Reduced Feature Set for Emotion Based Spoken Utterances of Normal and Special Children Using Multivariate Analysis and Decision Trees

Maria Andleeb Siddiqui
Department of Software Engineering
NED University of Engineering and Technology
Karachi, Pakistan
mandleeb@neduet.edu.pk

Syed Abbas Ali
Department of Computer Science & Information Technology
NED University of Engineering and Technology, Karachi, Pakistan
saaj@neduet.edu.pk

Najmi Ghani Haider
Department of Software Engineering
NED University of Engineering and Technology
Karachi, Pakistan
najmi@neduet.edu.pk

***Abstract*—The current paper deals with the use of multivariate data analysis and decision tree methods in order to reduce the feature set for the normal and special children speech in four different emotions: anger, happiness, neutral and sadness. Ten features were extracted, by an algorithm implemented in a previous study to classify the speech emotions of normal and special children. In the current study, the best features are selected using multivariate analysis: principal component analysis (PCA), factor analysis and decision tree. Step by step PCA is applied to reduce the feature set according to the variables that are collinear. The obtained reduced feature sets are applicable to both normal and special children samples. Experimental results revealed that PCA yields the feature set comprising pitch, intensity, formant, LPCC and rate of acceleration. Factor analysis provides three feature sets out of which the feature set comprising of Rasta PLP, MFCC, ZCR and intensity provides the best result. Decision tree yields a feature set comprising energy, pitch and LPCC.**

Keywords-speech emotions; PCA; factor analysis; decision tree; features

I. INTRODUCTION

Emotion recognition system identifies the emotional state from voice [1], therefore it is called speech emotion recognition (SER). There are four modules of SER: input, feature extraction, feature selection and classification of emotions [2]. Prosodic features, particularly pitch, intensity and duration were used in early research studies. Currently, LLD's features such as shimmer, jitter, harmonic to noise ratio (HNR) and cepstrum have been used extensively [3, 4]. LPCC and MFCC were also accompanied in the speech feature set [5]. In [6], 40 depressed patients and 40 control subjects were used in a study for speech feature analysis. Characteristics of depressed patients were found using ANOVA analysis and the results were linked to Gaussian mixture model (GMM) and support vector machine (SVM). Autism spectrum disorder comorbid for children (ASD-CC) psychometric properties were evaluated and developed in [7]. Confirmatory factor analysis (CFA) is used for the factor structure of the Korean version of ASD-CC.

In [8], ten features were extracted: frequency, pitch, intensity, rate of acceleration, formant frequencies, log power, log energy, rate of zero passages, Mel frequency cepstrum coefficient (MFCC), linear prediction cepstrum coefficient (LPCC). The extraction of frequency starts with the speech signal loading and the conversion of analog signal into numeric data. After loading, maximum and minimum frequency were set and fast Fourier transform (FFT) of windowed signal was performed as shown in (1). Followed by the cepstrum calculation the frequency is extracted as shown in (2).

$$ms1 = fft(x.*hamming(length(x))) \quad (1)$$

$$f = fs/(ms1 + fx - 1) \quad (2)$$

where *ms1*=windowed signal, *fft*=fast Fourier transform, *f*=frequency, *x*=input signal, *fs*=sampling frequency, *fx*=input frequency.

For pitch extraction, signal acquisition and signal processing are the same as in frequency extraction. FFT is applied on the processed signal. Discrete Fourier transform (DFT) is taken of the FFT signal as in (3). Then the log of FFT is calculated. After that the real cepstrum is calculated and it is the absolute value of filtered log of DFT as shown in (4). Then finally real cepstrum pitch is extracted given in (5).

$$dfty = abs(fft(x)) \quad (3)$$

$$dft = log10(dfty) \quad (4)$$

$$rcp = real\_ceps(16 : length(real\_ceps)) \quad (5)$$

(*dft*=discrete fourier transform, *rcp*=real cepstrum pitch).

The magnitude of the DFT signal with $\log_{10}$ is filtered as in (6). The conversion of the magnitude into decibels and transpose result of decibel gives out intensity as shown in (7).

$$M = 20*log10(abs(Y(1 : length(x))) + eps) \quad (6)$$

$$i = mag2db(M) \quad (7)$$

where *M*=magnitude, *Y*=filtered signal, *x*=input signal, *i*=intensity, *mag2db*=magnitude to decibels.

Processing of the signal involves the time instant calculation as given in (8). The derivative of speed gives the velocity in (9). The gradient of velocity by 0.01 yields the acceleration in (10). Finally the average rate of acceleration is calculated by taking the mean of acceleration.

$$x1 = (0.4 * t.\hat{}\,4) + (10.8 * t.\hat{}\,3) - (64.4 * t.\hat{}\,2) - (28.2 * t) + 4.4 \quad (8)$$

$$v = diff(x1) \quad (9)$$

$$acc = gradient(v, 0.01) \quad (10)$$

where *x1*=speed, *T*=time instant, *v*=velocity, *diff*=derivative, *acc*= acceleration.

For extracting the formant frequency, preprocessing involves setting the number of coefficients according to the rule of thumb for formant estimation given in (11). After that calculation of the linear prediction coefficients is carried out in (12). Then the frequencies are calculated using (13). The imaginary part of the root gives out formant frequencies.

$$ncoeff = 2 + Fs / 1000 \quad (11)$$

$$a = lpc(ncoeff) \quad (12)$$

$$r = roots(a) \quad (13)$$

where *ncoeff*=number of coefficients, *Fs*=sampling frequency, *lpc*=linear prediction cepstrum, *a*=linear prediction cepstrum coefficient, *r*=roots. Preprocessing involves the setting of the sampling rate and sampling window size for log power extraction. After that the frame size is calculated and windowing is applied. Then the average energy is extracted. Average power is yielded by dividing the average energy with window size as shown in (16).

$$windowsize = sampling\_rate * fs \quad (14)$$

$$Average\_Energy = sum(result.\hat{}\,2) \quad (15)$$

$$Lp = Average\_Energy / windowsize \quad (16)$$

where, *fs*=sampling frequency, sum=sum of frame size of windowed signal, *Lp*=Log power.

Preprocessing involves the same procedure as log power extraction. Average energy is calculated by applying windowing on signal according to window size and frame size, given in (17).

$$Average\_Energy = sum(result.\hat{}\,2) \quad (17)$$

Signal acquisition and signal processing are the same as in pitch and frequency estimation given in (18).

$$RZP = sum(abs(sign(y1) - sign(y2))/2)/windowsize \quad (18)$$

where *RZP*=rate of zero passages, *y1*=maximum frequency, *y2*=minimum frequency.

Preprocessing involves setting and analysis of frame duration and frame shift. It is followed by the setting of the pre emphasis coefficient alpha=0.97, the number of filter bank channels M=20 and lower and upper frequency limits. Then the Hertz to Mel- wrapping function is calculated. Application of DCT matrix routine and the magnitude of the spectrum are calculated in (19). The filter bank is applied to the unique part of the magnitude spectrum. Finally the calculation of cepstral liftering gives MFCC (20).

$$dctm = @(N, M)(sqrt(2.0/M) * cos(repmat([0:N-1].', 1, M) * repmat(pi * ([1:M] - 0.5)/M, N, 1)) \quad (19)$$

$$CL = @(N, L)(1 + 0.5 * L * sin(pi * [0:N-1]/L)) \quad (20)$$

where *dctm*=secrete cosine transform matrix, *N*=number of coefficients, *M*=number of filter bank channels, *L*=length of channel, *CL*= cepstrum lifter.

Preprocessing involves the estimation exponent of next high power according to signal size. Then the number of prediction paths 'p' is set. Calculation of the number of linear prediction of coefficients is carried out. Fourier transform is applied on X-lpc according to the number of shifts N as given in (21). The logarithm of LPC is taken and then the LPC coefficients are converted back to spectra. The number of cepstra is then set and the first and second derivative of LPCC features are estimated to have the value of coefficients in (22).

$$[x\_lpc] = lpc(x, p) \quad (21)$$

$$lpcc = fft(x\_lpc, N) \quad (22)$$

where, *Lpc*=linear prediction coefficients, *x*=input signal, *p*= prediction paths, *FFT*= fast Fourier transform, *lpcc*=linear prediction cepstrum coefficients, *N*= number of shifts.

After extracting these features, speech emotion recognition of normal and special children (SERNSC) is implemented. To make the algorithm run efficiently, dimension reduction is a valuable approach. The advantage of dimensionality reduction is that it helps to discover the grouping of features that for sure run the algorithm with improved accuracy [9]. Detection of projection subspace basis evaluation is suggested in [10]. For deduction it uses generalized hyperbolic mixture (HMMDR) fit. This method is well accepted along with discriminant analysis, model based classification and clustering analysis. Two SDR techniques are demonstrated in [11]. The relationship between partial least square (PLS) and principal component regression (PCR) is explained. Dimensionality reduction by joining features is one of the best strategies proposed so far [12]. Sparse partial least square regression (SPLSR) is investigated in depth in [13]. It is revealed that the recognition rate of SPLSR is up to 79.23% and it is superior when compared to other methods used for dimensionality reduction. In this paper, feature reduction is presented using

multivariate analysis that consists of PCA, factor analysis and decision tree method.

II. METHODOLOGY

Seven features and three coding schemes are taken into consideration for analysis: pitch, intensity, formant, rate of acceleration, zero crossing rate (ZCR), log energy, log power, Mel frequency cepstrum coefficient (MFCC), linear prediction cepstrum coefficient (LPCC), Relative spectrum transforms perceptual linear prediction (Rasta PLP).

A. Classification Algorithm

The classification algorithm used is speech emotion recognition for normal and special children (SERNC) and the data set is the same as in [6]. In system flow, the thresholding of the extracted features is done before data file preparation. The process of data file preparation for each sample file is applied for decision making of the algorithm. It starts by loading the sample file speech.mat. After loading, the file name, file path categories and emotions were extracted by path. Then, the file name is added in speech.mat. In addition to this information, files containing information of emotions and categories are also added in speech.mat. Finally, the thresholding feature extraction is performed using (23). A function for decision making is implemented in (24).

Features=allfeatures_extraction(wav_file) (23)

Function [Category, Category_%, Emotion, Emotion_%] = hybrid_Decision_making(wav_file) (24)

The test samples were classified by using hybrid classification function and then finding the maximum category score position. The process of hybrid classification is described as follows

1) Step#1:

If Result of position n=1
Set category = Normal and Category_Percentage = Categories_output(1).
Else category = Special and Category_Percentage = Categories_output(2).

2) Step#2:

After the category is decided as normal or special, the emotion is labeled on the category according to the classification accuracy. The maximum score position is to be estimated in (25)

$$[rn, cn] = find\left(strcmp\left(Emotion\_o/p,\ max\ \left(Emotions\_o/p\right)\right)\right) \quad (25)$$

3) Step#3:

If Result of position n=1
Set Emotion = Angry and Emotion _Percentage = Emotions _output(1).
Else If Result of position n=2
Set Emotion = Happy and Emotion _Percentage= Emotions _output(2).
Else If Result of position n=3
Set Emotion = Neutral and Emotion_Percentage = Emotions _output(3).
Else Set Emotion = Sad and Emotion _Percentage = Emotions _output (4).
[End of If condition]

B. Multivariate Analysis

When working with correlated variables, multivariate analysis is very valuable. Multivariate analysis is used when the data set involves more than one variable. Two methods of multivariate analysis are used in this study: PCA and factor analysis. These two methods are performed on MINITAB. The flow diagram of PCA and factor analysis is shown in Figure 1 and Figure 2 respectively.

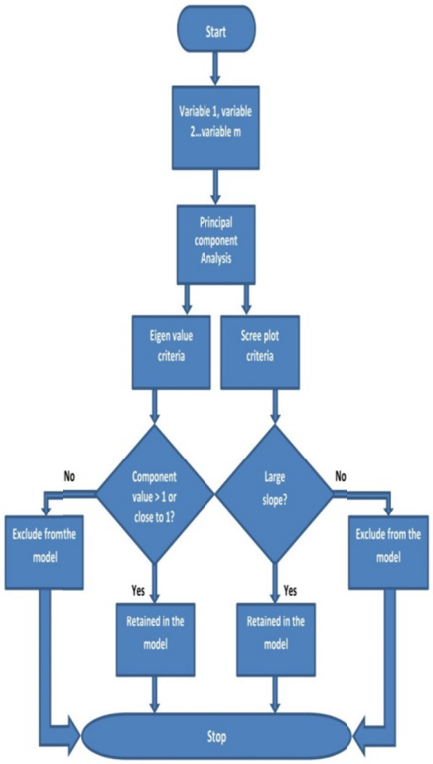

Fig. 1. PCA flow diagram

For PCA, the variables are the classification accuracy of speech features and coding schemes. Two methods are used for feature reduction: (a) Eigenvalue criteria and (b) Scree plot criteria. In (a) if the component value is greater than or close to 1 then the component (variable) is retained in the model, otherwise it is excluded from the model. In (b) if the slope is large then the component is retained in the model, otherwise it is excluded from the model. In factor analysis, the variables are the speech features and coding schemes. There are some underlying factors that have the effect on the communality of the variables. In this method two factors are extracted. There are two criteria: (a) communality criteria and (b) loading plot criteria. In (a) the communality of extracted factors for variables should be between 0.5 and 1 while in (b) the reduced feature sets based on the strong correlation between variables is provided.

Fig. 2. Factor analysis flow diagram

C. Decision Trees

In decision tree method a threshold is set for classification accuracy. The features on the two nodes (left and right) are set according to the threshold of classification accuracy. Higher classification accuracy features are retained in the model. The decision tree flow diagram is shown in Figure 3.

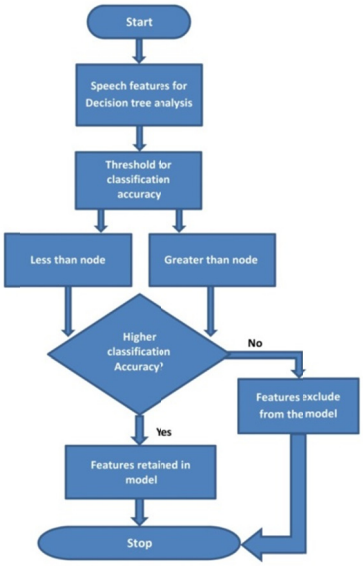

Fig. 3. Flow diagram of decision tree

III. Experimental Study

A. Experimental Design

The assumption is that these factors are collinear. PCA is used to reduce the number of factors that measure the accuracy of the speech emotion recognition. The number of components is the total considered factors. Factors are the same considered for PCA. The intersection of these factors is the focus in factor analysis. Likewise PCA, the goal of factorial analysis is to reduce the large number of variables into a smaller number of "factors". The effects of some underlying factors that can't be directly measured are assessed with the help of factorial analysis.

1) Experimental Results Based on PCA

The variance accounted for by a given component is represented by the eigenvalue of each component. The eigenvalue of the first component is usually greater than the others' as shown in Table I.

TABLE I. PCA STATISTICS

Factor/ Component	Eigenvalue	Percentage of variance (proportion)	Cumulative percentage
Pitch	3.6049	36	36%
Formant	2.2118	22.1	58.2%
Intensity	1.4581	14.6	72.7%
Rate of acceleration	1.4117	14.1	86.9%
LPCC	0.9818	9.8	96.7%
Rasta PLP	0.2159	2.2	98.8%
MFCC	0.1157	1.2	100%
ZCR	0.0000	0.0	100%
Energy	0.0000	0.0	100%
Power	0.0000	0.0	100%

- Eigenvalue Criteria

Eigenvalue is one of the most commonly used criteria for feature selection. Those components that have eigenvalue greater than 1 are retained in the model. If the successive variable difference of eigenvalue is so small then the eigenvalue will not serve as the best criterion for factor selection and the variable with eigenvalue 0.99 will be excluded from the model. Therefore the proportion and cumulative percentage of variance are also considered in the process of factor selection. Consider Table II, the first four factors have eigenvalues greater than 1 but the fifth component is very close to 1, it shows the 9.8% of the total variance and its cumulative percent of variance is 96.7%. So we can also include this variable into the model. According to the eigenvalue criteria the selected factors are pitch, intensity, formant, rate of acceleration (ROA) and LPCC.

- Scree Plot Criteria

Another method for factor selection is the scree plot. According to this criterion the factors having a big slope are on the cliff while the base of the cliff has the trivial factors. Consider Figure 4. It shows that the slope of the curve is quite

small from the starting of the sixth factor so all the factors after the sixth factor can be selected from the model.

2) Experimental Results Based on Factor Analysis

The factor loadings of each variable are shown in Table II.

TABLE II. FACTOR LOADINGS OF VARIABLES

Variable	Factor 1	Factor 2	Communality
Pitch	0.028	-0.726	0.528
Formant	-0.739	-0.104	0.556
Intensity	0.798	-0.223	0.686
ROA	-0.723	0.518	0.791
LPCC	0.102	-0.310	0.107
Rasta PLP	0.766	0.010	0.588
MFCC	0.687	-0.377	0.614
ZCR	0.882	-0.014	0.778
Energy	-0.382	0.592	0.497
Power	-0.331	-0.855	0.840
Variance	3.8089	2.1763	5.9852
% Var	0.381	0.218	0.599

The factor loadings for each variable are calculated. There are two extracted factors and the loading result tells that how much a variable is explained by the extracted factor. Loadings can be positive or negative. Large loading result means that the variable is strongly influenced by the factor whereas small loading result means that the variable is weakly influenced. "Variables" are the seven speech features and three coding schemes. "Factor 1" and "Factor 2" are the extracted factors by minitab that explains the 10 variables. The proportion of variables is explained by the extracted two factors and interpreted by the help of "communality" as shown in Table II. The assessment of the working of the model is calculated by the communality. In Table II, it is shown that the model works better for some variables. The model best explains power, and it is not bad for other variables. The model does not do a good job for LPCC and energy as they show only 10.7% and 49.7% communality which less than half variation of the variable. The loading plot of Factor analysis is shown in Figure 5.

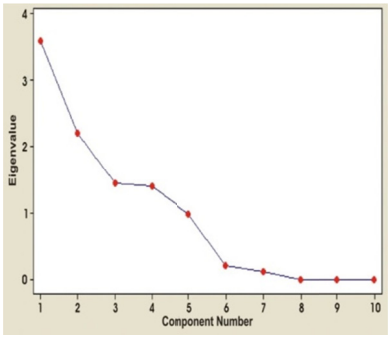

Fig. 4. Scree plot for factor reduction

Figure 7 reveals that:

- Rasta PLP, ZCR, intensity and MFCC lines are very close to each other and hence these factors are strongly correlated. So these four factors lead to the best classification accuracy.
- LPCC, pitch and power show strong correlation as the lines are in the same direction.
- ROA, energy and formant are also in the same direction and show strong correlation.

So these three sets of factors are closely correlated. The best correlation is seen in set 1 comprising rasta PLP, ZCR, intensity and MFCC. They are considered as a reduced feature set for higher classification accuracy.

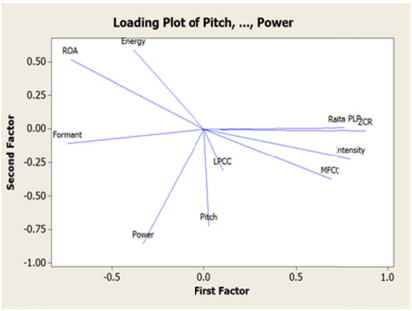

Fig. 5. Loading plot of factor analysis

3) Experimental Results based on Decision Trees

The factor analysis for reduction is done with the help of decision trees. The factor reduction is done with respect to the classification accuracy. The classification accuracy of the tree is shown in Figure 6. Energy, pitch and LPCC are the reduced feature in the node of the tree and hence they are formulating the reduced feature set.

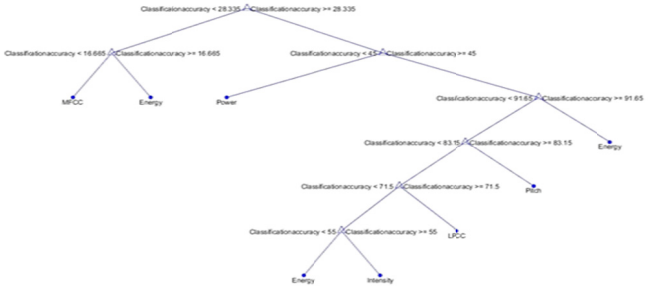

Fig. 6. Decision tree for factor reduction

The decision tree table is shown in Table III. It is shown that energy, pitch and LPCC show classification accuracy of 91.65%, 83.15% and 71.5%. The confusion matrix is shown in Figure 7. The top row shows the classification accuracy with true class energy. Predicted classes are shown in that column. The green cells show the true positive rate column. No other classification accuracy in the energy row is misclassified. The light green cells show the next higher similarity rate. Energy, pitch and LPCC have higher similarity rate in true and predicted class according to the accuracy in all emotions. The highest error rate is shown by dark red cells whereas the least error is shown by light red cells.

IV. CONCLUSION

In this study, the two techniques, multivariate analysis and decision tree were used to reduce the ten extracted features of emotion based spoken utterances for normal and special

children in four different emotions (anger, happiness, neutral and sadness). The reduced feature set and features are obtained from two multivariate techniques (PCA and factor analysis) and decision tree is applied on the emotion based spoken utterances in Urdu language for both normal and special children. The observations obtained from demonstrative experiments are: 1) PCA provides a reduced feature set comprising of pitch, intensity, formant, LPCC and ROA. 2) Factor analysis provides three sets of reduced features (rasta PLP, ZCR, intensity and MFCC), (energy, ROA and formant frequency) and (LPC, pitch and power) out of which the set comprising of rasta PLP, ZCR and MFCC shows better results. 3) Decision tree technique provides a reduced feature set comprising of energy, pitch and LPCC. Authors are focusing to develop an experiment to apply the reduced features and feature sets in order to analyze the classifier's accuracy.

TABLE III. DECISION TREE RESULTS

Node	Classification Accuracy Condition	Less Than Node	Greater Than Node
1	28.335>Accuracy>=28.335	Node 2	Node 3
2	16.6665>Accuracy>=16.6665	MFCC	ENERGY
3	45.00>Accuracy>=45.00	POWER	Node 4
4	91.65>Accuracy>=91.65	Node 5	ENERGY
5	83.15>Accuracy>=83.15	Node 6	PITCH
6	71.5>Accuracy>=71.5	Node 7	LPCC
7	55.00>Accuracy>=55.00	ENERGY	INTENSITY

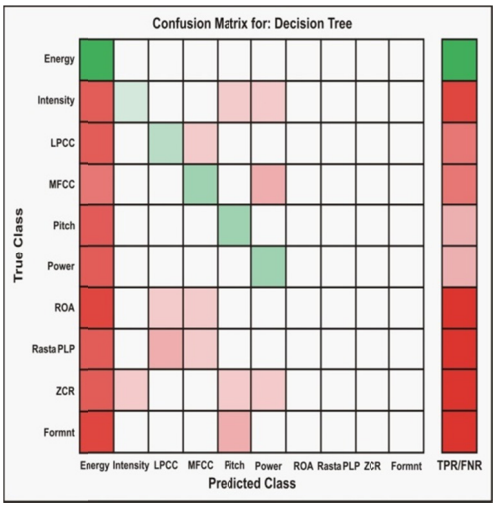

Fig. 7. Confusion matrix for decision tree factor reduction

REFERENCES

[1] S. Ramakrishnan, "Recognition of Emotion from Speech: A Review", in: Speech Enhancement, Modeling and Recognition- Algorithms and Applications, pp. 121-138, InTech, 2012

[2] S. Pahune, N. Mishra, "Emotion Recognition through Combination of Speech and Image Processing: A Review", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3, No. 2, pp. 134-137, 2015

[3] B. Schuller, A. Batliner, D. Seppi, S.Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functional", in: INTERSPEECH 2007, Antwerp, Belgium, pp. 2253-2256, August 27-31, 2007

[4] B. Schuller, G. Rigoll, "Recognizing interest in conversational speech–comparing bag of frames and supra-segmental features", INTERSPEECH, Brighton, UK, pp. 1999-2002, September 6-10, 2009

[5] Y. Zhou, Y. Sun, L. Yang, Y. Yan, "Applying articulatory features to speech emotion recognition", IEEE 9th International Conference on Research Challenges in Computer Science, Shanghai, China, December 28-29, 2009

[6] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, M. Breakspear, "Characterizing Depressed Speech for Classification", INTERSPEECH, Florence, Italy, pp. 2534-2538, August 25-29, 2013

[7] K. M.Chung, D. Jung, "Validity and reliability of the Korean version of autism spectrum disorders comorbid for children (ASD-CC)", Research in Autism Spectrum Disorders, Vol. 39, pp.1-10, 2017

[8] M. A. Siddiqui, N. G. Haider, S. A. Ali, S. Hina, "A: Novel Approach for Features Extraction towards Classifying Normal and Special Children Speech Emotions in Urdu Language", International Journal of Computer Science and Network Security, Vol. 17, No. 7, pp. 188-195, 2017

[9] L. E. Aik, L. C Kiang, Z. B. Mohamed, T. W Hong, "A review on the multivariate statistical methods for dimensional reduction studies", in: AIP Conference Proceedings, Perlis, Malaysia, Vol. 1847, No. 1, AIP Publishing, 2017

[10] K. Morris, P. D. McNicholas, "Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures", Computational Statistics and Data Analysis, Vol. 97, pp. 133-150, 2016

[11] Y. W. Lin, B. C Deng, Q. S Xu,Y. H. Yun, Y. Z. Liang, "The equivalence of partial least squares and principal component regression in the sufficient dimension reduction framework", Chemometrics and Intelligent Laboratory Systems, Vol. 150, pp. 58-64, 2016

[12] K. Mallick, S. Bhattacharyya, "Uncorrelated Local Maximum Margin Criterion: An Efficient Dimensionality reduction Method for Text Classification", Procedia Technology, Vol. 4, pp. 370-374, 2012

[13] Y. Jingjie, X. Wang, W. Gu, L. Ma, "Speech Emotion Recognition Based on Sparse Representation", Archives of Acoustics, Vol. 38, No. 4, pp. 465–470, 2013