TOWARDS AFFECT RECOGNITION: AN ICA APPROACH[1]

Mandar A. Rahurkar and John H.L. Hansen

Robust Speech Processing Group
Center for Spoken Language Research
University of Colorado
Boulder, CO-80303
{rahurkar, jhlh}@cslr.colorado.edu
Web: http://cslr.colorado.edu

ABSTRACT

The speech signal can be thought of as multi-spatial signal where each signal subspace corresponds to different attributes that affect the speech. Language, accent, speaker, emotions are few potential sub-spaces. Each space can further be projected into n-dimensions, as there are different languages, accents and emotions, which combine together to make their sub-space. In this paper, we investigate detecting emotions in speech by projecting an emotional sub-space into a 2-dimensional space. The basis of this space is neutral and stress emotions. A previously formulated stress dependent TEO based feature is employed. To identify the weights of each component we use independent component analysis. This approach was used to exploit the non-Gaussianity of the data. Evaluations are conducted on the SOQ stress database using ICA sub-space decomposition. The results here suggest that dimensionality reduction offers a promising new approach to TEO based stress classification.

1. INTRODUCTION

It is easy to think of emotion as a luxury, an attribute that seems entirely unnecessary for intelligent functioning and hence why bother having a machine with emotional intelligence? Research in neuroscience, psychology and cognitive sciences has suggested that too little of emotion impairs rational decision-making. It may not be necessary for machine to have all the emotional capabilities of human; however having a small subset of these skills would enable them to interact in a much more intelligent manner thus making Human-computer interaction natural and social. The problem of detecting emotion in speech has been the subject of a number of studies [1, 2, 3,14]. Much of the current effort on detecting emotions has been aimed at detecting emotion for improving the robustness of speech recognition algorithms. However, depending on the type of emotion or task induced stress condition, reliable detection, even in noise free environments, continues to be a challenging task. Reliable stress detection requires that a speaker change their neutral speech production process in a consistent manner so that extracted features can detect and perhaps quantify the change. However, there is significant variability in how different speakers convey stress or emotion. A previous study on stress speech classification [1] resulted in the formation of a nonlinear Teager Energy Operator (TEO) based feature TEO-CB-AutoEnv. We use this feature to build the framework for our emotion sub-space projection theory.

The speech signal emotional sub-space is considered as mixture of weighted emotions. The goal of the Blind Source Separation (BSS) is to recover independent sources given only sensor observations that are linear mixtures of independent source signals. Then term blind indicates that both the source signals and the way the signals are mixed is unknown. Independent component analysis is the algorithm for solving the blind source separation problem.

2. TEAGER ENERGY OPERATOR

Historically, most approaches to speech modeling have taken a linear plane wave point of view. While features derived from such analysis can be effective for speech coding and recognition, they are clearly removed from physical speech modeling. Teager did extensive research on nonlinear speech modeling and pioneered the importance of analyzing speech signals from an energy point of view [5,6]. He devised a simple nonlinear, energy tracking operator, for a continuous time signal *x(t)* as follows:

[1] This research was funded by DARPA through SPAWAR under Grant No. N66001-00-2-8906

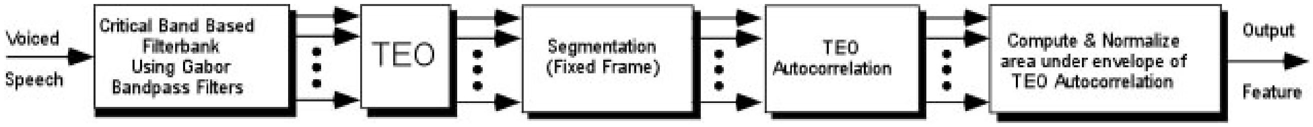

Voiced
Speech
Critical Band Based Filterbank Using Gabor Bandpass Filters
TEO
Segmentation (Fixed Frame)
TEO Autocorrelation
Compute & Normalize area under envelope of TEO Autocorrelation
Output
Feature

Figure 1:Feature Extraction

$$\varphi_c[x(t)] = [\frac{\partial(t)}{\partial t}]^2 - x(t)\frac{\partial^2(t)}{\partial t^2}], \qquad (1)$$

and for a discrete-time signal x (n) as:

$$\varphi[x(n)] = x^2(n) - x(n+1)x(n-1), \qquad (2)$$

where $\varphi[.]$ is the Teager Energy Operator (TEO). These operators were first introduced systematically by Kaiser [7,8].

It has been observed [1] that under stressful conditions, a speech signal fundamental frequency will change and hence the distribution pattern of pitch harmonics across critical bands will be different then for speech under neutral conditions. Therefore, for finer resolution of frequencies, the entire audible frequency range can be partitioned into many critical bands. Each critical band possesses a narrow bandwidth, (i.e., typically 100-400Hz), thus making this new feature independent of the accuracy of median F0 estimation. This is essential as reliable pitch estimation in emotional speech is difficult, since pitch can increase by more then 200 percent in some high stress situations [11].

2.1 TEO-CB-AutoEnv: Critical Band Based TEO Autocorrelation Envelope

We can summarize the feature extraction procedure mathematically as follows using band pass filters (BPF) centered at critical band frequency locations,

$$u_j(n) = s(n) * g_j(n),$$

$$\varphi_j(n) = \varphi[u_j(n)] = u_j^2(n) - u_j(n-1)u_j(n+1),$$

$$R_{\varphi_j^{(i)}(n)}(k) = \sum \varphi_j^{(i)}(n)\varphi_j^{(i)}(n+k),$$

where,

$g_j(n)$, j = 1, 2, 3,…17, is the BPF filter response,

$u_j(n)$, j = 1, 2, 3,…17, is the output of each BPF,

$R_{\varphi_j^{(i)}(n)}(k) =$ Autocorrelation function of the *i*th frame of the TEO profile from the *j*th critical band, $\varphi_j^{(i)}(n)$,

and, N = Analysis Frame length.

Fig.1 shows a flow diagram of the feature extraction process. The TEO-CB-AutoEnv feature has been shown to reflect variations in excitation characteristics including pitch harmonics, due to its finer frequency resolution. However, we believe that the variation in excitation structure is not uniform across all the bands. In our previous work [10], we proposed a new weighted scheme, which supported our hypothesis.

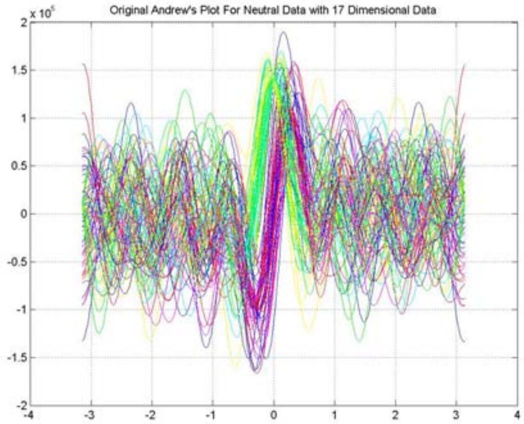
3. INITIAL EXAMINIATION OF FEATURE VECTOR

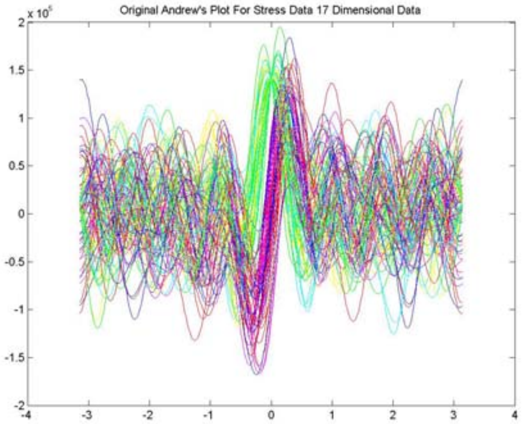
3.1 Andrews Plots

This procedure is essentially very simple; each of the 17 d-dimensional feature vectors, are mapped into a function of form:

$$x(t) = x_1/\sqrt{2} + x_2\sin(t) + x_3\cos(t) + x_4\sin(2t) + ...$$

This function is now plotted for values of t ranging from -pi to +pi. The set of multivariate observations will now appears as a set of lines on the plot. The usefulness of this particular representation lies in the fact that this function preserves Euclidean distances, in the sense that observations close together in 17-dimensional space will correspond to points on the plot that remain close together for all values of t: points far apart in the space will be represented by lines which remain apart for at least some values of t. This property allows the plots to be examined for distinct groups of observations, outlying observations and so on. We look at the original Andrews plot for first 150 frames.





As can be seen from these plots of neutral(top) and stress(bottom) data, it is very difficult if not impossible to draw any meaningful conclusions. Thus, issue of excess feature dimensionality illustrates a lack of a useful classification sub-space.

3.2 Kurtosis

The kurtosis is the fourth central moment divided by fourth power of the standard deviation. Random variables that have kurtosis greater then 3 are called super Gaussian, while the ones having kurtosis less then 3 are called sub-Gaussian. Thus, the kurtosis is 3 for gaussian random variable. In statistical literature, corresponding terms platykurtic and leptokurtic are also used. Kurtosis, or rather its absolute value, has been widely used as a measure of nonGaussianity in ICA and related fields. The main reason is its simplicity, both computational and theoretical. Computationally, kurtosis can be estimated simply by using the fourth moment of the sample data. Theoretical analysis is simplified because of the following linearity property: If x_1 and x_2 are two independent random variables, it holds:

$$kurt(x_1 + x_2) = kurt(x_1) + kurt(x_2)$$

and,

$$kurt(\alpha x_1) = \alpha^4 kurt(x_1)\,.$$

The kurtosis for the stress and neutral data was found to be 3.2325 and 3.1892, which shows that data is non-gaussian. The histogram of stress data and neutral is shown in Figure 2.

Thus as can be seen from the probability distribution functions the pdf resembles Laplacian density rather then Gaussian and hence ICA approach can be used.

4. INDEPENDENT COMPONENT ANANLYSIS

Assume that we observe n linear mixtures $x_1 \ldots x_n$ of n independent emotions,

$$\mathbf{x_j} = \mathbf{a_{j1} s_1} + \ldots + \mathbf{a_{jn} s_n}\,, \quad \text{for all j.} \tag{3}$$

We assume that each mixture x_j as well as each independent component s_k is a random variable, instead of a proper time signal. The observed values $x_j(t)$, in our case, is an emotional signal space in the speech signal. Without loss of generality, we can assume that both the mixture variables and the independent components have zero mean: If this is not true, then the observable variables x_i can always be centered by subtracting the sample mean, which makes the model zero-mean.

It is convenient to use vector-matrix notation instead of sums like in the previous equation. Let us denote by $\mathbf{x}$ the random vector whose elements are the mixtures $x_1 \ldots x_n$, and likewise by $\mathbf{s}$ the random vector with elements $s_1 \ldots s_n$. Let us denote by $\mathbf{A}$ the matrix with elements a_{ij}.

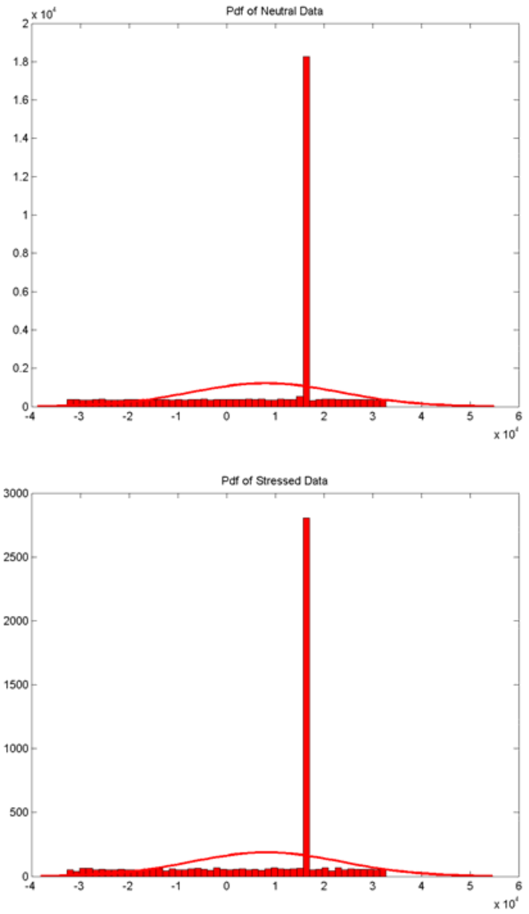


Figure 2: Pdf of Neutral and Stress Data.

Bold lower case letters indicate vectors and bold upper-case letters denote matrices. All vectors are understood as column vectors; thus $\mathbf{x}^\mathbf{T}$, or the transpose of $\mathbf{x}$, is a row vector. Using this vector-matrix notation, the above mixing model is written as:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{4}$$

The statistical model in Equation 5 is called independent component analysis, or ICA model. The ICA model is a generative model, which means that it describes how the observed data are generated by a process of mixing the components s_i. The independent components are latent variables, meaning that they cannot be directly observed. Also the mixing matrix is assumed to be unknown. All we observe is the random vector $\mathbf{x}$, and we must estimate both $\mathbf{A}$ and $\mathbf{s}$ using it. This must be performed under as general a set of assumptions as possible.

The starting point for ICA is the very simple assumption that the components s_i are statistically independent. It will be seen below that we must also assume that the independent component must have nonGaussian distributions. However, in the basic model we do not assume that these distributions are known (if they are known, the problem is considerably simplified.) For simplicity, we also assume that the unknown mixing matrix is square, but this assumption at times be relaxed. After estimating the matrix $\mathbf{A}$, we can compute its inverse, say $\mathbf{W}$, and obtain the independent component simply by:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \qquad (5)$$

ICA is very closely related to the method called *blind source separation* (BSS) or blind signal separation. A "source" means here an original signal, (i.e. independent component, like the speaker in a cocktail party problem). "Blind" indicates that we know very little, if anything, on the mixing matrix, and make little assumptions on the source signals. ICA is one method, perhaps the most widely used, for performing blind source separation. [12].

In our work, we assume *S1… S2* to be two emotions whose mixtures are available, and we have to separate these two emotions.

5. EVALUATIONS

Evaluations were carried out on the SOQ speech corpus. The reason for using this corpus was the availability of biometric measures to confirm the integrity of the data. Evaluations carried out were speaker dependent since we need to have speaker dependent models before we can progress towards the speaker independent models.

Using the fixed point Fast-ICA algorithm, two independent components were calculated for both stress and neutral speech.

5.1. Soldier of the Quarter Board (SOQ) Speech Corpus

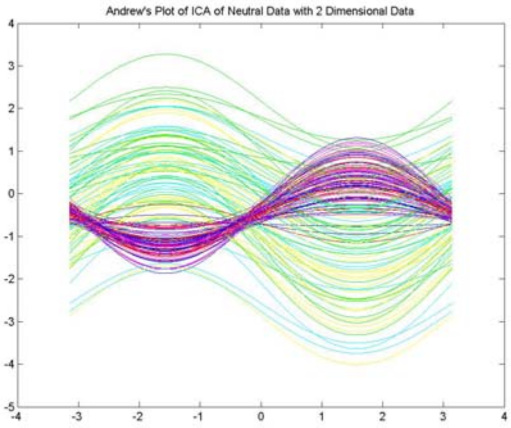
A speech under stress corpus was collected at the Walter Reed Army Institute of Research. The speech corpus was constructed in WRAIR Soldier of the Quarter paradigm [11,12], by recording the spoken response of 6 individual soldiers to questions in a neutral setting, as well as while seated in front of a seven person military evaluation board (all board members had military rank much above the soldier who faced the panel).

The SOQ board is a training exercise and a competition used to prepare soldiers for actual promotion boards. Subjects in this study were candidates in the competition who volunteered to be studied after giving informed consent. Table 1 summarizes average speaker conditions for 6 speakers and 7 speech data collection phases before "Day of Board (DOB)" (A, B, C), during DOB (D), and after DOB (E,F,G). Changes in mean heart rate (HR), blood pressure (sBP, dBP) and pitch (F0) all confirm a change in speaker state between {A,B,C,E,F,G} and D. Results confirm a significant shift in biometric measures from the assumed neutral conditions (A,B,C),(E,F,G), versus the assumed stress condition (D). Each soldier was asked to answer all questions by responding ``the answer to this question is NO". Each speaker was asked the same set of 6 different militarily relevant questions on seven occasions. For our evaluations, we focused our analysis on the word 'NO'.

Summary Of Mean Biometrics For SOQ Subjects					
Measure	A B -7 Day	C -20 min	**D** **Board**	E +20 min	F G +7 Day
HR	70.3	70.8	**93.2**	69.5	67.2
sBP	118	146	**178**	154	117
dBP	77.5	74.8	**89.7**	71.2	69.5
F0	103.4	102.7	**136.9**	104.3	103.1

Table 1: HR - heart rate (in beats per minute), **SBP** -Systolic blood pressure in mm, **dBP** – Dystolic blood pressure in mm, **F0** – Fundamental frequency in Hz.

5.2. Andrews Plots of Independent Components



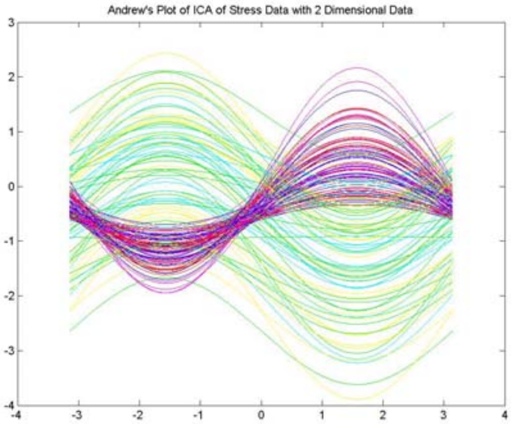


Figure 3:Andrews plot for neutral and stress data with 2 Independent Components after ICA.

We plot the Andrews plot for 2 independent components obtained after using Fast-ICA algorithm. The feature dimension was reduced to two. Figures 3 shows the plots for neutral and stress data respectively. When compared with the original plots one can easily see that they are much easier to interpret. The structure suggests the use of either cluster analysis for analyzing the two components separately.

5.3. Classification using Independent Components

We used 2 methods for classification:
Method 1. Preprocessing using DSLVQ (Distinctive Sensitive learning vector quantization) and classification using least means square algorithm. [13]

Method 2. Preprocessing using fuzzy k-means and classification using voted perceptron algorithm [13,15]

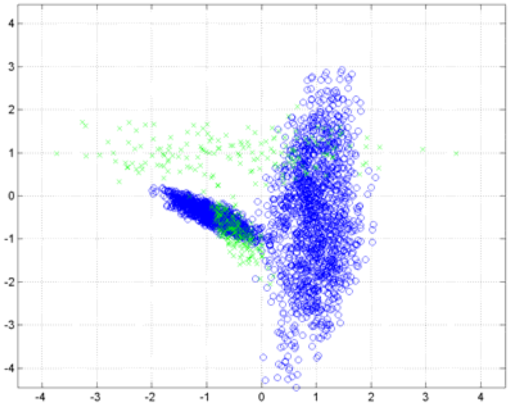


Figure 4:Joint distribution of 2 independent components

Figure 4 shows the original joint distribution between the 2 independent components. Figure 5 shows the classification regions as classified by method 1 while Figure 6 shows the classification regions using method 2. The error rates are shown in the Table 2. The circles (blue) represent the neutral data while the cross (green) represents the stress data point.

Evaluation	%Neutral Error	% Stress Error
Method 1	31.00%	47.00%
Method 2	8.4%	65.00%

Table 2: Classification Results

6. DISCUSSION

We explored a completely new approach to affect recognition and compared few of the multivariate statistical techniques for classification purposes. We believe that instead of applying ICA directly, if some transform could be applied or instead use ICA mixture models, the recognition rate would probably improve. Error rates are high for stress detection. One of the reasons would be that size of the stress data is comparatively smaller than the neutral data. Also it is very difficult to quantify stress and hence high levels of stressed speech detection error might indicate that a portion of this speech is actually mild stress, and thus stress data is probably closer to neutral data instead of stress.

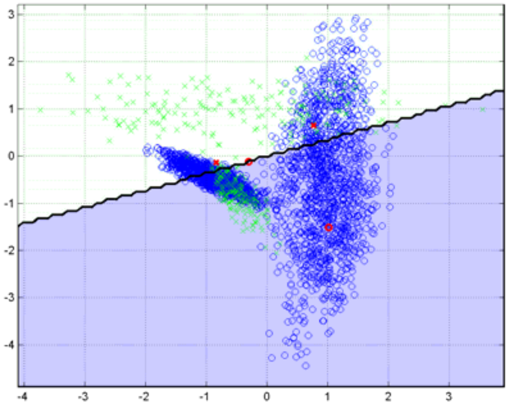


Figure 5:Classification using method 1.

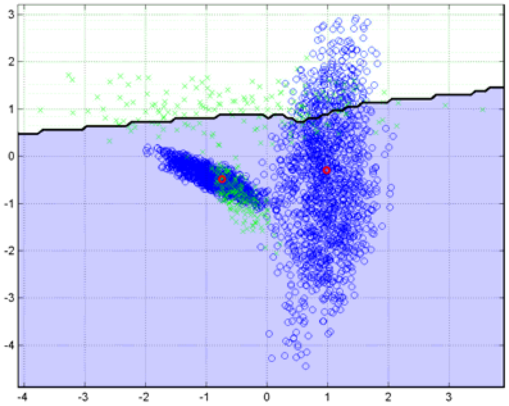


Figure 6:Classification using method 2.

7. CONCLUSION

In this paper we explored a new approach to affect recognition by decomposing the speech signal space. Instead of attempting to present a state of the art system at this stage, we look at the promising properties of ICA for solving this problem. We point out that affect classification is more challenging problem in speech, since ground truth is sometimes, difficult to determine (i.e., for speech or speaker recognition, you are either wrong or right; however for stress or emotion, it is extremely difficult to quantify emotion).

8. ACKNOWLEDGEMENTS

We would like to thank our partners, James Meyerhoff, George Saviolakis, and Michael Koenig, Affiliated with Dept. of Neuroendocrinology, Div. Neuroscience, Walter Reed Army Institute of Research (WRAIR), Silver Spring, Maryland for providing the SOQ speech corpus.

Mandar Rahurkar would also like to thank, Pongtep Angititrakul and Rahul Patil for introducing him to field of multivariate statistics.

8. REFERENCES

[1] G.Zhou, J.H.L. Hansen, and J.F.Kaiser, “Nonlinear Feature Based Classification of Speech under Stress”, *IEEE Trans. Speech & Audio Process*, 9(3): 201-216, Mar. 2001.

[2] J. H. L. Hansen, B.D. Womack, “Feature Analysis and Neural Network Based Classification of Speech Under Stress”, *IEEE Trans. Speech Audio Process*. (4): 307-313, 1996.

[3] D. A. Cairns, J. H. L. Hansen, “Nonlinear Analysis and Detection of Speech under Stressed Conditions”, *J. Acoust. Soc. Am*. (96}(6): 3392-3400, 1994.

[4] J.H.L. Hansen, “Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition”, *Speech Communication,* vol. 20(2), pp. 151-170, November 1996.

[5] H. Teager, “Some Observations on Oral Air Flow During Phonation”, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol.ASSP-28, No.5, pp. 599-601, Oct. 1990.

[6] H. Teager, S. Teager, “Evidence for Nonlinear Production Mechanisms in the Vocal Tract”, Speech Production and Speech Modeling, NATO Advanced Study Institute, vol. 55, Kluwer Academic Pub., pp. 241-261, 1990.

[7] J.F. Kaiser, “On a Simple Algorithm to Calculate the `Energy' of a Signal”, ICASSP-90, pp. 381-384, 1990.

[8] J.F. Kaiser, “On Teager's Energy Algorithm, its Generalization to Continuous Signals,” in Proc. 4th IEEE Digital Signal Processing Workshop, Sept 1990.

[9] M.A.Oleshansky, and J.L. Meyerhoff, Acute catecholaminergic responses to mental and physical stressors in man. Stress Medicine 8:175-179, 1992.

[10] M. Rahurkar, J.H.L. Hansen, M.A.Oleshansky, J.L. Meyerhoff, M. Koenig “Frequency Band Analysis for Stress Detection using a Teager Energy Operator Based Feature”, ICSLP-02, Denver, Colorado.

[11] J.H.L. Hansen, C. Swail, A.J. South, R.K. Moore, H. Steeneken, E.J. Cupples, T. Anderson, C.R.A. Vloeberghs, I. Trancoso, P. Verlinde, *“The Impact of Speech Under `Stress' on Military Speech Technology”* published by NATO Research & Technology Organization RTO-TR-10, AC/323(IST)TP/5 IST/TG-01, March 2000 (ISBN: 92-837-1027-4).

[12] A. Hyvarinen and E. Oja, *“Independent Component Analysis: A Tutorial”*

[13] Pattern Classification (2nd ed), R.O. Duda, P.E. Hart and D.G. Stork.

[14] R.W Picard, E Vyzas, J. Healy, “Toward Machine Emotional Intelligence: Analysis of Affective Physiological State”, *IEEE Trans on PAMI*, vol. 23(10), October 2001.

[15] Y. Freund, R.E. Schapire, “Large Margin classification using the perceptron algorithm”, AT&T Labs.