Comparative Study of Filter Characteristics as Statistical Vocal Correlates of Clinical Psychiatric State in Human

Thaweesak Yingthawornsuk[1] and Chusak Thanawattano[2]

[1]Department of Electrical Technology Education, King Mongkut's University of Technology Thonburi, Thailand
[2]National Electronic and Computer Technology Center, Pathumthani, Thailand

Open Science Index, Biomedical and Biological Engineering Vol:4, No:12, 2010 publications.waset.org/14337.pdf

***Abstract*—** Acoustical properties of speech have been shown to be related to mental states of speaker with symptoms: depression and remission. This paper describes way to address the issue of distinguishing depressed patients from remitted subjects based on measureable acoustics change of their spoken sound. The vocal-tract related frequency characteristics of speech samples from female remitted and depressed patients were analyzed via speech processing techniques and consequently, evaluated statistically by cross-validation with Support Vector Machine. Our results comparatively show the classifier's performance with effectively correct separation of 93% determined from testing with the subject-based feature model and 88% from the frame-based model based on the same speech samples collected from hospital visiting interview sessions between patients and psychiatrists.

***Keywords*—**Depression, SVM, Vocal Extract, Vocal Tract

I. Introduction

Persons who is clinically depressed due to psychiatric disorders, under strenuously living circumstances of social competition and/or experiencing unpleasant childhood in the past could be jeopardized to suicidal risk due to the symptom threatening his/her mind, thought and emotional behavior for a long time without properly clinical treatment under the experienced physician's supervision. Clinical depression has been evidently reported for its prominent precursor related to the risk of suicide in human, in which number of important public heath problems has statistical link to this psychiatric illness. The relationship between the suicidal risk and the major depression has been formerly studied by a group of collaborative researchers from the divergent disciplines of Psychiatry at Vanderbilt Medical Center, Biomedical Engineering at Vanderbilt University in US and Electrical and Computer Engineering in Thailand. The main objective of this study attempts to investigate the relationship between the acoustical properties of patient's speech and his/her severity of mental states, clinically diagnosed of being depressed or suicidal risk by psychiatrists. In the past, the experimental studies have been proposed that the vocal parameters in human speech can assist observation on affection of recognizing pattern and consequently assess level of mental severity of depressive speaker. Attempting methods to identify persons who are severely depressed and likely to loss their life in killing themselves or committing suicide are still in great need in clinical practice and healthcare center. Presently, the most common methods to assess, if patients were at severe state of depression or even at elevated risk of suicide, are self-scored patient survey, report by others, clinical interviews and rating scales, such as the Hamilton depression rating scale [1]. Diagnosis and decision making on clinical categories patient belongs to are the time-consuming procedure in which practitioners must involve information gathering, database checking for background profile, hospital visiting records, diagnosing progress, crime related report by police, and healthcare hotline consultation. The rapid symptom diagnosis to determine if patients were psychologically safe from suicidal risk or clinically identified for one of symptom categories, presently necessitates for physician to respond with correct decision making on admission and treatment for patient in better way of feasibility and availability for diagnosing tool in clinical practice.

As formerly published in pilot studies [2,3,4,5], several analytical techniques were designed and developed for measuring any particular changes in vocal parameters corresponding to the patient's speech production system affectively manipulated by the underlying symptom. Research reports have concluded that the suicidal speech is very similar to depressive one, but its tonal quality of speech significantly changes when speaker is in a moment of near-term suicidal risk. Several acoustical properties of speech parameter extracted by several speech processing techniques such as Glottal ratio/spectrum, pitch contour, frequency distribution

Y. Thaweesak is with King Mongkut's University of Technology Thonburi, Thailand (e-mail: thaweesak.yin@kmutt.ac.th)
C. Thanawattano, is with National Electronic and Computer Technology Center, Pathumthani, Thailand (e-mail: chusakt@gmail.com)

of spectral energy, speech jitters and shimmers, vocal-tract characteristics (i.e., Formants: F_{1-4}), and statistical properties of fundamental frequency (i.e., skewness, kurtosis, and coefficients of variation in F_o) have been studied for vocal properties that can indicate symptom in patients related to major depression [2,3,6]. The relationship between affective characteristics of vocal tract filter response and major depression studied by Tolkmit et al. has been suggested for an existence of identical phonetic context in formant information of vowels from the patient's speech found in clinical observation on patients during their recovery period from depression [7]. Another group of discriminative features is a set of low order mel-cepstral coefficients formerly studied by Ozdas et al in attempt to separate suicidal patients from depressed patients and control subjects. Her study reported the finding of highly significant different measures in set of coefficients among three subject groups and her results on feature model validation came out with the highly correct classification percentages via Gaussian mixture modeling on studied feature samples [3]. All these research results have suggested for the certain change in studied speech parameters being able to determine, which can be used to represent as vocal correlate of the speaker's mental state.

This work is an ongoing project continuing the research quest to determine optimal acoustical parameters with high class discriminative power and to employ as indicator or reflector of mental severity speaker suffers. Alternative speech parameter representing the characteristics of vocal-tract filter system such as formant pattern information was investigated as well for their significant vocal cue in predicting psychiatric state of studied subjects [2]. Alternative way to estimate vocal-tract parameters is explained and implemented instead of employing the conventional method, Linear Prediction Coding, for frequency response of filter, which is model-based approach and can provide inaccurate measure of spectral structure for formant estimation. Therefore, the probabilistic representation of spectral structure of vocal tract filter response is our main study to be focused in analysis on female speech data, which is organized into comparative studies by classification validation between depressed patients and another group of subjects with diagnosis of remission (patients under clinically approval as recovered from previously being depressed) with two different feature sample models in performance evaluation: testing with frame-based feature model and another with subject-based model. In addition, two different types of acoustically controlled audio sample: interviewing speech, recorded during interviewing sessions between patients and clinicians, and reading speech, recorded from post-interviews which patient reads a prepared text passage, were studied to distinguish any posture factors speaker made during audio recording. Analyzed results of classifying speech samples collected from these two different recording conditions will provide more suggestion on speech acoustics affecting class separation corresponding to ways that subjects producing speech, rather than focusing only on the determination of optimal feature with the high class separation.

II. Methodology

A. Database Information

The database used in this study consists of (1) female speech samples recorded from pre-session of interviews with physician and (2) post-session of reading text-contents "Rainbow passage". This passage contains all normal sounds in spoken English with balanced phonetics [8].The categorized groups of depressed and remitted patients comprise, respectively, of eighteen and fifteen females. The patients' ages were within a range from 25 to 65 years during a time of audio recordings. Each subject completed the Beck Depression Inventory-II, (BDI-II), while being interviewed by physician. BDI-II inventory is for mood measure which is known as a standard, brief and self-score questionnaire regarding of mental as well as physical depression related to symptom. Total 21 questions provide a numerical score ranging from 0 to 64 for patients' responses, where the higher scores relate to more suicidal risk [9]. The preprocessing was carried out by first digitizing all speech signals through a 16-bit analog to digital converter at a sampling rate of 10 KHz via a 5 KHz anti-aliasing low-pass filter. All preprocessed speech samples were then screened over to eliminate other voices or any sound artifacts rather than the patients' only voice, as well as silences longer than 0.5 seconds via GoldWave v.5.8 sound editing program. For each female's speech sample, the 8-minute continuous speech collected from interview session was used to represent for interviewing speech sample and 120-second speech from reading session for reading sample in feature extraction procedure. All speech samples were stored for further subsequent analysis.

B. Extraction of Vocal Features

Preprocessed speech samples were first tested and classified into three groups of voiced, unvoiced, and silence through weighting of their estimate of energy on the energy-level thresholds. Only voiced portion of speech being classified were then statistically normalized to adjustment of suitable amplitude to a group baseline for all voiced segments in database. Important tools we used in this project are Matlab scripts that extract the patients' voiced sample and eliminate all silences and other sounds that don't involve the vocal cords (i.e., the sounds of 's' and 'th'). Then the first four dominant Gaussians estimated from the Gaussian mixture model fitted to each 51.2ms frame-based estimate of the cepstral structure of voiced samples were estimated by following procedure described as:

- Divide each voiced speech into segments of 512 points.
- Compute the log-scale cepstrum for each segment.
- Lifter the low-time section of estimated cepstrum by

using a length of window within pitch period detected.

- Convert estimate of the low-time cepstrum to normalized probability density function over 0-5KHz.
- Estimate the Maximum Likelihood (ML) for each GMM distribution via Expectation-Maximization algorithm with modification of frequency component multiplication on each mixture, providing fast iterative convergence in fitting state [10].
- Determine the individual Gaussian's mean, variance and mixture probability from the most four dominant fitted GMM's, discarding those with low ratio of mixture weight and standard deviation.
- Calculate mean and S.D. for every 200, 300 and 400 samples by frame to be used as feature sample input in cross-validation of SVM.
- Evaluate and sort all means and S.D. based on their F-ratio; defined as ratio of the between-class variance and the within-class variance.

C. Feature Classification

The Support Vector Machine (SVM) [11,12] has been a powerful method that outperforms most other classification systems in a wide variety of applications. It achieves relatively robust pattern recognition performance using well established concepts in optimization theory. SVM separates an input $\boldsymbol{x} \in \boldsymbol{R}^d$ into two classes. A decision function of SVM separates two classes by $f(\boldsymbol{x}) > 0$ or $f(\boldsymbol{x}) < 0$. The training data which is used in training phase is $\{x_i, y_i\}$, for $i = 1, \ldots, l$ where $\boldsymbol{x}_i \in \boldsymbol{R}^d$ is the input pattern for the *i*th sample and $y_i \in \{-1, +1\}$ is the class label. Support Vector Classifier maps $\boldsymbol{x}_i$ into some new space of higher dimensionality which depends on a nonlinear function $\phi(\boldsymbol{x})$ and looks for a hyperplane in that new space. The separating hyperplane is optimized by maximization of the margin. Therefore, SVM can be solved as the following quadratic programming problem,

$$\max_{\alpha_i} \{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j \, y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \} \qquad (1)$$

Subject to $0 \le \alpha_i \le C$ and $\sum_{i=1}^{l} \alpha_i \boldsymbol{y}_i = 0$

where C is a parameter to be chosen by user, a larger C corresponding to assigning a higher penalty to errors, and $\alpha \ge 0$ are Lagrange multipliers.

When the optimization problem has solved, system provides many $\alpha_i > 0$ which are the required support vector. Note that the Kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi^T(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j)$ where $\phi(\cdot)$ is a non linear operator mapping input vector $\boldsymbol{x} \in \boldsymbol{R}^d$ to a higher dimensional space. In this work, we choose the polynomial kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = < \boldsymbol{x}_i, \boldsymbol{x}_j >^d$ as the kernel function, where $d \in N$. In addition, other kernels can also be applied. Classification consists of two steps: training and testing. In the training phase, SVM receives some feature patterns as input. These patterns are the extracted speech features represented by N feature parameters that can be seen as points in N-dimensional space. In this paper we extracted twelve features from each frame of voiced sample so that they are multi-dimensional. Then the classifying machine becomes able to find the labels of new vectors by comparing them with those used in the training phase.

III. Results and Discussion

The extracted means, variances and probabilities obtained from Gaussian mixture model fitting on the vocal-tract cepstra were used in training and testing. Feature samples were tested for its sample distribution between two classes of subject through the observable histograms in comparison. Figure 1 shows histograms of the originally extracted mixture probability belonging to the selected Gaussians that suggest us about sample distribution and possible discrimination between studied classes of symptom, clearly with observable tendency of separating distance between two class sample distributions. In addition, the discriminant scores calculated from class sample distributions were measured and compared. It is clear to notify for significant difference in means between two classes. Figure 2 indicates the bar-shaped histograms in term of discriminant score by following a basis of Fisher's discriminant function with pooled covariance between class covariance matrices [13]. The reason of using the single unbiased estimate of population covariance in calculation is simplicity and reasonably high efficiency across a wide variety of feature population models which is suitable assumption for our studied data. As mentioned before in the section of feature extraction, we processed all speech samples on the basis of 51.2ms-length speech frames. Therefore, the distance scores depicted in figure 2 were calculated from a big gathering of all extracted frames from both studied classes of speech sample with blind identifying of numbers of sample frames used to represent for individual subject. Distributions of the weighting probabilities extracted from two classes of speech sample are plotted in discriminant score showing the significant difference in sample means. The sample squared distance between two sample means can be followed by

$$d^2 = (\bar{x}_{rm} - \bar{x}_{dp})' S^{-1} (\bar{x}_{rm} - \bar{x}_{dp}) \qquad (2)$$

where $\bar{x}_{dp}$ is sample mean of vocal extracts of depressed speech population, $\bar{x}_{rm}$ represents sample mean of remitted speech population, and S represents sample pooled covariance matrix combined from two covariance matrices of depressed and remitted populations. As shown in comparative performances validated from SVM classification, all evaluated performances are increasing for all cases of feature combination added in cross-validation. We found that total number of nine sorted features with f-ratio ranking makes classifier's performance gradually degraded. For more insightful understanding on how significantly different the frame lengths of sample used in processing can effects the classification performance, we decided to classify our samples with three different frame lengths in estimating mean and S.D., which are 400, 300 and 200 samples per frame. As our result on performances shown, few features in combination clearly differentiate SVM performances among variety of frame lengths, but the tendency of more similarity in performance score can be noticed at higher combination of six features or more. At more features formed in model combination in case of classifying speech samples with 200samples/frame, performance indicates to be the highest median score of 88% when we performed the frame-based testing on interviewing samples. By applying mean and S.D. of speech extracts calculated from 200 samples/frame processing in subject-based classification with the same interviewing speech samples, performance shows to be the highest score of 93% in median among studying cases.

Therefore, we have decided to use such frame ratio in our investigating process in which we possibly expect the good result of classification performance. More comparative studies on interviewing speech data set revealed that, when we used either the frame-based samples or the subject-based samples (subjects represented by an average of all frames of feature collected individually from that subject) in testing state of SVM cross-validation, the increasing and outperforming performances are obviously be observable form classifying speech samples with 200 samples/frame, compared to other cases of different sample/frame ratios. Figure 5 illustrates the tendencies of means and standard errors obtained from testing classifier with different samples/frame ratios and different features combinations. Performances from classifying mean and S.D. parameters are represented in blue, black, and red lines for applying 200, 300 and 400 samples/frame respectively in estimating for input mean and S.D samples for classification. In figure 5, 6 and 7 small standard errors can be determined from classification scores which are interpreted for very highly reliable statistic of the confidence interval in our classifier's performances. This possibly lead us to the conclusion in that the acoustical features representing the affective characteristics of vocal-tract response which are mediated with spoken speech caused by depression can be utilized for assessment of mental severity in suffered speaker.

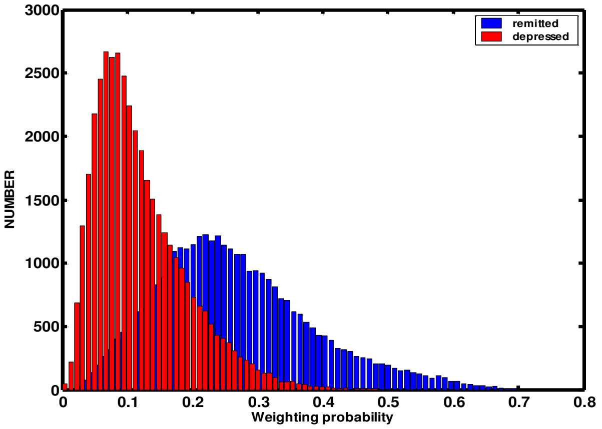

Fig. 1 Distribution histograms of the extracted probabilities

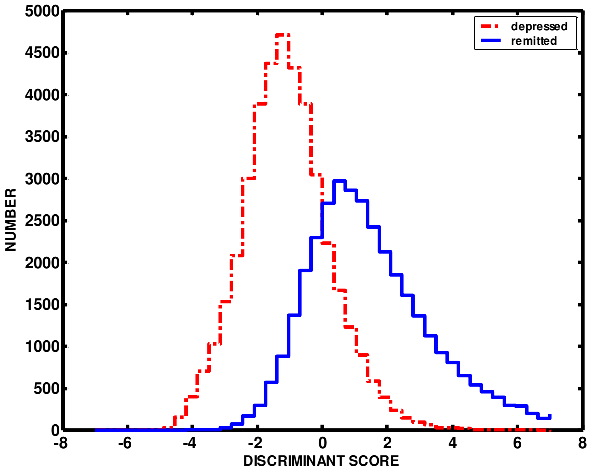

Fig. 2 Discriminant scores calculated from extracted probabilities between depressed (dash line) and remitted (solid line)

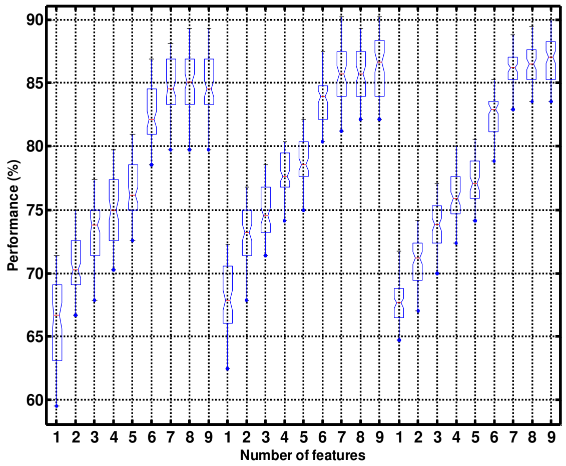

Fig. 3 Comparative performances with respect to different frame lengths of 400, 300 and 200 samples in frame-based classification

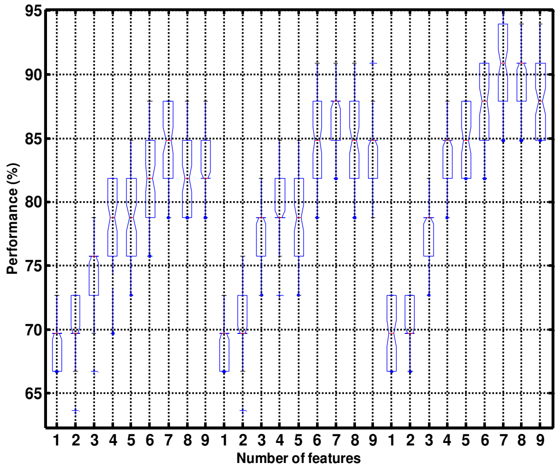

Fig. 4 Comparative performances with respect to different frame lengths of 400, 300 and 200 samples in subject-based classification

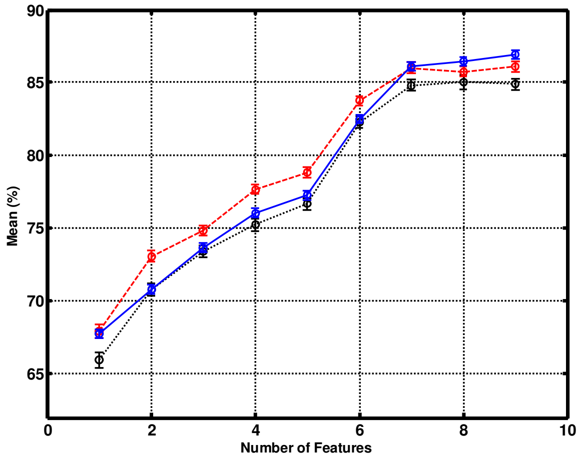

Fig. 5 Comparison of performance mean and S.E. values from frame-based testing with different frame lengths of: 400 (dot line), 300 (dash line) and 200 (solid line) samples

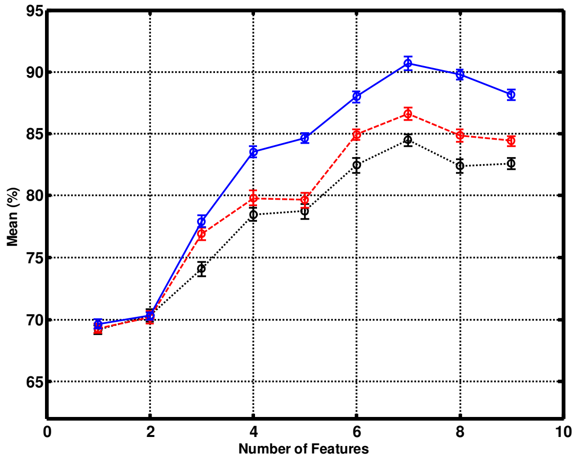

Fig. 6 Comparison of performance mean and S.E. values from subject-based testing with different frame lengths of 400 (dot line), 300 (dash line) and 200 (solid line) samples

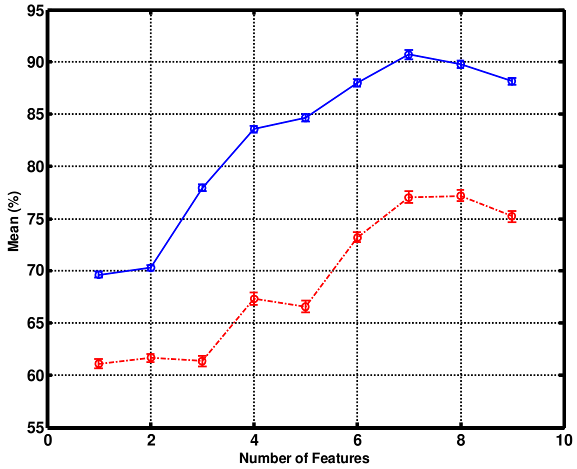

Fig. 7 Overall tendencies of performance mean and S.E. values from classifying interviewing (solid line)and reading speech (dash line)

IV. ONCULSION

This paper demonstrated that probabilistic mixture based vocal parameters representing characteristics of filter/vocal tract extracted from the affective speech samples with depression, in terms of mean, variance and mixture probability achieve for being good indicators as combination for discriminating between clinically diagnosed patients being depressed and patients recovered from formerly being depressed based on extracts of their vocal outcome measurements. Different techniques of audio recording during interviews provide the different results of performance evaluation. In case study of testing interviewing speech samples classifier tends to provide better performance, when compared to that of reading case. Difference in performance suggests speaker having different relative articulation in collaboration between speech production and nervous system acting alternatively between spontaneous speaking and text-reading speaking. Degradation in classifier's performance can be achieved through the measure of F-ratio statistics on speech extracts, which helps organize all speech features in order ranking with their class discriminating power and reduce the dimension of feature model used in cross-validation. This possibly effects the statistical interpretation on analyzed results when size of sample is not large enough to represent a population of data. Results from our empirical study along with findings reported from other research studies claim an existence of impairment in the pathway of speech production due to symptom. More identifiable and quantifiable study possibly helps us to gain more understanding of pathophysiological impact on speech production system.

REFERENCES

[1] M. Hamilton, "A rating scale for depression", Journal of Neurology, Neurosurgery and Psychiatry, Vol. 23, pp. 56-62, 1960

[2] France, D.J., et al., "Acoustical properties of speech as indicators of depression and suicide", *IEEE transactions on BME*, 2000. 47:p 829-837.

[3] Ozdas, A., et al., "Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment", Meth.Info.in Medicine, 2004. 43: p. 36-38.

[4] Ozdas, A., et al., "Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk", *IEEE Transactions on BME*, 2004. 51: p. 1530-1540.

[5] T. Yingthawornsuk, H. Kaymaz Keskinpala, D. France, D. M. Wilkes, R. G. Shiavi, R.M. Salomon, "Objective Estimation of Suicidal Risk using Vocal Output Characteristics", *International Conference on Spoken Language Processing* (*ICSLP-Interspeech* 2006), 2006, pp. 649-652.

[6] T. Yingthawornsuk, et al., "Direct Acuostic Feature using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Risk", *Interspeech 2007*, Antwerp, Belguim.

[7] F. Tolkmitt, H. Helfrich, R. Standke, K.R. Scherer, "Vocal Indicators of Psychiatric Treatment Effects in Depressives and Schizophrenics", *J. Communication Disorders*, Vol.15, pp.209-222, 1982.

[8] G. Fairbanks, *Voice and Articulation Drillbook.* Harper &Row, New York, 1960.

[9] A.T. Beck, st al., "An inventory for measuring depression", *Arch Gen Psychiatry,* 1961. 4:p. 561-571

[10] Dempster, A.P., et al., "Maximum likelihood from incomplete data via the EM algorithm", J. *Royal Stat. Soc. Series B*, 39:1–38, 1977.

[11] V.N. Vapnik, *The Natural of Statistical Learning Theory.* 2nd ed., Springer Verlag (New York), Dec 1999

[12] C. Cortes and V.N. Vapnik , “Support vector networks”, *Machine Learning*, vol.20, pp. 1-25, 1995.

[13] A.J. Richard, *Applied Multivariate Statistical Analysis.* 3th ed., Prentice hall, New Jersey, 1992