Identification of Affection in Voice Using an Enhanced Supervised Locally Linear Embedding Technique

Cui Yeqin, Gao Jianguo

College of Mathematics and Information Science
Langfang Normal University
Langfang 065000, China

***Abstract* — To enhance the identification of affection in voice, it is necessary to conduct nonlinear dimensionality reduction for voice attribute samples lying on a nonlinear manifold embedded in high-dimensional voice space. Supervised locally linear embedding is an outstanding supervised manifold learning method for nonlinear dimensionality reduction. Taking this technique existing drawbacks into account, this paper proposes an enhanced version of supervised locally linear embedding method, which enhances the identification ability of low-dimensional embedded samples and to provide the best generalization power. The proposed method can be used to conduct nonlinear dimensionality reduction for 36-dimensional voice affection attribute samples including prosody and speech character attributes, and draws low-dimensional embedded identification attributes so as to identify four affections including indignation, happiness, sadness and neutrality. Experimental results on natural voice affection database show that the proposed method achieves a high accuracy of 91.56% with only less than 8 embedded attributes, acquiring 14.35% enhancement over the proposed method. Thus, the method can greatly enhance voice affection identification while being applied for reducing dimensionality of voice affection attribute samples.**

Keywords - Voice Affection Identification; Nonlinear Dimensionality Reduction; Manifold Learning; Supervised Locally Linear Embedding

I. INTRODUCTION

Emotional calculating can be viewed as super-computers with the affection ability and survey ability like persons, which comes into being many types of emotional features to usable computers, affected each other [1]. Since voice, one of the most important communication medium to human, carries sufficient affection information from speakers, it makes the research, which is about analysis and extracting affective characteristics in voice signals so that computers can identify the affective state of speakers automatically, especially important, and it has a great application values in the neighborhood human-computer interaction [2], call center [3], Intelligent robot [4], etc.

Currently, files have been discovered that attribute samples of voice data can be on the non-linear high-dimensional embedding in speech characteristic distance [5]. This makes manifold learning methods aiming at searching intrinsic topology structure in high-dimensional samples, which is utilized in non-linear manifold learning for voice attribute samples, for instance low dimensional voice [6] and voice identification [7]. The focus purpose of dimension reduction is acquiring the principal attribute for lower dimensional inset. The two important manifold learning methods for nonlinear dimensionality are local linear embedding (LLE) [8] and Isometric Mapping (Isomap) [9]. Although these two algorithms can availably achieve low-dimensional visualization for voice samples, they are ineffective for voice identification, and even worse than traditional principal component analysis (PCA) [10]. The reason for this is that these two manifold learning algorithms are in the unsupervised way for dimensionality reduction, which means they don't consider the class information among the existing samples points which is helpful in classifying.

To put up with the drawback of unsupervised manifold learning methods in identification, Reference [11] presents an enhanced supervised locally linear embedding (SLLE) method by utilizing the space, which takes the classification labels in samples into accounts, modifying the neighboring samples to seek in the LLE. The SLLE method can be extensively utilized in kinds of areas such as face feature identification [12] and face identification. It has a superior pattern identification property. Nevertheless, there are still three inefficiencies in this algorithm: (1) The supervision distance that SLLE is using is linear, which will make the inner-class distance in the desirable samples points grow in step with class separation distance, and it'll weaken the discriminating power of the low-dimensional embedded samples generated from SLLE, which makes it not beneficial to samples classification; (2) SLLE runs the existing training sample samples in batch, however, it not able to solve the generalization for new test sample samples, because the input from low-dimensional discriminating samples SLLE extracts from the existing training sample samples to new test sample samples can't give a reasonable embedding output directly, which is called the lack of generalization ability; (3) The constant factors that constitute the supervision distance in SLLE have an important influence on SLLE's ability of generalization.

However, with how to optimize constant factors from the distance of SLLE, the existing studies mostly implement fussy and repetitive artificial search test on a specific target dimension to get the constant factor's optimum value, and fix the value to use dimensional reduction in every different target dimension. In fact, this constant factor's optimum value can be easily influenced by different target dimension for dimensional reduction. So, the constant factor's optimum value that is extracted from a target dimension is not the best to other target dimension for dimension reduction.

We present the **main contributions** as follows in this paper:

We study the drawback of unsupervised manifold learning methods and the weakness of SLLE method for identifying voice affection.

We present a non-linear space that could improve a different ability of lower dimensional embedding samples to take the place of the linear space, which solves the weakness of SLLE method for identifying voice affection.

We design an automatic algorithm which majorizes the invariable parameter adaptively in kinds of dimensionality to design an enhanced SLLE that has good capability. We utilize the Enhanced-SLLE to make non-linear dimension for voice affection attribute factor lying on a high dimension distance and acquiring lower dimension embedding attribute with enhanced significant ability to enhance voice affection identification results in lower dimension embedding attribute space.

The conducted experiments on the natural voice affection database demonstrate that the proposed algorithm obtains the highest accuracy of 91.56% with only less 8 embedded attributes, making 14.35% enhancement over SLLE algorithm. Therefore, the proposed algorithm can significantly enhance voice affection identification results while applied for reducing dimensionality of voice affection attribute samples.

The rest of the paper is concluded in the following part. Section 2 presents our proposed Enhanced-SLLE algorithm. Section 3 gives the results of experiments. Finally, we conclude the paper in Section 4.

II. Enhanced-SLLE

Set N input samples data of D dimensionality equivalence to X_i ($X_i \in R^D$, $i \in [1, N]$), type number equivalence to L_i , embedding output dimensionality $(d, d \le D)$. N samples data equivalence to $Y_i\left(Y_i \in R^d, i \in \left[1, N\right]\right)$.

Enhanced-SLLE has three steps:

Discover each samples data $x_{ij} = x_j^{\min} + rand(0,1)\left(x_j^{\max} - x_j^{\min}\right)$ neighboring data by computing the non-linear supervised space between samples data.

Compute the reconstruction W matrix by the samples data neighboring data. Whereas, the optimal reconstruction W matrix p needs to minimize function:

$$p \tag{1}$$

It requires p; while p is not p's neighborhood point, f_1=0.

Compute the samples data values by local reconstruction W matrix and neighboring data. Computing the embedding samples data on lower dimension distance to minimize lower dimension reconstruction error, which can minimize cost function：

$$f_2 \; f_3 \tag{2}$$

where f_4 and f_5 contains attribute factor.

In the Step(1) of Enhanced-SLLE, the non-linear supervised space formal can be as follows:

$$\Delta^{'} = \begin{cases} \sqrt{1 - e^{-\Delta^2/\beta}}, & L_i = L_j \\ \sqrt{e^{\Delta^2/\beta}} - \alpha, & L_i \ne L_j \end{cases} \tag{3}$$

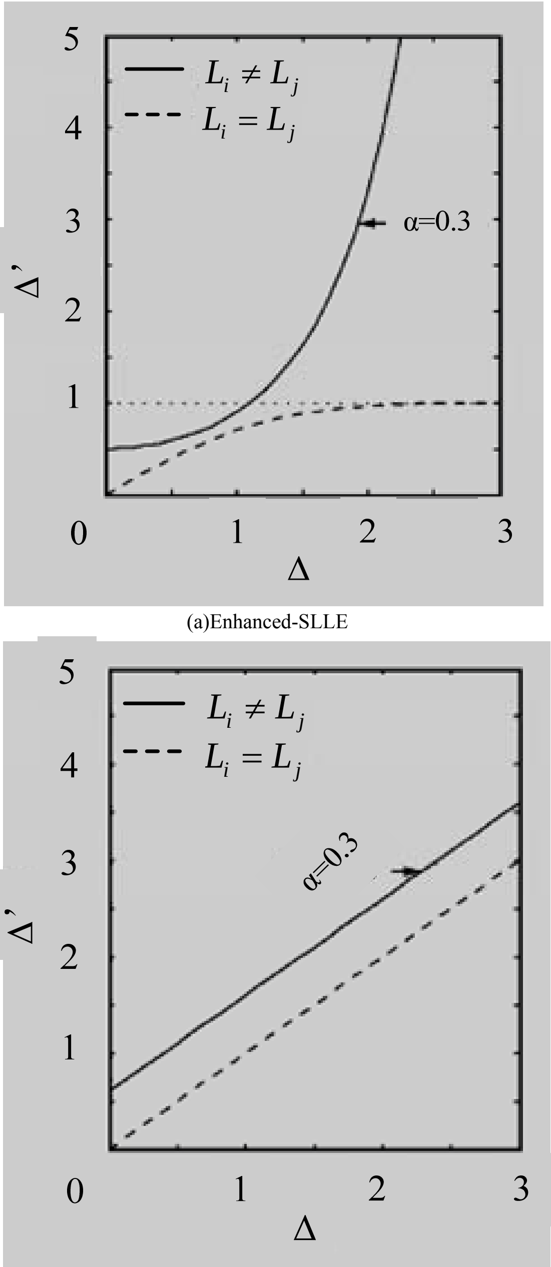
where f_7 denotes the space computed by samples data class label; whereas, f_8 denotes the initiate Euclidean space. f_9 denotes f_{10} of exponent function for increasing excessively rapid

Compared to Enhanced-SLLE, the linear supervised distance formula that the original SLLE calculate the distance between point and point is:

$$\Delta^{'} = \begin{cases} \Delta, & L_i = L_j \\ \Delta + \alpha \cdot \max\left(\Delta\right), & L_i \ne L_j \end{cases} \tag{4}$$

where max(Δ) stands for maximum Euclidean distance while $\alpha\left(0 \le \alpha \le 1\right)$ is used to control the degree of quantity combination for samples point class information from distance calculating.

In order to gain a better understanding of the superiority for the nonlinear supervised distance in Enhanced-SLLE, figure 1 made an example. It compared the different change characteristic of distance curve between Enhanced-SLLE and SLLE. The original Euclidean distance, is increasing linearly in specified interval [0,3]. Both of them set the α equals to 0.3.



(a)Enhanced-SLLE

(b) SLLE

Figure 1. Comparison of two algorithms' distance curve.

From Fig. 1(a) we can see that Enhanced-SLLE use nonlinear supervised distance, so that the class separation distance of different class samples points is growing rapidly exponentially and inner-class distance of same class samples points is in a slow-growth in the interval [0,1]. This make the specific value of class separation distance and inner-class distance grow in step with the growth of distance, which make the growth of discriminating power of low-dimensional embedded samples also go with the growth of distance, and this is beneficial to analysis for embedded samples. In the opposite, SLLE don't have this benefit because of its linear supervised distance. As the Fig. 1(b) shows, while the inner-class distance is growing rapidly, the class separation distance grows in step with it, so that the specific value of class separation distance and inner-class distance remain unchanged. And it weakens the discriminating power of lower dimension embedding samples, which is bad for the samples classification.

To acquire the better power for Enhanced-SLLE, the proposed method utilizes Nystrom to design the algorithm of Enhanced-SLLE. To enhance availably the property on voice affection identification, it is necessary to conduct nonlinear dimensionality reduction for voice attribute samples lying on a nonlinear manifold embedded in high-dimensional voice space. Supervised locally linear embedding is an outstanding supervised manifold learning method for nonlinear dimensionality reduction. Taking its existing drawbacks into account, this paper proposes an enhanced version of supervised locally linear embedding method, which enhances the identification ability of low-dimensional embedded samples and is provided with the best generalization power. The proposed method can be used to conduct nonlinear dimensionality reduction for 36-dimensional voice affection attribute samples including prosody and speech character attributes, and draws low-dimensional embedded identification attributes so as to identify four affections including indignation, happy, broken-hearted and neutral. The method can greatly enhance voice affection identification results while being applied for reducing dimensionality of voice affection attribute samples.

The description of Algorithm 1 can be as follows:

Step 1: Input: training samples X , maximum number of dimensionality $d_{\max}$ ($d_{\max} \leq D$), initiate $\alpha = 0$, the initiate value $error-rate = 1$, initiate dimensionality for dimension reduction $d = 2$, optimal number α equivalence to α .

Step 2: Dividing samples X : 50% for training, training samples equivalence to X_1 ; 50% for testing, test samples equivalence to X_2 .

Step 3: Conduct the following:

```
for $d$ to $d_{\max}$ , $d = 2,3,\cdots,\mathrm{d}_{\max}$
for $\alpha$ to 1, $\alpha = 0,0.1,0.2\cdots,1$
(a) Using Enhanced-SLLE to compute $X_1$ 's $d-D$ embedding samples $X_{1d}$ ;
(b) Using the algorithm of Enhanced-SLLE to compute $X_{2d}$ ;
(c) Using K-Nearest-Neighbor [15] (K=1) to train $X_{1d}$ and test $X_{2d}$ ; thus, compute the identification $error-rate$ of $d-D$ embedding samples $new-error$ and compare $new-error$ and $error-rate$ .
        if $new-error < error-rate$
            $error-rate = new-error$ ;
            $best\alpha = \alpha$ ;
    (d) optimized $\alpha$ corresponding to each $d-D$ and $error-rate$
            end for
end for.
```

III. VOICE AFFECTION IDENTIFICATION EXPERIMENT RESULTS

To evaluate the voice affection identification ability of the proposed algorithm, Enhanced-SLLE, PCA, LLE, Isomap and SLLE will be utilized to conduct dimension reduction for voice attribute samples; then compare voice affection identification results of the 5 methods with different dimensionality.

IV. NATURAL AFFECTION VOICE DATABASE

Nowadays domestic and foreign researchers mostly do the voice affection identification research by artificial simulate affection voice database. The naturalness of voice affection in simulate database, however, is still not able to be compared with that in real affection, which makes it is doubted. So the research about real affection voice identification in men's real lives seems to be more practical and meaningful. This article has built a Chinese voice database that contains the natural affection of speaker for that. We build a Chinese affection database with 800 sentences, which is irrelevant to the speaker and with high-naturalness, by video backgrounds of 20 TV interview programs. There are at least two person talks casually about typical social phenomenon, family conflict or moving stories in each program, and they don't have previous lecture while they are talking about the topic. They just do an off-the-cuff discussion. So the expression not obvious of these people while they are discussing is very real. Because of the restrain of topic range, the number of the video that contain the affection type, such as indignation, happiness, broken-hearted and neutral is large. Use the professional Adobe Audition CoolEdit (http://www. mp3-converter. com/cool_edit_2000. htm) to extract the audio file from the whole video, delete the affection express of characters in it and video clips with much noise, then cut out a whole affection voice of different person in a short time from an ideal video clips to use as affection analysis corpus. We finally have extracted 800-sentences WAV database with sampling rate of 16 kHz, single track of 16 bit from 53 persons (female 37, male 16). Then asked 4 audiences to test randomly and do the delete and re-extracting to the sentences with less obvious affection attribute.

V. EXTRACT VOICE AFFECTION ATTRIBUTE PARAMETER

Nowadays, researchers found that the affection attribute parameters related to the pronunciation of speakers mainly include the rhythm attributes like fundamental frequency, amplitude (or energy) and time of pronunciation, the voice quality like formant, and the distribution of spectrum energy and HNR. So this article has extracted 48 rhythm attributes and voice quality attribute parameters from each voice in natural affection voice database as Table 1 shows.

TABLE I. VOICE AFFECTION ATTRIBUTE PARAMETERS

Attribute type	Attribute group	Gene frequency
Rhythm attribute	Gene frequency	(1) $f_{\max}$; (2) $f_{\min}$; (3) $f_d = f_{\max} - f_{\min}$; (4) Upper quartile: $f_{0.75}$; (5) Median: $f_{0.5}$; (6) Lower quartile: $f_{0.25}$; (7) $f_i = f_{0.75} - f_{0.25}$; (8) Average value: f_m ; (9) Standard deviation: σ_ρ ; (10) The average absolute pitch: M_β ;
	Amplitude	(11) A_m ; (12) σ_ρ ; (13) $A_{\max}$; (14) $A_{\min}$; (15) $A_d = A_{\max -} A_{\min}$; (16) $A_{0.75}$; (17) $A_{0.5}$; (18) $A_{0.25}$; (19) A_i ;
	Pronunciation continuous time	T_s ; T_v ; T_u ; $T_{vu} = T_v / \mathrm{T}_u$; $T_{vs} = T_v / \mathrm{T}_s$; T_{us}
	Formant F_1 - F_3	Average value F_1 ; F_2 ; F_3 ; standard deviation： F_1 ; F_2 ; F_3 ； median： F_1 ; F_2 ; F_3
Voice attribute	Frequency band energy distribution	0-500hz: SED_{500} ;50-1000Hz: SED_{1000} ;2500-4000Hz: SED_{4000} ; 4000-5000Hz: SED_{5000}
	HNR	(1) $H_{\max}$;(2) $H_{\min}$ (3) $H_d = H_{\max} - H_{\min}$ (4) H_{m} (5) σ_H
	Short-term jitter parameters	Jitter; Shimmer

VI. EXPERIMENT TEST AND ANALYSIS OF RESULT

Experiment 1: Don't conduct any dimensionality reduction on the extracted original 48D voice attribute samples, and does the affection identification experiment directly. The identification results are shown in Table 2.

TABLE II. DESCRIPTION RESULTS

Affection type	Indignation	Happiness	Broken-hearted	Neutral
Indignation	166	24	10	0
Happiness	36	140	20	4
Broken-hearted	16	40	123	21
Neutral	0	20	8	172

We can know that the identification results about indignation and neutral are more satisfactory from table 3, and the correct identification rate is up to 83% and 86%. The average correct identification rate of 4 kinds of affection is 75.13%. While the correct identification rate of happiness and broken-hearted is lower, in which the one about happiness is 70% and that about broken-hearted is 61.5%. The main reason lead to this is that the rhythm attributes of happiness are similar to that of indignation and the voice quality attributes of broken-hearted are similar to that of indignation. So that we mix the two paired affection (happiness and broken-hearted, broken-hearted and indignation) up easily.

Experiment 2: Conduct PCA, LLE, Isomap, SLLE and Enhanced-SLLE. The related generation algorithm on the dimensionality reduction for 48D voice attribute samples, and does the affection identification test to the after-dimensionality reduction low-dimensional discriminating attribute samples, then compare identification results. The generation algorithm of LLE, Isomap, and SLLE is similar to Enhanced-SLLE, while that in PCA can get directly by multiplying linear mapping matrix from training sample and new test sample. The target dimension range for dimensionality reduction is $2 \le d \le 20$. It conducts better while the neighborhood number of LLE, Isomap, SLLE and Enhanced-SLLE $k = 12$. Use automatic optimization algorithm to get the optimized value of constant factor α corresponding to Enhanced-SLLE and SLLE in each dimension in every cross check experiment. From table 3 and 4, we can know that the optimized average value of Enhanced-SLLE's constant factor α in different dimension is very stable that its value is generally less than 0.5. But the optimized average value of SLLE's constant factor α changes obviously. It's because that there's a parameter β existing in nonlinear supervised distance that Enhanced-SLLE is using. β's value equals to the average value of all paired samples points' Euclidean distance, so that β has equilibrium activity on the change of α.

Table 1 gives us the every dimension's voice identification results got by five kinds of dimensionality reduction ways. Table 5 lists the comparison of the best property gained by various ways in different dimension, in which the "Original" stands for the identification results gained without conducting dimensionality reduction on attribute samples of original 36D (as Table 2 shows)

In addition, SLLE, after the generalization and he optimization of constant factor α, gained a identification property better than PCA, LLE and ISOMP, as a supervised dimensionality reduction way, SLLE is better than unsupervised PCA, LLE and Isomap. While use the un-generalized SLLE to do the affection identification, the results of identification is bad that only can be up to22.45%. Furthermore, compared to LLE and Isomap, PCA get a better property for identification, which means the voice attribute samples lying on the nonlinear distance is not that high. And it makes linear PCA can still extract the low-dimensional embedded samples with better discriminating power than that of nonlinear LLE and Isomap. Besides, Isomap conducts better than LLE. Isomap is a global dimensionality reduction method, which remains the global structure information while embedding. LLE is a partial dimensionality reduction method, which remains the partial structure information while embedding. And it shows that to remain the global structure information is more effective than to remain partial structure information. In addition, identification property of all the algorithms grow obviously step with dimension at beginning, but while the dimension gets higher, its property will be falling and finally be tending towards stability .This is because the intrinsic dimension of inner-structure information embedded in 48D voice attribute space samples is exactly in the dimension range [2,20], which shows the dimension range [2,20] in dimensionality reduction is reasonable.

TABLE III. OPTIMIZED AVERAGE VALUE OF ENHANCED-SLLE'S CONSTANT FACTOR α IN DIFFERENT DIMENSION

d	2	3	4	5	6	7	8	9	10	11
α	0.16	0.27	0.33	0.42	0.25	0.36	0.21	0.59	0.12	0.24
Error rate %	52.95	48.18	44.32	32.23	35.68	33.86	31.23	30.85	27.50	28.06
D	12	13	14	15	16	17	18	19	20	
α	0.26	0.23	0.14	0.38	0.22	0.12	0.36	0.41	0.23	
Error rate %	29.32	30.36	32.59	34.32	35.45	35.77	36.52	36.09	37.82	

TABLE IV. OPTIMIZED AVERAGE VALUE OF SLLE'S CONSTANT FACTOR α IN DIFFERENT DIMENSION

D	2	3	4	5	6	7	8	9	10	11
α	0.62	0.35	0.41	0.72	0.44	0.76	0.53	0.35	0.78	0.24
Error rate %	54.09	46.04	43.86	38.63	35.82	34.14	33.50	29.54	30.90	28.61
D	12	13	14	15	16	17	18	19	20	--
α	0.55	0.14	0.28	0.67	0.21	0.37	0.45	0.62	0.42	--
Error rate %	31.80	32.90	34.09	38.60	40.90	38.33	39.71	38.60	42.04	--

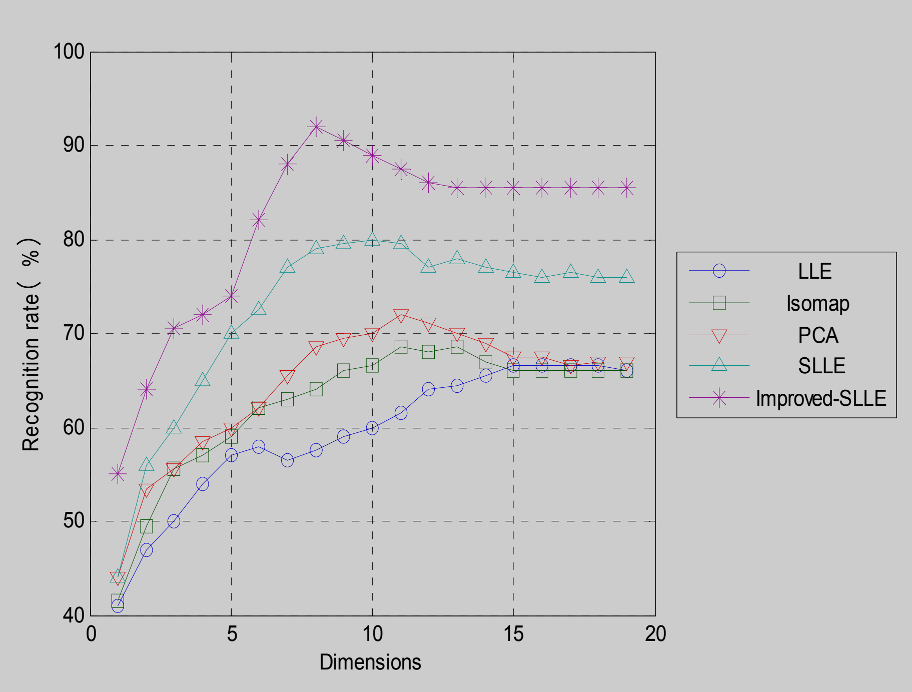


Figure 2. Voice affection identification result for different dimensions

VII. Conclusions

This paper proposes an enhanced supervised partial linear embedding algorithm called Enhanced-SLLE. It extracts low-dimensional embedded discriminating attribute to do voice affection identification and got 91.56% correct identification rate by Enhanced-SLLE. Compared to other algorithms, Enhanced-SLLE has best identification property. The research of voice affection identification is now still in primary stage. To develop a manifold learning algorithm better than Enhanced-SLLE is significant for further research of voice affection identification.

Acknowledgements

This work is supported by Project of Natural Science Foundation, Heilongjiang Province, China (No. F200804); Project of Science Technology, Heilongjiang Province, China (No. 11531056).

References

[1] Errity, A., and McKenna, J. "An investigation of manifold learning for speech analysis". The 9th International Conference on Spoken Language Processing (ICSLP'06), Pittsburgh, PA, USA, vol. 15, no. 3, pp: 2506-2509., 2006.

[2] Goddard, J., Schlotthauer, G., and Torres, M., et al. "Dimensionality reduction for visualization of normal and pathological voice samples". Biomedical Signal Processing and Control, vol. 4, no. 3, pp: 194-201, 2009.

[3] Jolliffe, I. T.. "Principal component analysis". New York: Springer, vol. 12, no. 5, pp: 150-165, 2002.

[4] Jones, C., and Deeming, A.. "Affective human-robotic interaction; Affect and affection in Human-Computer Interaction", Springer, Lecture Notes in Computer Science, vol. 48, no. 5, pp: 175-185, 2013.

[5] Liang, D., Yang, J., and Zheng, Z., et al. "A facial expression identification system based on supervised locally linear embedding". Pattern Identification Letters, vol. 26, no. 22, pp: 2374-2389, 2012.

[6] Morrison, D. R. W., and De Silva, L. C. "Ensemble methods for spoken affection identification in call-canters". Voice Communication, vol. 49, no. 2, pp: 98-112., 2011.

[7] Picard, R. 'Affective computing". MIT Press, Cambridge, MA, vol. 20, no. 5, pp: 1-24., 1997.

[8] Picard, R.. "Robots with affection intelligence. The 4th ACM/IEEE International Conference on Human Robot Interaction", California, vol. 32, no. 10, pp: 98-112, 2009.

[9] Ridder, D. D., Kouropteva, O., and Okun, O., et al. "Supervised locally linear embedding. Artificial Neural Networks and Neural Information Processing-ICANN/ICONIP"-2013, Springer, Lecture Notes in Computer Science, vol. 2714, no. 10, pp: 333-341, 2013.

[10] Roweis, S. T., and Saul, L. K.. “Nonlinear dimensionality reduction by locally linear embedding”. Science, vol. 290, no. 5500, pp: 2323-2326, 2000.

[11] Tenenbaum, J. B., “Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction”. Science, vol. 290, no. 5500, pp: 2319-2323, 2000.

[12] Yu, D.. “The application of manifold based visual voice units for visual voice identification”. [Ph.D.dissertation], Dublin City University, vol. 29, no. 55, pp: 23-28, 2008.