

# Speaker verification using probabilistic 2D CLAFIC

Chong Lee Ying<sup>a)</sup> and Andrew BJ Teoh<sup>b)</sup>

Faculty of Information Science and Technology, Multimedia University,  
Jalan Ayer Keroh, 75450 Melaka, Malaysia

a) [lychong@mmu.edu.my](mailto:lychong@mmu.edu.my)

b) [bjteoh@mmu.edu.my](mailto:bjteoh@mmu.edu.my)

**Abstract:** User-specific subspace method, such as CLAFIC concatenates the speech matrices side by side leads to a high dimensional feature space before forming the correlation matrix. We propose an enhanced CLAFIC method in speaker verification, coined as Probabilistic 2D CLAFIC which apply a straightforward 2D speech matrix projection technique for feature dimension reduction and GMM is used to construct the speaker model to boost up the performance. Our results show the superior performance and low computation time in the proposed system.

**Keywords:** Speaker verification, 2D CLAFIC, Gaussian Mixture Model

**Classification:** Science and engineering for electronics

## References

- [1] D. A. Reynolds, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- [2] S. Watanabe and N. Pakvasa, "Subspace Methods of Pattern Recognition," *Proc. 1st Int. Conf. on Pattern Recognition*, pp. 25–32, 1973.
- [3] Y. Ariki and N. Ishikawa, "Integration of Face and Speaker Recognition by Subspace Method," *Proc. ICPR96*, pp. 456–460, 1996.
- [4] M. B. Gülmezoglu, V. Dzhafarov, M. Keskin, and A. Barkana, "A Novel Approach to Isolated Word recognition," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 620–628, 1999.
- [5] M. A. El-Gamal, M. F. Abu El-Yazeed, and M. M. H. El Ayadi, "Dimensionality Reduction For Text-Independent Speaker Identification Using Gaussian Mixture Model," *Proc. IEEE Int. Midwest Symp. Circuits Syst.*, pp. 625–628, 2003.
- [6] W. Zhang, Y. Yang, Z. Wu, and L. Sang, "Experimental Evaluation of a New Speaker Identification Framework using PCA," *IEEE Int. Conf. Syst., Man, Cybern.*, vol. 5, pp. 4147–4152, 2003.
- [7] A. Higgins, "YOHO speaker verification," *Speech Research Symp.*, 1990.

## 1 Introduction

Biometric authentication for speaker verification is based on individual's voice

to determine an identity claimed. High dimension feature set is required for construction of speaker model. However, the presence of parameters which contain irrelevant information and improper choice of speech features have led to the loss of its discrimination power.

The subspace method such as CLAFIC emphasises that every class owns its subspace which represents the distinct spectral characteristic for each speaker. The conspicuous feature is extracted to construct a low dimension speaker subspace for each speaker, thus solving the problem induced by high dimension feature set.

CLAFIC concatenates the speech matrices side by side before forming the correlation matrix. The resulting matrix lead to a high dimensional feature space where it is difficult to evaluate the correlation matrix accurately due to its huge size compared to small order of speech feature. This causes the computation of eigenvalues and eigenvectors to be problematic and numerically unstable.

Hence, Probabilistic 2D CLAFIC is proposed to overcome the drawback of the subspace method. As opposed to CLAFIC, Probabilistic 2D CLAFIC is based on the average of 2D matrices rather than long concatenation of speech feature matrices. That is, a correlation matrix can be constructed directly using the original speech feature matrices. Thus, the discrimination capability is improved and computational complexity is reduced. Gaussian Mixture Model (GMM) [1] is used to replace the conventional CLAFIC matching method and boost up the performance of the speaker verification.

## 2 Related Works

Several subspace methods have been used to construct the speaker subspace, the earliest subspace method was CLAss-Featuring Information Compression (CLAFIC) proposed by Watanabe et al. [2]. This method employs the Karhunen-Loeve Transform (KLT) to generate the basis vectors for subspace construction. The modified CLAFIC which was proposed by Y. Ariki [3] translates the subspace origin to the centroid of feature space to increase its discrimination power.

Common Vector Approach (CVA) was introduced by Gülmezoğlu et al. [4] to estimate different subspaces which represent the feature caused by intra-speaker difference. Principal Component Analysis (PCA) [5, 6] is a conventional dimensional reduction technique which extracts the eigenvectors from the covariance matrix of the feature set. The high dimensional features are orthogonally projected into the low dimensional subspace.

CLAFIC is different from PCA. PCA constructs a low dimensional subspace representing feature space of the *entire class*. This causes the dimension of entire feature spaces to be too large to cover all the discriminatory features of each class. In contrast, every class in CLAFIC forms a low dimensional subspace distincts from the subspaces spanned by the other classes. Thus, the performance of speaker model by using the entire feature space might be less efficient than the speaker model employing the subspace method for

individual class.

### 3 Probabilistic 2D CALFIC

The proposed method consists of 3 stages: (a) Subspace formulation (b) Probabilistic modeling (c) Probabilistic matching. In the subspace formulation stage, the 2D individual's speech feature matrix,  $X$  is retained after the feature extraction. Let each individual holds  $H$  training speech samples, the  $i$ th training sample is denoted by  $n \times m$  matrix where  $n$  represents the order of feature coefficient and  $m$  represents the number of frames,  $X_i \in \mathbb{R}^{n \times m}$ , where  $i = 1, \dots, H$ . The construction of 2D individual's speech feature matrix,  $X$  is then extended to include  $C$  speakers,  $\{X_i^j \in \mathbb{R}^{n \times m} \mid i = 1, \dots, H, j = 1, \dots, C\}$  where  $i$  refers to number of training samples for each speaker and  $j$  refers to number of speaker. The class mean matrix of  $j$ th speaker is denoted by  $\bar{X}^j$ . The class correlation matrix  $\{\Phi_{2D}^j \in \mathbb{R}^{n \times n} \mid j = 1, \dots, C\}$  is computed for  $j$ th class as,

$$\Phi_{2D}^j = \frac{1}{H} \sum_{i=1}^H (X_i^j - \bar{X}^j) (X_i^j - \bar{X}^j)^T. \quad (1)$$

We use Singular Value Decomposition (SVD) for eigenvalue decomposition to decompose class correlation matrix defined by,  $\Phi_{2D}^j = U_j \Sigma V_j^T$  where matrix  $U_j = \{u_1^j, \dots, u_n^j\}$  is an orthogonal matrix consists of a set of eigenvectors  $u_n$  corresponding to its eigenvalues. By selecting the first  $k$  basis vectors from matrix  $U_j$ , the matrix  $\Omega_j = \{u_1^j, \dots, u_k^j\}$  with size of  $n \times k$  is formed.

In probabilistic modeling stage, a set of training samples  $X^j = \{x_1, \dots, x_H\}$  is projected onto the corresponding matrix  $\Omega_j$  expressed by,  $Y^j = \Omega_j^T X^j$  where  $Y^j = \{y_1, \dots, y_H\}$  is a new set of low dimension projected training samples. This set of new training samples is fed into GMM to construct a speaker model.

During verification, a test sample  $X_{test}$  is transformed by a claimed matrix  $\Omega_j$  to produce a low dimension test sample,  $Y = \Omega_j^T X_{test}$ . This new projected test sample is input into GMM for probabilistic matching. A likelihood ratio test is used to produce a match score. If the match score is larger than the decision threshold, the claimed speaker is accepted as a true user. Otherwise, the system will reject the claimed speaker.

### 4 Database and Experimental Setting

The experiments are conducted by using YOHO speech corpus [7]. The training set is composed of (138x5) samples: 5 speech samples per speaker are randomly chosen. The test set consists of (138x8) samples: 8 speech samples per speaker are randomly picked. There is no overlapping between the two sets.

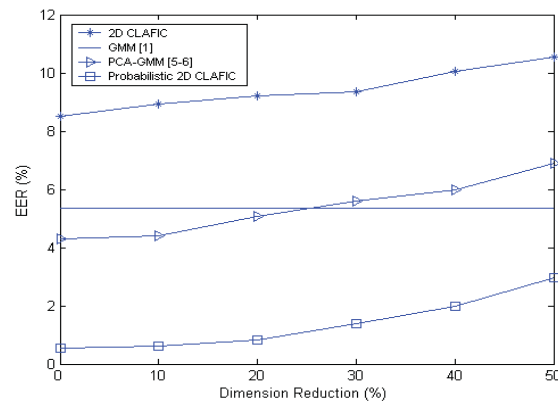
The speech signal is blocked into 240 sample frames with 160 overlapping with adjacent frames. The linear predictive (LP) cepstrum with Hamming Window is used to extract the speech feature matrix. The dimension of

speech feature set is  $n \times m$ , where  $n$  is Linear Prediction Coefficients (LPC) order which is fixed to 20 and  $m$  refers to the number of frames.

## 5 Experimental Results

Experiment is carried out with the dimension of feature set reduced steadily from  $n = 20$  to  $n = 10$  to construct the speaker subspace. Speaker model is fixed to 20 Gaussians mixture order.

From Fig. 1, it is clearly showed that the Probabilistic 2D CLAFIC obtains the best verification rate among all the tested methods. It performs well with Equal Error Rate (EER) = 2.98% in the 50% dimension reduction as compared to the full feature dimension for GMM (EER = 5.36%) and PCA-GMM (EER = 4.29%). This indicates that Probabilistic 2D CLAFIC preserves its high performance even with 50% feature reduction. Besides, it attains lowest error rate with EER = 0.56% in full feature dimension.



**Fig. 1.** Performance comparison of several subspace methods with different feature dimensions

The experiment results show that Probabilistic 2D CLAFIC outperforms PCA-GMM with maximum difference of EER = 5.46%. This indicates that Probabilistic 2D CLAFIC, which represents the separate individual subspace is having high discrimination capability as compared to GMM-PCA which operates in common subspace.

The poor performance of 2D CLAFIC is due to the conventional CLAFIC matching method which does not contribute much to the accuracy. The PCA-GMM performs better than 2D CLAFIC because its discrimination capability is boosted up by GMM.

Table I shows the comparison of training time and matrix size among GMM [1], PCA-GMM [5, 6] and Probabilistic 2D CLAFIC. GMM is the baseline system with feature dimension  $n = 20$ . The dimension of feature set  $n$  is reduced from 20 to 10 for PCA-GMM and Probabilistic 2D CLAFIC. The total number of training speech samples is 690. The LPC order for each training speech sample is fixed at  $n = 20$ . Thus, the size of each training sample is  $20 \times m$ , where  $m$  is the number of frames different from each speaker.

**Table I.** Comparison of matrix size and training time for several subspace methods using GMM as speaker modeling (CPU: Pentium IV 3.00GHz, RAM: 1 GB)

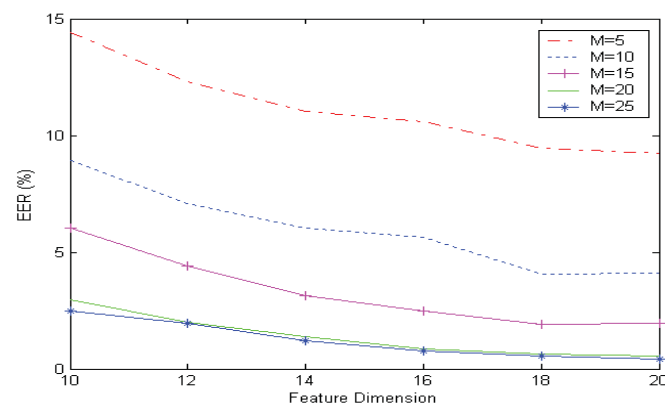
	GMM [1]	PCA-GMM [5, 6]	Probabilistic 2D CLAFIC
Feature dimension	20	10	10
EER (%)	5.36	6.91	2.98
Training time (sec)	1944.6	1055.5	1126.6*
Matrix size	—	690 x 690	20 x 20

\* The training time of 1126.6 sec involves 138 speakers

The PCA-GMM involves the calculation of the eigenvectors of a 690 x 690 matrix for common subspace. Whereas, the Probabilistic 2D CLAFIC involves calculating the eigenvectors of a 20 x 20 matrix for individual speaker subspace. It is showed that the Probabilistic 2D CLAFIC is more efficient than the PCA-GMM because the matrix size of Probabilistic 2D CLAFIC is much smaller. Both PCA-GMM and Probabilistic 2D CLAFIC exhibit shorter training time compared to GMM. This is due to the feature reduction from  $n = 20$  to  $n = 10$  for both methods.

The experiment in Table I involves the training of 138 speakers. Thus, the time for Probabilistic 2D CLAFIC to construct the matrix and find the eigenvectors for 138 speakers is taken separately. However, the practical employment involves only one speaker in each enrollment. The Probabilistic 2D CLAFIC trains the subspace for one speaker during each enrollment stage. Whereas, PCA-GMM involves the retraining of the entire common subspace every time when a new user registers into the system in enrollment stage. Thus, Probabilistic 2D CLAFIC is superior to PCA-GMM in term of computational efficiency and training time for practical employment. This is due to the Probabilistic 2D CLAFIC that does not involve the retraining of the entire subspace and thus having much shorter training time than PCA-GMM.

The variation of Gaussians mixture order when the feature dimension is reduced from 20 to 10 for Probabilistic 2D CLAFIC is illustrated in Fig. 2.



**Fig. 2.** Dimension reduction of Probabilistic 2D CLAFIC upon different value of Gaussians mixture order

Gaussians mixture order ranges from 5 to 25 are used and  $M$  indicates the Gaussians mixture order. The performance increased when larger value of Gaussians mixture order is used. However, the increasing performance is levelled off when the Gaussians mixture order is within 20 to 10.

## 6 Conclusion

In this paper, the Probabilistic 2D CLAFIC has been shown to outperform all state of the arts speaker verification when the best EER = 0.56% with full dimension feature and EER = 2.98% with 50% of reduction of feature dimension is attained. This is because GMM is used as the probabilistic modeling that boosts up the discrimination capability. The experimental results show the effectiveness of the proposed method for improving the verification performance of the text-independent speaker verification.