# Using the conformal embedding analysis to compensate the channel effect in the i-vector based speaker verification system

Z. Boulkenafet[1], M.Bengherabi[1], O.Nouali[2], M.Cheriet[3]

Centre de Développement des Technologies Avancées (CDTA) Algeria[1]
Centre de Recherche sur lÍnformation Scientifique et Technique (CERIST) Algeria[2]
Ecole de Technologie Supérieure ETS, Canada
(zboulkenafet, mbengherabi)@cdta.dz[1]
onouali@cerist.dz[2]
mohamed.cheriet@etsmtl.ca[3]

**Abstract:** The I-vector approach to speaker recognition has become the prevalent paradigm over the past 2 years, showing top performance in NIST evaluations. This success is due mainly to the capability of the I-vector to capture and compress the speaker characteristics at low dimension and the subsequent channel compensation techniques that minimize channel variability. The Linear Discriminative Analysis (LDA) followed by Within-Class Covariance Normalization (WCCN ) and Cosine Similarity Scoring (CSS) represents the best compromise between performance and computational complexity. In this paper, we propose to use Conformal Embedding Analysis (CEA ); a recently proposed manifold leaning technique; to tackle the main limitations of LDA which are: the Gaussian assumption on the classes distribution, the inability to preserve the local geometric relationships of the data-space and its reliance on the Euclidean distance for characterizing the relationships between feature vectors. Experimental results on the challenging MOBIO-voice database show that CEA+WCCN outperforms LDA+WCCN for both male and female speakers at all operating points.

## 1 Introduction

The Gaussian Mixture Model-Universal Background Model (GMM-UBM) framework [RQD00] forms the core of the state-of-the art speaker recognition systems. Starting from the Joint Factor Analysis (JFA), proposed by Kenny et al [KBOD04] to model jointly speaker and session components to the recently proposed total variability paradigm dubbed I- vector [DKD$^+$11] . The experiences performed by Dehak et al [DKD$^+$11] demonstrated that the space of channel component in the JFA method contains information which can be used to discriminate between speakers. Consequently, they proposed a new representation called Identity vector or Intermediate vector (I-vector), in which the speaker and the channel sub-spaces are represented by a single total variability space. Unlike the JFA method which models the channel variability during the training stage, I-vector takes into account

the channel compensation during the scoring stage.

In the original I-vector system [DKD+11], the channel compensation is done by using Linear Discriminant Analysis (LDA) projection followed by Within-Class Covariance Normalization (WCCN). The motivation for using LDA is to maximize the inter-speakers variability and minimize the intra-speaker variability, which is an important point in the speaker recognition. The LDA projection measures the euclidean distance between the input vectors and assumes that each class vectors have a Gaussian distribution. In general, this approach suppose that the testing data drawn from the same underlying distribution as the training data. Unfortunately, it is usually hard to guarantee this assumption. Therefore, recent studies reveal that the local features and intrinsic geometric structures [SR03] [WH00] in the input data can further improve the discriminative power. Such techniques suppose that the targeted space is a sub-manifold of low dimension embedded in a high dimensional ambient space.

The representative non-linear manifold learning such as Local Linear Embedding (LLE) [RS00], Isometric feature mapping (ISOMAP) [TSL00], Laplacian Eigenmaps (LE) [BN03], etc. aims to map data into a low dimensional manifold which preserves the local topological structure of neighbors connections. These embedding methods are designed to describe a fixed set of data and not to generalize to novel data (test data). To cover the new data some techniques suggest to use the linear approximation of the non-linear method such as Locality Preserving Projections (LPP) [HYH+05], Locally Embedded Analysis (LEA) [FH05] and Neighborhood Preserving Embedding (NPE) [HCYZ05]. In these techniques the projection from a high dimensional space to a low dimensional space is described by a transformation matrix instead of using a nonlinear mapping method defined on the training set. Hence, it is easy to apply the transformation to unseen data. However, these transformations focus on preserving data localities and similarities so the discrimination between classes can not be sufficiently guaranteed. To deal with this problem, some methods such as Local Discriminant Embedding(LDE) [CCL05] and Locality Sensitive Discriminant Analysis (LSDA) [CHZ+07] propose to use Fisher criterion and Kernel transformation to boost the discrimination power. Furthermore, and starting from the fact that the Euclidean metric is incapable of capturing the intrinsic similarities. Some recent researches have suggested to use the cosine distance for better discrimination and robustness [FLH07].

In this work, we propose the use of Conformal Embedding Analysis (CEA ) [FLH07]; a recently proposed manifold leaning technique; as an alternative of LDA. The main motivations for using this dimensionality reduction technique are: 1) The CEA has no assumption about the distribution of the input data. 2) The CEA preserves the local geometric relationships of the data-space and increase the inter-class discrimination using the cosine distance for characterizing the relationships between feature vectors projected on a unit sphere. Knowing that the cosine similarity scoring is found to be the most appropriate for I-vector, we could expect that CEA will boost the system performance.

The rest of this paper is organized as follows. In Section 2 we present the I-vector system. The CEA approach is described in Section 3. Section 4 is dedicated to experimental results while the main conclusions and possible research perspectives will be presented in section 5.

## 2 Overview of the I-vector system

Inspired by the earlier use of JFA speaker factors directly as features for SVM classification, Dehak et al [DKD+11] have recently proposed a new approach of speaker modeling called I-vector. Unlike the JFA, the I-vector method represents the speaker and the channel subspaces with a single total variability space (equation 1). This representation is motivated by the fact that the channel space of JFA system contains information which can be used to discriminate between speakers.

$$m(s) = M + Tw(s) \tag{1}$$

where $M$ is the UBM supervector (a supervector is constructed by concatenating all the mean vectors of the GMM model), $w$ is a latent variable with a standard normal distribution and $T$ is a low rank variability matrix

As the total variability space represented by $T$ contains both speaker and channel information, the I-vector method requires additional techniques to attenuate the effect of session variability. Using the Linear Discriminant Analysis (LDA) followed by the Within Class Covariance Normalization (WCNN) in [DKD+11] ; has given the best performances.

### 2.1 Scoring

In i-vector system, we use a simple method of similarity measure, which calculates the Cosine Similarity Score (CSS) between the enrollment speaker I-vector ($w_E$) and the test I-vector ($w_T$). With the use of the LDA and WCCN projection matrices ($A$ and $S$ respectively) the cosine similarity will be given by:

$$score(w_E, w_T) = \frac{(A^t w_E)^t}{\sqrt{(A^t w_E)^t S^{-1} (A^t w_E)}} S^{-1} \frac{(A^t w_T)}{\sqrt{(A^t w_T)^t S^{-1} (A^t w_T)}} \tag{2}$$

The use of the angle between the two vectors make this scoring method more robust to the channel and the session effects.

## 3 Conformal Embedding Analysis CEA

Given a data set of I-vectors $X = [x_1, x_2, ...., x_n]$ with a high dimensionality and the corresponding class labels $L = [l_1, l_2, ...., l_n]$, where each $x_i$ belongs to a class $l_i$. The CEA objectives are:

1. Preserve the same-class conformal affinity while keeping away the diff-class conformal affinity after the embedding.
2. If two original high dimensional i-vectors are close (large conformal affinity), then the embedded low-dimensional points are close as well.

3. The embedded sub-manifold can better reflect the class relations with respect to the labeling information.

To achieve these objectives the computation of the CEA projection matrix $P$ passes by the following steps:

- Scale each i-vector $x_i$ to be a norm-one:

$$x_i = \frac{x_i}{||x_i||} \tag{3}$$

- Construct intrinsic graph $G_s$ and penalty graph $G_d$ both with $n$ nodes (each node corresponds to an i-vector). For $G_s$, we only consider each pair of data $x_i$ and $x_j$ from the same class ($l_i = l_j$). An edge is constructed between nodes $i$ and $j$ if $x_j$ is among the $k_s$ largest conformal neighbors of $x_i$ and vice versa. For $G_p$, we only consider each pair of data $x_i$ and $x_j$ from different classes ($l_i \neq l_j$). An edge is constructed between nodes $i$ and $j$ if $x_j$ is among the $k_d$ largest conformal neighbors of $x_i$ and vice versa.

- Define the conformal affinity weight matrices $W_s$ and $W_d$ for $G_s$ and $G_d$, respectively. If the two nodes $i$ and $j$ are connected then the weight of the edge between $i$ and $j$ is set by: $w_{i,j} = w_1$, Otherwise, if $i$ and $j$ are not connected the weight $w_{i,j} = w_2$.

  In [FLH07], the authors present tree types of weight: Balanced Rigid Weights, Unbalanced Soft Weights and Balanced Soft Weights. The last one is used in our study and it is described by:

$$w_{i,j} = \begin{cases} exp((cos(\theta_{(i,j)}) - 1)/t)) & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0 & \text{Otherwise} \end{cases} \tag{4}$$

  where $t$ is a constant and $\theta_{(i,j)}$ is the angle between the vectors i and j

- Compute the CEA projection matrix $P = [p_1, p_2, ..., p_N]$ by finding the $N$ eigenvectors corresponding to the $N$ largest eigenvalues of the matrix:

$$B = ((X(D_i - W_i)X^t))^{-1}(X(D_p - W_p)X^t). \tag{5}$$

  where $D$ is a diagonal matrix: $D(i,i) = \sum_j w_{i,j}$.

## 4 Experiments

In this section, we give a brief description of the database and the associated benchmarking protocol, followed by the feature extraction module. Finally, we present and discuss the obtained results.

### 4.1 Database description

The MOBIO database [MMH+12] contains audiovisual recordings of 152 people (100 males and 52 females) from five European countries. These recordings were registered through two phases, each one consists of 6 sessions. In the first phase data, the speakers were asked to answer a set of 21 questions with the question types ranging from: Short Response Questions, Short Response Free Speech, Set Speech, and Free Speech wherein the second phase data the speakers were asked to answer a set 11 questions with the question types ranging from: Short Response Questions, Set Speech, and Free Speech.

The MOBIO database was recorded using two mobile devices: NOKIA N93i mobile phone and MacBook (2008) laptop computer. Since this database was acquired with mobile devices, it had a significant amount of noise [SCP+10]. About 10% of utterances had an SNR less than 5 dB, while 60% had SNR between 5 to 10 dB. Also the utterances had limited amount of speech. About 25% of utterances had less than 2 seconds of speech while 35% had between 2 to 3 seconds of speech.

The MOBIO database was partitioned in three sets: the training set, the development set and the evaluation set. the training set utterances were used to learn the background parameters such as the UBM model and the subspace matrices. The development set data are used to tune the meta-parameters such as number of Gaussians and the subspace dimensions, this data set is divided into two parts, the enrollment set which is used to generate the client models (5 utterances for each speaker) and the probe set which is used in scoring. The evaluation set has the same structure as the the development and it is used for compute the final evaluation performance.

### 4.2 Experimental Setup

In our experiments, we use 19 MFCC coefficients extracted using 20 ms Hamming window taken every 10 ms. These features were augmented with the log energy, delta and double delta coefficients to produce 60 dimensional feature vector. To reduce the channel effect, we apply the Cepstral Mean Subtraction (CMS) normalization [Ata74] to the features. We eliminated the no speech segments using the voice activity detection (VAD) algorithm described in [RSB+04].

In our case, the UBM is gender independent. It was trained using 100 utterances from each speaker, and it is composed with 256 Gaussian components with diagonal covariance matrice. For the i-vector experiments, the dimension of the total variability subspace is 400 while the dimensions of the CEA projection matrix is 200. In the Balanced Soft Weights described in section 4, the constant $t = 0.1$ yields the best performance.

System performance is assessed using both equal error rate (EER) of the development set and the half total error rate (HTER) of the evaluation set. The EER corresponds to the point defined by some threshold $\theta$, where false acceptance rate (FAR) is equal to false rejection rate (FRR). HTER is the mean of FAR and FRR of the evaluation set, at the threshold $\theta$ previously tuned in the development set. Further more we plot DET curves which allow

Table 1: EER and HTER of CEA+WCCN and LDA+WCCN compensation techniques on MOBIO database.

| | Male | | Female | |
|---|---|---|---|---|
| Method | EER | HTER | EER | HTER |
| CEA+WCCN | 10.992% | 20.651% | 12.474% | 25.298% |
| LDA+WCCN | 12.337% | 20.982% | 14.402% | 28.614% |

the comparison of many systems at different operating points.

### 4.3 Results and discussion

In this section we present the results of the verification system using the two techniques of channel compensation: LDA +WCCN and CEA+ WCCN. The performances are given by the EER and HTER values (table 1) and represented by the DET curves (figures 1 and 2).

We notice from table1 that in the development set the EER value of the male gender decreases from 12.337% with the LDA+ WCCN technique to 10.992% with the CEA+WCCN technique. Thus, the CEA+WCCN yields a relative error reduction rate of 10.90% compared to the baseline LDA+WCCN. For female speakers, the reduction is more pronounced ( 13.38% ), where the EER decreases from 14.402% with LDA+WCCN to 12.474 % with CEA+WCCN. The same in the evaluation set the HTER of the male gender decreases from 20.982% with LDA+WCCN technique to 20.651% with CEA+WCN technique (reduction of 1.57%) and for the female gender the HTER decreases from 28.614% with LDA+WCCN to 25.298% with CEA+WCCN (reduction of 11.58%).

The comparison of the two systems at different operating points are illustrated through the DET curves in figures 1 and 2 for the male and female cases respectively. We notice that the CEA+WCCN outperforms the LDA+WCCN at all operating points.

## 5 Conclusion

In this work we investigated the use of the CEA sub-manifold learning method to compensate the channel variability in the i-vectors based speaker verification systems. Unlike LDA, CEA has no assumption about the distribution of the input data and it uses both Conformal embedding nature and discriminating criterion to compute the projection matrix. Experiments on the MOBIO-voice database shows that the CEA+WCCN technique performs better than the LDA+WCCN technique. Future work includes further experimental analysis on other databases to confirm obtained results.

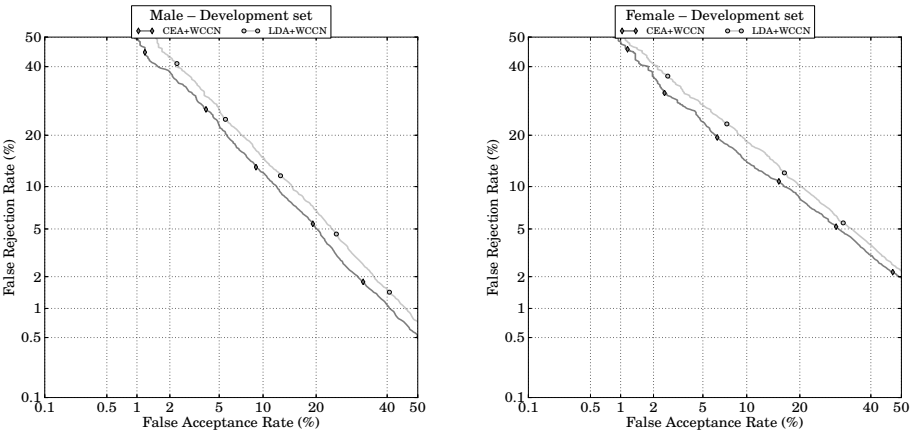


Figure 1: DET curves of CEA+WCCN and LDA+WCCN techniques of male and female for the development set.

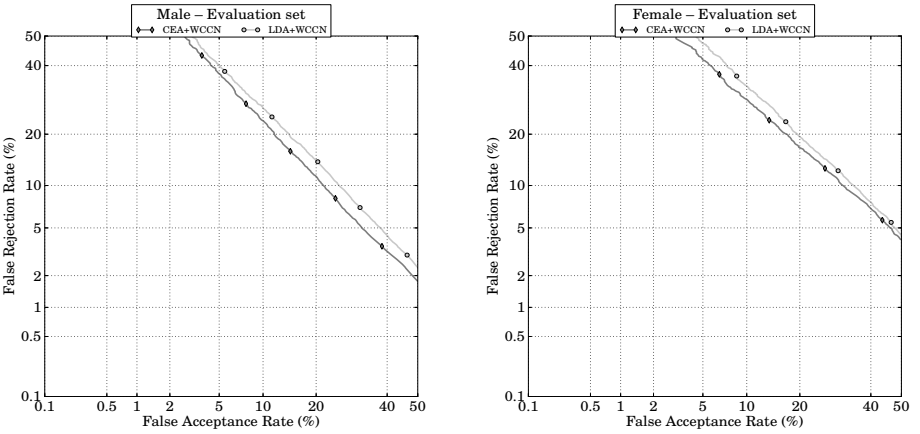


Figure 2: DET curves of CEA+WCCN and LDA+WCCN techniques of male and female for the evaluation set.

# References

[Ata74] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974.

[BN03] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, 2003.

[CCL05] H. T. Chen, H. W. Chang, and T. L. Liu. Local discriminant embedding and its variants. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer*

*Society Conference on*, volume 2, pages 846–853 vol. 2, 2005.

[CHZ+07] D. Cai, X. F. He, K. Zhou, J. W. Han, and H. J. Bao. Locality Sensitive Discriminant Analysis. In *IJCAI conf*, 2007.

[DKD+11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front end factor anlysis for speaker verification. *IEEE Trans. Audio, Speech, and Languge Processing*, 16(4):788–798, 2011.

[FH05] Y. Fu and T. S. Huang. Locally Linear Embedded Eigenspace. *IEEE Trans. on PAMI*, 27(3):328–340, 2005.

[FLH07] Y. Fu, M. Liu, and T. S. Huang. Conformal Embedding Analysis with Local Graph Modeling on the Unit Hypersphere. In *CVPR*, 2007.

[HCYZ05] X. F. He, D. Cai, S. C. Yan, and H. J. Zhang. Neighborhood Preserving Embedding. In *ICCV conf*, pages 208–1213, 2005.

[HYH+05] X. F. He, S.C. Yan, Y.X. Hu, P. Niyogi, and H.-J. Zhang. Face Recognition Using Laplacianfaces. *IEEE Trans. on PAMI*, 27(3):328–340, 2005.

[KBOD04] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Speaker adaptation using an eigenphone basis. *IEEE Trans on Speech Audio Processing*, 12(6):579 – 589, 2004.

[MMH+12] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J. F. Bonastre, P. Tresadern, and T. Cootes. Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data. In *IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, July 2012.

[RQD00] D. Reynolds, T. Quatieri, and R. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[RS00] S.T. Roweis and L.K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[RSB+04] J. Ramirez, J. Segura, C. Benitez, A. Torre, and A. Rubio. Eficient voice activity detection algorithm using long-term speech information. *speech communication*, 42(3-4):271â287, 2004.

[SCP+10] Marcel Sébastien, McCool Chris, Matějka Pavel, Ahonen Timo, Černocký Jan, Chakraborty Shayok, Balasubramanian Vineeth, Panchanathan Sethuraman, Chan Chi Ho, Kittler Josef, Poh Norman, Fauve Benoit, Glembek Ondřej, Plchot Oldřich, Jančík Zdeněk, Larcher Anthony, Lévy Christophe, Matrouf Driss, Bonastre Jean-Francois, Lee Ping-Han, Hung Jui-Yu, Wu Si-Wei, Hung Yi-Ping, and Machlica Lukáš. On the Results of the First Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation. *Lecture Notes in Computer Science*, Neuveden:210–225, 2010.

[SR03] L.K. Saul and S.T. Roweis. Think Globally, Fit Locally:Unsupervised Learning of Low Dimensional Manifolds. *Machine Learning Research*, 4:119–155, 2003.

[TSL00] J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[WH00] A. Weingessel and K. Hornik. Local PCA Algorithms. *IEEE Trans. Neural Networks*, 11(6):1242–1250, 2000.