Efficient Speech Emotion Recognition Based on Multisurface Proximal Support Vector Machine

Chengfu Yang
Computational Intelligence Laboratory
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu 610054, P.R.China
Sichuan University of Arts and Science
Dazhou 635000, P.R.China
chengfu.yang.uestc@gmail.com

Xiaorong Pu, Xiaobin Wang
Computational Intelligence Laboratory
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu 610054, P.R. China
puxiaor@gmail.com, xbw@uestc.edu.cn

***Abstract*—An efficient speech emotion recognition method based on Multisurface Proximal Support Vector Machine (MPSVM) is presented in this paper. Seven primary human emotions including anger, boredom, disgust, fear/anxiety, happiness, neutral, sadness are investigated using cepstral and spectral features. These novel and robust acoustic features and the multisurface proximal support vector machine classifier based on the Gaussian Mixture Models (GMM) are proposed to yield more correct result. In order to get the normal features in speech emotion space, the corpus of Berlin Database of Emotional Speech is used to train the system, and a simple speech emotion corpus in English, French, Slovenian and Spanish recorded by 2 non-professional speakers are used to test the classifiers. The results achieved by MPSVM are compared by that of the Standard Support Vector Machine (SSVM) classifier. The more efficient and more accurate results are achieved.**

I. Introduction

Researches about emotion in the fields of psychology and physiology have been done for a long time. More recently, Automatic Emotion Recognition (AER) has been an interesting and research topic in the Human-Computer Interaction (HCI) field [2][3][6][8]. In today's HCI systems, machines can know who is speaking and what he or she is speaking of by the speaker and speech recognition system. When the machines are equipped with emotion recognition techniques, they can know how he or she is speaking, and they can react more appropriately and more naturally. There are many potential applications of AER, such as psychiatric diagnosis, intelligent toys, lie detection, learning environment, customer service, educational software and detection of the emotional state in telephone call center conversations to provide feedback to an operator or a supervisor for monitoring purposes [4][12].

People can perceive each other's emotion by the way they talk because the auditory channel carries speech and vocal intonation. The task of AER system is to get the emotion underlying the speech. There are many difficulties in dealing with this problem. The first is the categories definition of the human emotion. In [8], 4 categories of neutral, anger, Lombard and loud were defined. Six different human emotion states were categorized as anger, happiness, fear, surprise, sadness and disgust in [12]. In this work, we are automatically categorizing seven different human emotional states: anger, boredom, disgust, fear/anxiety, happiness, neutral, sadness based on the definition in [10].

The second problem is how to model and express the defined emotions using computer vision. In developing speech emotion recognition system, acoustic features, prosodic features and their statistical characteristics are normally utilized [1][4]. Some work focus on the combination of acoustic features and textual content or linguistic information [6][7]. In our work, some speaker-independent acoustic features are used because our goal is to get the almost same features reflecting the emotion in different languages.

The next problem is the classifiers used in this research field. There is no universal classifier in the literature which is guaranteed to work well with all kinds of data sets. Various different classifiers have adopted to category the emotional states, such as Hidden Markov Model (HMM) [5], Artificial Neural Networks (ANNs) [1] and Support Vector Machines (SVMs) [7][8]. Some work combined the probability model and discriminative model classifier to get more correct results, such as [9]. In [12], the reformulation of SVMs named Least Square Support Vector Machines (LS-SVMs) [16] has been used to get more quickly classification. In this contribution, the Multisurface Proximal Support Vector Machines (MPSVM) [17] is used to get more simply and quickly classification as well as to get more generalization, especially more expediently to extend the classification from binary-class to multi-class.

The rest work is organized as follows. In Section II, the used speech emotion database, Berlin Database of Emotional Speech and two simple testing corpus, is described in brief. In Section III, the MPSVM algorithm is introduced and extend it to be used in speech emotion. Speech emotion recognition system proposed in this work is detailed in Section IV. In Section V, the experimental results are presented, and the last section is the conclusion.

Some notations about this paper: All vectors are column vectors unless transposed to a row vector by a prime super-

978–1–4244–1676–9/08/$25.00 © 2008 IEEE

script '. *x'y* denote the scalar (inner) product of the vector *x* and *y* in the *n*-dimensional real space R^n. $\|x\|$ denote the 2-norm of *x*. For a matrix $B \in R^{m\times n}$, B_i is the *i*th row of *B* and $B_{.j}$ is the *j*th column of *B*. A column vector of ones of arbitrary dimension will be denoted by *e* and the identity matrix of arbitrary order will be denoted by *I*. For $A \in R^{m\times n}$ and $B \in R^{n\times k}$, a kernel *K(A,B)* maps $R^{m\times n} \times R^{n\times k}$ into $R^{m\times k}$.

II. The Used Speech Emotion Database

In this work, we used two emotion databases. The one created in [10] is a professional reference database recorded in German. In the database, there are the sound files itself, the label files (syllable label files and phone label files), information about the results of different perception tests (including the recognition of emotions, the evaluation of naturalness, the syllable stress and the strength of the displayed emotions) as well as some results of the measurements of fundamental frequency, energy, loudness, duration, stress and rhythm. The sound files were recorded by five actors and five actress. The content of the sound files include ten independent text spoken by different actor or actress in seven emotion states.

The other speech emotion corpus [13] is recorded in English, French, Slovenian and Spanish by two non-professional speaker which One is a man and another is a woman. There are four emotions included in this corpus that are anger, sadness, neutral and joy. The text content of the speech are composed of words and long sentences. All the utterances are approved by two experts in order to be genuine.

III. Multisurface Proximal Support Vector Machines

A. Binary Nonlinear MPSVM

Multisurface Proximal Support Vector Machine (MPSVM) combines the standard support vector machines with the generalized eigenvalue problem. MPSVM can get two nonparallel proximal planes which each plane is as close as possible to one of the data sets and as far as possible from the other data sets. Each of the nonparallel proximal planes is generated by an eigenvector corresponding to a smallest eigenvalue of each of the generalized eigenvalue problems which are $Gz = \lambda Hz$ and $Lz = \lambda Mz$ [17]. In this work, we first introduce the nonlinear MPSVM, and then extend it to the multi-class MPSVM for speech emotion recognition.

Let the m points in the n-dimensional real space R^n to be classified into two patterns, one be denoted by the $m1 \times n$ matrix $C1$ belonging to class 1 and another be represented by the $m2 \times n$ matrix $C2$ belonging to class 2, with $m1 + m2 = m$. The nonlinear MPSVM will get two following kernel-generated proximal surface:

$$K(x', C')w^1 - b^1 = 0, \;\; K(x', C')w^2 - b^2 = 0 \qquad (1)$$

where C=[C1; C2], and K is an arbitrary kernel as defined:

$$(K(A,B))_{ij} = e^{-\mu\|A_i' - B_{.j}\|^2} \qquad (2)$$

where $A \in R^{m\times n}$, $B \in R^{n\times k}$, and μ is a positive constant.

In order to obtain the first plane of (1), we minimize the sum of the squares of two-norm distances between each of the points of class 1 to the plane divided by the squares of two-norm distances between each of the points of class 2 to the plane. This lead to the minimization problem as following:

$$\min_{(w,b)\neq 0} \frac{\|K(C1,C')w - eb\|^2 + \varepsilon\|[w;b]\|^2}{\|K(C2,C')w - eb\|^2}. \qquad (3)$$

On the other hand, in order to obtain the second plane of (1), we minimize the sum of the squares of two-norm distances between each of the points of class 2 to the plane divided by the squares of two-norm distances between each of the points of class 1 to the plane. This lead to the minimization problem as following:

$$\min_{(w,b)\neq 0} \frac{\|K(C2,C')w - eb\|^2 + \varepsilon\|[w;b]\|^2}{\|K(C1,C')w - eb\|^2}. \qquad (4)$$

By making the definitions:

$$\begin{cases} G = [K(C1,C') - e]'[K(C1,C') - e] + \varepsilon I \\ H = [K(C2,C') - e]'[K(C2,C') - e] \end{cases} \qquad (5)$$

and

$$\begin{cases} L = [K(C2,C') - e]'[K(C2,C') - e] + \varepsilon I \\ M = [K(C1,C') - e]'[K(C1,C') - e]. \end{cases} \qquad (6)$$

Define $z = [w; b]$, then the optimization problems become:

$$\begin{cases} \min\limits_{z\neq 0} \; r(z) = \dfrac{z'Gz}{z'Hz} \\ \min\limits_{z\neq 0} \; s(z) = \dfrac{z'Lz}{z'Mz}. \end{cases} \qquad (7)$$

This leads to the respective generalized eigenvalue problems:

$$\begin{cases} Gz = \lambda Hz, \;\; z \neq 0 \\ Lz = \lambda Mz, \;\; z \neq 0. \end{cases} \qquad (8)$$

By the two MATLAB commands: eig(G,H) and eig(L,M), we can get the proximal surface (1) corresponding to the smallest eigenvalues of the (8).

B. Modifications of MPSVM for Speech Emotion Recognition

In order to apply the MPSVM to speech emotion recognition, there are some modifications need to be achieved. The first step is to extend the binary MPSVM to multi-class MPSVM. According to the characters of MPSVM that there is a nonparallel plane for every pattern, we use the one-to-rest method to solve this problem. In training phase, the proximal plane of one emotion class is obtained by regarding the other emotions as another emotion class, and the same procession to all emotion classes. In this way, we get seven nonparallel proximal planes corresponding to seven emotions to be recognized in this work.

Because of the number of the samples $(m = m1+m2+...)$ is very large in the training phase, the techniques of the Reduced Support Vector Machine (RSVM) [15] classification must be applied to reduce the dimensionality m of the generalized eigenvalue problem to $\overline{m}$. This is easily achieved by replacing the kernels $K(Ci, C')$ by the reduced kernels $K(Ci, \overline{C}')$, where $i = 1 \sim 7$ and $\overline{C}$ is the matrix formed by taking a small random samples of the rows of C. Another problem of the sample unbalance in training phase can be solved at the same time by the method of RSVM.

IV. Speech Emotion Recognition System

A. Preprocessing on Speech Signal

The first stage of preprocessing is to de-noise by soft-thresholds the wavelet coefficients after three level of decomposition. The next processing for de-noised speech signal is endpoint detecting. In this stage, the silent parts of the speech signals will not be eliminated by the threshold of the amount of energy in the small intervals of the signal because this is probably applying some emotions. The pre-emphasis, frame-partition and Hamming-windowing are the next three stages in the preprocessing. The pre-emphasis is to get balance between low frequency and high frequency signals. Because we focus on acoustic features in this work, the frame-partition and the Hamming-windowing are the necessary stages in the preprocessing.

B. Features Extraction

The prosodic features and their statistical characteristics of speech were proposed in classifying the emotions in most former works[1][9]. In this contribution, we are using spectral features, cepstral features listed in Table I and time-domain features listed in Table II. All the features are extracted from the speech signals after the stages of preprocessing. The mean and standard deviation for each feature based on frame-partition are considered to constitute the feature vector. Then we get the speech emotion space of 150 dimensions. Every training emotion speech utterance can be projected into this space and to get a series of vectors which will be classified with the classifier MPSVM.

After the foregoing steps, the vectors in feature space are sent to the GMM algorithm to get a series of vectors respective to different speech emotions. These vectors will be the training samples to train the MPSVM. This step will increase the robust of the recognition.

TABLE I
The Spectral and Cepstral Features Used for Speech Emotion Recognition

Spectral Features	**Cepstral Features**
16 LPC	12 LPCC
pitch maximum gradient	16 PLP
pitch relative position of maximum	20 MFCC
pitch mean value gradient	12 ASCC
pitch mean value	
pitch relative maximum	
pitch range	
pitch relative position of minimum	
pitch relative minimum	
pitch mean distance between reversal points	
spectral energy below 250 Hz	
spectral energy below 650 Hz	

TABLE II
The Time-domain Features Used for Speech Emotion Recognition

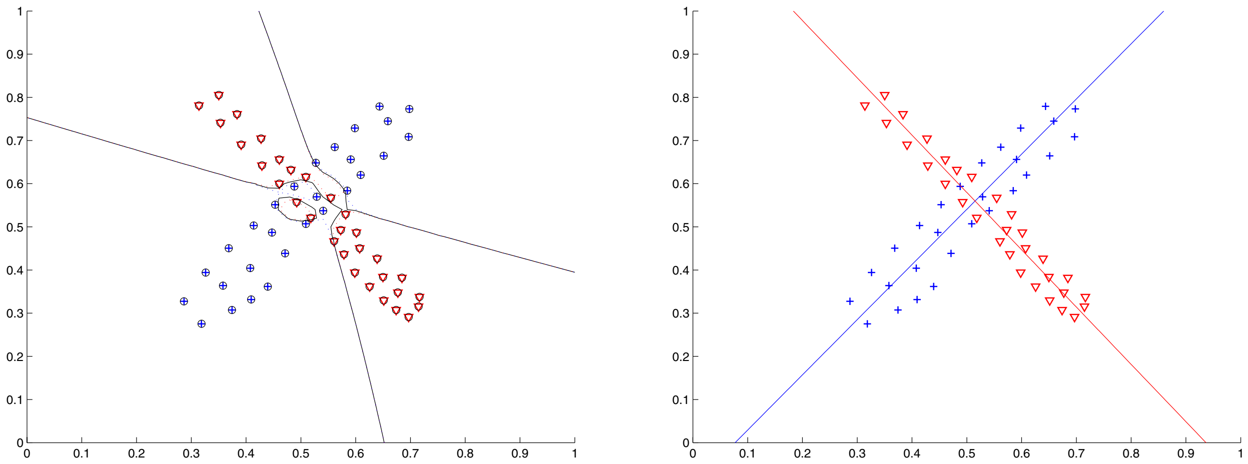
Time-domain Features
energy mean distance between reversal points
duration mean value of voiced sounds
energy mean of fall-time
energy median of fall-time
energy mean value
energy mean of rise-time
energy median of rise-time duration of silences mean value
signal number of zero-crossings
energy relative maximum
energy relative position of maximum
energy maximum gradient
duration of silences median

C. Classification Based on MPSVM

The standard support vector machine (SSVM) [14] show a high generalization capability based on their structural risk minimization optimized principle. At the same time, due to the support vectors, the discriminative plan is spanned by the sparse references from the training samples. The kernel functions map the data points from low dimension space to high dimension space in which the data points can be linearly separable in underlying mode. On the other hand, a quadratic function subject to linear inequality constraints must be solved in cost of computational complexity. Many multi-class SSVM were proposed to apply this technology into practical applications. The Least Square-Support Vector Machine (LS-SVMs) [16] replaced the inequality constraints with equality constraints. The solution follows from solving a set of linear equations instead of a quadratic programming problem, accompanied with losing sparse reference results. In this work, we use the Multisurface Proximal Support Vector Machine (MPSVM) [17] as the classifier. MPSVM can get the classifying planes for every pattern class in a simple MATLAB command with less computational complexity. The MPSVM combines the principles of SSVM and LS-SVMs, and it can easily be extended to multi-class case.

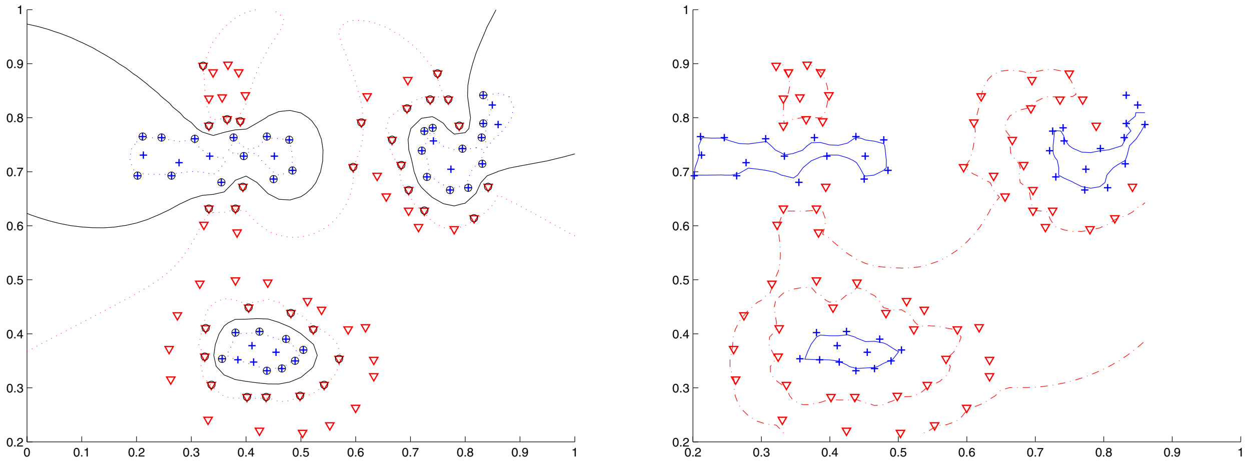
The structure of the proposed speech emotion recognition system is depicted in Fig.1.

The results of linear and nonlinear distribution for binary problem solved in SSVM and MPSVM are displayed in Fig.2. and Fig.3.. From the Fig.3., we can see that for nonlinear problem, the MPSVM can get more generalization results. This case is suitable for speech emotion classification.



(a) SSVM with Gaussian Kernel (b) MPSVM with Linear Kernel

Fig. 2. Comparison Between SSVM and MPSVM in Linear Distribution



(a) SSVM with Gaussian Kernel (b) MPSVM with Exponential RBF Kernel

Fig. 3. Comparison Between SSVM and MPSVM in Nonlinear Distribution

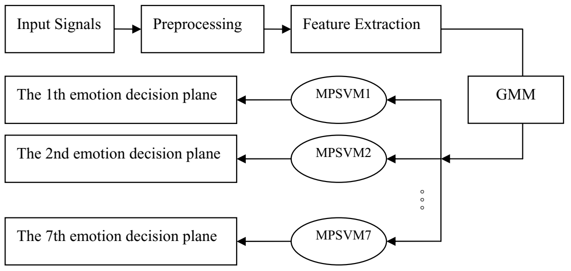

Input Signals
Preprocessing
Feature Extraction
GMM
MPSVM1
MPSVM2
MPSVM7
The 1th emotion decision plane
The 2nd emotion decision plane
The 7th emotion decision plane

Fig. 1. The Structure of the Speech Emotion Recognition

The Fig.4. show the multi-class classification result in MPSVM. The simulations in later sections proved that the multi-class speech emotion classification based on MPSVM can get more generalization results than other SVM algorithms.

V. Experimental Results

Our experiments are based on two speech emotion databases. One is a professional reference database recorded in German [10]. The database consist of 700 instances in seven emotions recorded by five actors and five actresses. In order to test the robust of speech emotion recognition in different languages, we used another speech emotion database [13], recorded in English, Spanish, Slovenian and French which includes four emotions. We used it as test database to evaluate the classification plane we have get.

The first test is to evaluate the correctness of in-set speech emotional instance by using the training set as testing set. In this test, we use 70 instances in [10], 10 for every emotion which was spoken by one actor or actress, to train and test the system. AS it is shown in Table III, the high correctness means that the classification is feasible in training phase.

The second test is to evaluate the correctness of out-set speech emotional instance in same language by using the rest

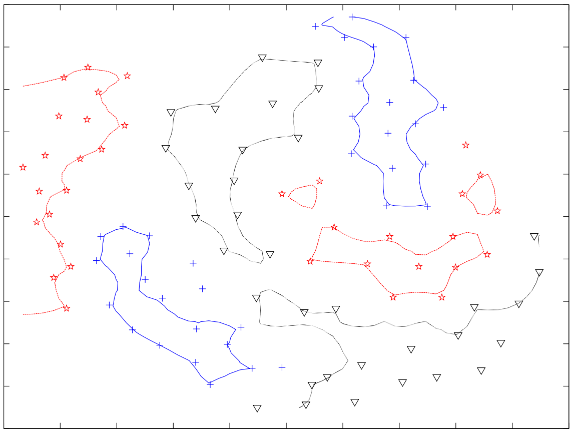


Fig. 4. MPSVM for Multi-class Classification with Exponential RBF Kernel

TABLE III
THE PERFORMANCES IN IN-SET AND OUT-SET TEST

Accuracy(%)	In-set Test	Out-set Test
anger	100.0	98.3
happiness	98.4	96.2
sadness	94.2	89.2
neutral	95.1	82.3
boredom	89.6	75.6
disgust	87.8	74.3
anxiety/fear	93.1	87.6
training time	20(s)	20(s)
test time	30(s)	10(s)

speech instances except for the training set in database [10]. In this test, we use 70 instances to train the system and other 30 instances to test the system. As it is shown in Table III, the correctness is lower than the in-set test, and it is natural. The best performance belongs to the speech emotion of anger.

The third test is to evaluate the correctness of speech emotional instances in different language. In this test, 40 instances of anger, happiness, sadness and neutral emotion in [10] were used to train the system based on MPSVM. And for every testing language, we selected 5 instances in [13] for every emotion to test the system. AS it shown in Table IV, the performance is not very high which charges upon the features we have picked up are not very robust for speech emotions in different language.

The fourth test is to compare the performance based on SSVM and MPSVM. In this test, we used 70 instances in [10] to training the system based on SSVM and MPSVM, and other 70 instances in [10] to test the system. By using the same training and testing speech instances, as it is shown in Table V, the algorithm based on MPSVM can get high performance in efficient way.

VI. CONCLUSION

In our contribution, we test a set of acoustic features which are abstracted from German, English, French, Slovenian and Spanish, and are used in speech emotion for the first time. For classification, we used the MPSVM which is a newly and powerful classifier with quick and simple solution as well as with good generalization. We achieved an overall classification accuracy as same as the other system, but with less computational time. In the future work, we will extend to find more robust features for speech emotion among different languages.

TABLE IV
THE PERFORMANCES IN DIFFERENT LANGUAGE

Accuracy(%)	German	English	Spanish	Slovenian	French
anger	98.7	85.6	69.1	72.5	84.1
happiness	96.4	82.5	67.2	73.4	82.3
sadness	89.1	68.1	62.1	64.5	74.2
neutral	82.5	73.4	64.3	65.8	72.3

TABLE V
THE COMPARISON PERFORMANCES BETWEEN SSVM AND MPSVM

Accuracy(%)	SSVM	MPSVM
anger	93.2	98.2
happiness	87.7	96.2
sadness	89.2	88.4
neutral	90.5	86.7
boredom	86.2	85.2
disgust	84.3	86.2
anxiety/fear	88.3	90.6
training time	250(s)	20(s)
test time	25(s)	25(s)

ACKNOWLEDGMENT

This work was supported by Chinese 863 High-Tech Program Under Grant 2007AA01Z321.

REFERENCES

[1] J. Nicholson, K. Takahashi and R. Nakatsu, *Emotion recognition in speech using Nueral Networks*, Proceedings of the 6th International Conference on Neural Information Processing, vol. 2, pp. 495-501, 1999.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G. Taylor, *Emotion recognition in Human-Computer Interaction*, IEEE Signal Proc. Mag., 18(1), pp. 32-80, 2000.

[3] O.W. Kwon, K. Chan, J. Hao and T.W. Lee, *Emotion Recognition by Speech Signals*, Eurospeech, pp. 125-128, 2003.

[4] C.D. Park and K.B. Sim, *Emotion Recognition and Acoustic Analysis from Speech Signal*, Proceedings of IJCNN, pp. 2594-2597, 2003.

[5] B. Schuller, G. Rigoll and M. Lang, *Hidden Markov model-based speech emotion recognition*, Proceedings of the International Conference on Multimedia and Expo-Volume 2 pp. 401 - 404, 2003 .

[6] Z.J. Chuang and C.H. Wu, *Emotion recognition using acoustic features and textual content*, IEEE international conference on multimedia and expo, Vol. 1, pp. 53-56, 27-3 June 2004.

[7] B. Schuller, G. Rigoll and M. Lang, *Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture*, Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing, vol.1, pp. I-577-580, 17-21 May, 2004.

[8] T. Nguyen and I. Bass, *Investigation of combining SVM and decision tree for emotion classification*, 7th IEEE International Symposium on Multimedia, pp. 540-544, 2005.

[9] Y.L. Lin and G. Wei, *Speech emotion recognition based on HMM and SVM*, Proceeding of International Conference on Machine Learning and Cybernetics, Vol. 8, pp. 4898-4901, 18-21 Aug. 2005.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, *A Database of German Emotional Speech*, Proceedings Interspeech, Lissabon, Portugal, 2005.

[11] B. Schuller, S. Reiter and G. Rigoll, *Evolutionary Feature Generation in Speech Emotion Recognition*, IEEE International Conference on Multimedia and Expo, pp. 5 - 8, 9-12 July 2006.

[12] T.S. Tabatabaei, S. Krishnana and A. Guergachi, *Emotion Recognition Using Novel Speech Signal Features* , IEEE International Symposium on Circuits and Systems, ISCAS 2007. pp. 345 - 348, 27-30 May 2007.

[13] DSPLAB, *http://wwwbox.uni-mb.si/eSpeech/*, University of Maribor, Faculty of Electrical Engineering and Computer Science, Slovenia.

[14] V.N. Vapnik, *The Nature of Statistical Learning Theory*, second ed. New York: Springer, 2000.

[15] Y.-J. Lee and O.L. Mangasarian, *RSVM: Reduced Support Vector Machines*, Proc. First SIAM Int'l Conf. Data Mining, Apr. 2001.

[16] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, *Least Squares Support Vector Machines*, second ed. Singapore: World Scientific Publishing Co., 2002.

[17] O.L. Mangasarian and E.W. Wild, *Multisurface Proximal Support Vector Machine Classification via Generalized Eigenvalues*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, No. 1, pp. 69-74, January 2006.