of Electrical and Computer Engineering*, vol/issue: 8(5), pp. 2804-2811, 2018.
[2] S. Singh. "High Level Speaker Specific Features as an Efficiency Enhancing Parameters in Speaker Recognition System," *International Journal of Electrical and Computer Engineering*, vol/issue: 9(4), 2019.
[3] S. Singh, "The Role of Speech Technology in Biometrics, Forensics and Man-Machine Interface," *International Journal of Electrical and Computer Engineering*, vol/issue: 9(1), pp. 281-288, 2019.

[4] S. Singh, et al., "Short Duration Voice Data Speaker Recognition System Using Novel Fuzzy Vector Quantization Algorithms," *2016 IEEE International Instrumentation and Measurement Technology Conference*, Taipei, Taiwan pp. 1-6, 2016.
[5] P. Belin, et al., "Voice-selective areas in human auditory cortex," *Nature*, vol. 403, pp. 309-312, 2000.
[6] S. Singh, "Evaluation of Sparsification algorithm and Its Application in Speaker Recognition System," *International Journal of Applied Engineering Research*, vol/issue: 13(17), pp. 13015-13021, 2018.
[7] E. Formisano, et al., "Who' is saying 'what'? Brainbased decoding of human voice and speech," *Science*, vol. 322, pp. 970-973, 2008.
[8] D. A. Reynolds, et al., "Speaker verification usingvadapted Gaussian mixture models," *Digital Signal Process*, vol/issue: 10(1-3), pp. 19-41, 2000.
[9] M. Zissman, "Language identification using phoneme recognition and phonotactic language modeling," *Proc. ICASSP*, pp. 3503-3506, 1995.
[10] S. Singh, "Support Vector Machine Based Approaches For Real Time Automatic Speaker Recognition System," *International Journal of Applied Engineering Research*, vol/issue: 13(10), pp. 8561-8567, 2018.
[11] S. Singh, et al., "Speaker Specific Phone Sequence and Support Vector Machines Telephonic Based Speaker Recognition System," *International Journal of Applied Engineering Research*, vol/issue: 12(19), pp. 8026-8033, 2017.
[12] H. Li, et al., "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE 101*, pp. 1136-1159, 2013.
[13] P. T. Carrasquillo, et al., "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," *Proc. ICSLP*, pp. 89-92, 2002.
[14] S. M. Siniscalchi, et al., "Exploiting context-dependency and acoustic resolution of universal speech attribute models in spoken language recognition," *Proc. INTERSPEECH*, pp. 2718-2721, 2010.
[15] G. Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition," *The shared views of four research groups. IEEE Signal Processing Magazine*, vol. 29, pp. 82-97, 2012.
[16] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1060-1089, 2013.
[17] S. Singh, "Speaker Recognition by Gaussian Filter Based Feature Extraction and Proposed Fuzzy Vector Quantization Modeling Technique," *International Journal of Applied Engineering Research*, vol/issue: 13(16), pp. 12798-12804, 2018.
[18] P. Kenny, et al., "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 980-988, 2008.
[19] W. Campbell, et al., "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308-311, 2006.
[20] L. Burget, et al., "Discriminative training techniques for acoustic language identification," *Proc. ICASSP*, 2006.
[21] S. Singh, et al., "Short Duration Voice Data Speaker Recognition System Using Novel Fuzzy Vector Quantization Algorithms," *IEEE International Instrumentation and Measurement Technology Conference*, 2016.
[22] S. Singh, et al., "Efficient Modelling Technique based Speaker Recognition under Limited Speech Data," *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, vol/issue: 8(11), pp. 41-48, 2016.
[23] D. Martinez, et al., "Language recognition in ivectors space," *Proc. INTERSPEECH*, pp. 861-864, 2011.
[24] S. Singh, et al., "A Novel Algorithm of Sparse Representations for Speech Compression/Enhancement and Its Application in Speaker Recognition System," *International Journal of Computational and Applied Mathematics*, vol/issue: 11(1), pp. 89-104, 2016.
[25] S. Singh and A. Singh, "Accuracy Comparison using Different Modeling Techniques under Limited Speech Data of Speaker Recognition Systems," *Global Journal of Science Frontier Research: F Mathematics and Decision Sciences*, vol/issue: 16(2), pp. 1-17, 2016.