Feature Space Dimension Reduction in Speech Emotion Recognition Using Support Vector Machine

Bo-Chang Chiou and Chia-Ping Chen
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
Email: m013040072@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw

***Abstract*—We report implementations of automatic speech emotion recognition systems based on support vector machines in this paper. While common systems often extract a very large feature set per utterance for emotion classification, we conjecture that the dimension of the feature space can be greatly reduced without severe degradation of accuracy. Consequently, we systematically reduce the number of features via feature selection and principal component analysis. The evaluation is carried out on the Berlin Database of Emotional Speech, also known as EMO-DB, which consists of 10 speakers and 7 emotions. The results show that we can trim the feature set to 37 features and still maintain an accuracy of 80%. This means a reduction of more than 99% compared to the baseline system which uses more than 6,000 features.**

I. Introduction

Affective computing is an emerging field of researches, covering a broad range of expertise, e.g., computer sciences, psychology, cognitive science, and engineering [1]. It is concerned with the recognition and synthesis of human emotional expressions, and the integration of such in computational systems. Emotional expressions convey information of the internal states of human, and they play important roles in achieving high levels of user satisfaction in interactive systems, such as information agents or computer games. The basic means to the recognition of emotion is through the visual (expression and color of faces) or acoustic cues (voices), as in the PHYSTA project [2]. In addition, lexical and discourse information can be integrated as well [3]. Besides, emotion can be described by arm gesture to present music emotion [4]. In emotion researches, how to collect the real emotion is a problem. When collecting frustrated and delighted smiles, nature emotion is different with acted emotion [5]. Participants are at a unfamiliar site sometimes also cannot express the real emotion. Crowdsourcing which collects data through Internet can let participants at their room or any they feel comfortable. [6].

The research in this paper belongs to the area of speech emotion recognition (SER). Many approaches have been researched on SER. An SER system generally is equipped with a front-end feature extraction module and a back-end recognition/classification module. A survey on SER with respect to features, classifiers, and databases can be found in [7]. Common acoustic features include spectral features, energy features, cepstral features, quality features, prosodic features, and pitch-related features. For example, the Interspeech 2009 Emotional Speech Challenge [8] adopted a feature set containing prosodic, spectral and HMM-based features. Common recognizers include model-based classifiers such as hidden Markov models (HMM) and artificial neural networks (ANN), and examplar-based classifiers such as support vector machines (SVM). Comparison across databases has been carried out in [9], where evaluation results on 9 databases in combination of different recognizers are reported.

Our system is evaluated on the Berlin Database of Emotional Speech (EMO-DB). EMO-DB is a popular database, on which many works have been reported. In [10], SVM, multiple layer perceptron (MLP), and probabilistic neural networks (PNN) on the same feature set were experimented, with SVM getting 78% accuracy. In [11], half of the 4,368 features adopted in [8] are reduced for each emotion by Kolmogorov-Smirnov test, while still achieving accuracy as high as 88% for certain emotion classes. In [12], temporal interval time information is studied, and achieving 91.6% accuracy with the optimal feature on MLP. In [13], a small feature set of segment-level features such as the fundamental frequency (F0) and mel-frequency cepstral coefficients (MFCC) with global statistics is used. In contrast, a huge feature set containing 6,552 acoustic features is used in [9]. On EMO-DB, this large set of features achieves 85.2% accuracy.

In this research, we aim to find an optimal feature set with the minimum number of features while maintaining a decent recognition accuracy for SER on EMO-DB. In SER, the common approach is to extract a very large feature set and use it to train a recognizer. However, not all these features have a positive impact on recognition. Too many features not only reduce performance of recognition, but also increase computing time. Thus, it is of interest to investigate the relationship between the number of features and the performance level. Efforts for finding optimal features have been attempted in the past. In [13], the linear discriminant analysis (LDA) is applied to reduce the feature space. In [14], a scheme called

iterative feature normalization (IFN) is explored to reduce speaker variability by normalizing the acoustic features from neutral utterances. At this moment, the SER community still have not converged to a best default feature set [15].

The baseline feature set [9] used in this paper contains many acoustic features and functionals. First, we explore the delta regression's effects on SER. Second, we try to find out the optimal acoustic feature set and functional set. Finally, the principal component analysis (PCA) is employed to aggressively reduce the number of features. Our goal is to reduce feature number under 100, while keeping the accuracy above 80%.

This paper is organized as follow: Section 2 introduce Support Vector Machine and the common low-level descriptors. Section 3 is split 3 part: First, we describe database information; second part lists the detail of baseline feature set; third part shows how we reduced the feature set in different ways. Last, our conclusion are drawn in the final section.

II. Frond-End Feature and Back-end

A. Basic Acoustic Feature

We introduce zero-crossing rate (ZCR), signal energy, pitch, and MFCC, which are common features in speech recognition.

ZCR is the sign change rate of waveform within a given frame and is often used with energy for end-point detection.

Signal energy is variation of voice intensity. Under fixed window size, signal energy is:

$$E = \left(\sum_{n=0}^{N} {x_n}^2 \right) /N \tag{1}$$

where N is PCM frame size and PCM frame values x_n, $n = 1..N$. Because human do not hear loudness on a linear scale, it is more similar to log scale. Log energy formula is:

$$E_l = log \left[\left(\sum_{n=0}^{N} {x_n}^2 \right) /N \right] \tag{2}$$

Pitch means the human perception of audio signal which can be represented by fundamental frequency or equivalently, the reciprocal of the fundamental period of voiced audio signals. In our system, we use Auto-correction function to estimate pitch.

In speech recognition, the most common feature is MFCC. Human auditory system are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Mel scale is to match closely to human. The formula for converting from frequency to Mel scale is:

$$M(f) = 1125 \ln(1 + f/700) \tag{3}$$

MFCCs consider this characteristic so that they are suit to speech recognition. In the implementation steps, the most important thing is to multiply the magnitude frequency response by a set of 20-40 (standard 26) triangular bandpass filters to get the energy of each triangular bandpass filter. It applies Mel frequency to get the filter banks. The reason why using triangular bandpass filters is to smooth the magnitude spectrum in order to obtain the envelop of the spectrum with harmonics. Therefore, MFCCs will not contain pitch of utterances information.

B. Support Vector Machine

SVM is to find the optimal separating hyperplane which separates two different label sets. Given a set of data $\{x_i, y_i\}, i = 1..n$ where $x_i \in R^d$ denotes the input vector, $y_i \in \{+1, -1\}$ denotes the output value. Optimal separating hyperplane formula is as:

$$w \cdot \phi(x) + b = 0 \tag{4}$$

where x is input vector, w is weight vector, and $\phi()$ is a mapping function in non-linear SVM. When linear SVM cannot solve the problem, non-linear SVM uses kernel function to project vector to higher dimensional space[16]. Then SVM finds a linear separation linear hyperplane from high dimensions. SVM can be formulated as following optimal problem:

Minimize

$$\phi(w) = \frac{1}{2}||w||^2$$

Subject to

$$y_i (w \cdot \phi(x_i) + b) \geq 1$$

Above the optimal question have solution if and only if it exists one optimal separating hyperplane to separate data perfectly. For non-separable case, it must add a slack variable ξ to release the restrict condition. Then, new optimal problem is:

Minimize

$$\phi(w, \xi_i) = \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} \xi_i$$

Subject to

$$y_i (w \cdot \phi(x) + b) \geq 1 - \xi_i, \ \xi_i \geq 0 \quad \forall i$$

where C is the penalty parameter of the error term. The decision function of SVM is defined as:

$$f(x) = sign \left(\sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b \right) \tag{5}$$

where α_i is Lagrange multipliers and $k()$ is kernel function.

III. Experiment

We use the common database EMO-DB [17]. This database contains utterances from ten native speakers of German, 5 male and 5 female speakers. Each speaker speaks 10 utterance with 7 emotion states: anger, boredom, disgust, anxiety/fear, happiness, sadness, and neutral. In order to provide more reliable data, each utterance have been taken perception-test by 20 evaluators. Most utterances in database have more than 80% recognition and 60% naturalness. We used all 535 utterances in our evaluation.

TABLE I
56 LOW-LEVEL DESCRIPTORS

Feature Group	Feature in Group	# of LLD
Raw Signal	Zero-crossing-rate	1
Signal energy	logarithm	1
Pitch	Fundamental frequency F0 in Hz via Cep strum and Autocorrelation (ACF). Exponentially smoothed F0 envelope.	2
Voice Quality	Probability of voicing ($\frac{ACF(T0)}{ACF(0}$)	1
Spectral	Energy in bands 0 - 250 Hz, 0 - 650 Hz, 250 - 650 Hz, 1 - 4 kHz 25 %, 50 %, 75 %, 90% roll-off point, centroid, flux, and rel. pos. of spectrum max. and min.	12
Mel-spectrum	Band 1-26	26
Cepstral	MFCC 0-12	13

TABLE II
39 FUNCTIONALS APPLIED TO LLD

Functionals, etc.	# of functionals
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. and min. value - arithmetic mean	2
Arithmetic mean, Quadratic mean	2
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95% and 98% percentile	2
Std. deviation, variance, kurtosis, skewness	4
Centroid	1
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. mean	4
Linear regression coefficients and corresp. approximation error	4
Quadratic regression coefficients and corresp. approximation error	5

A. Baseline

The baseline feature set is used in [9]. OpenEAR toolkit is used to extract 6552 features consisting of 39 functionals of 56 acoustic low-level descriptors (LLD) along with the first and second order delta regression coefficients. TABLE I shows all LLD and TABLE II shows all functionals mapping LLD from time series onto a statical value. Note that these features are utterance-level.

B. Experimental Setup

Speech input is processed using 25ms Hamming window, with a frame rate of 10ms. Feature values are normalized before training or prediction. We apply the Leave-One-Speaker-Out method to speaker independent experiments. The back-end classifier is support vector machine with polynomial kernel based on Sequential Minimal Optimization (SMO) [18]. We first remove certain regressions, feature groups, and functionals. Then we apply PCA for further feature dimension reduction.

C. Reducing Regression Features

The baseline set includes first and second order delta regression coefficients. The delta coefficients are computed using (6).

$$d^t = \frac{\sum_{i=1}^{W} i\left(x^{t+i} - x^{t-i}\right)}{2\sum_{i=1}^{W} i^2} \tag{6}$$

x^t is the tth frame, and W specifies half the size of the window to be used to computation of the regression coefficients, and we set W to 2. Equation (6) relies on past and future speech parameter. Performance of speech recognition can be prevented by adding time derivatives. To know whether delta regression is good for SER or not, we design three experiments: baseline set without second order delta, baseline set without first order delta, and baseline set without first and second order delta regression. The first result gets 86.1% accuracy; the second result gets 83.7% accuracy; the third result gets 83.3% accuracy. TABLE III shows 3 experimental results. According to the above results, we assume that second delta regression can be delete to get a smaller feature set. From comparing the second result with the first result, we assume that first delta regression benefits SER. In following section, most feature set of experiments do not contain second delta regression coefficients. Here we reduce 1/3 feature set.

D. Reducing Feature Groups

In TABLE I, the baseline set has 7 feature groups. We only have 6 feature groups because we combine voice quality group into pitch group. This new group is called $vpitch$. We experiment with different feature sets which include at most 3

TABLE III
3 EXPERIMENTS IN REDUCING REGRESSION FEATURE. 'O' - ORIGINAL DATA; 'Δ' - DELTA REGRESSION

Feature	# of Features	Accuracy
O+Δ	4368	86.1
O+Δ^2	4368	83.7
O	2184	83.3

feature groups. TABLE IV presents the 18 feature group sets that perform more than 80% accuracy, and TABLE V presents the times of each feature group showing in TABLE IV. In these 18 experiments, 16 feature sets include the MFCC group. However, the times of other 5 feature groups are only 5 to 7. We can assume that MFCC group is more efficient than other 5 feature groups. Because they may be somewhat redundant. Therefore, we run three experiments each only using one feature group to compare these three groups. The results also show MFCC group is better than other two groups. Therefore, we delete Mel-spectrum and spectral groups. Then, we select the feature sets which have at least 83% accuracy and do not include Mel-spectrum and spectral groups from TABLE IV, and we get MFCC+$vpitch$+ZCR, MFCC+energy, and MFCC+ZCR sets. We also add MFCC+energy+$vpitch$+ZCR set to include all feature groups. These four feature group sets are applied to following experiments. In order to reduce more feature, we try to delete some LLD in energy, $vpitch$, and ZCR groups. These 3 groups have 5 LLD: F0, F0 envelope, log energy, voice probability, and ZCR. In order to compare these 5 LLD, we perform 5 experiments with each one. TABLE VI shows these 5 results. We delete F0 and F0 envelope because these 2 LLD are the worst of 5 LLD. The first table of TABLE VII shows our 4 feature group sets with only reducing feature group and LLD; the second table of TABLE VII shows reducing with feature group, LLD and delta regression.

TABLE IV
18 EXPERIMENTS FROM 41 FEATURE GROUP EXPERIMENTS ABOVE 80% ACCUARCY

Feature Group Set	# of Features	Accuracy
MFCC+Mel-spectrum+spectral	3978	85.6
MFCC+Mel-spectrum+$vpitch$	3276	84.4
MFCC+spectral+$vpitch$	2184	83.7
MFCC+Mel-spectrum+ZCR	3120	83.5
MFCC+spectral	1950	83.5
MFCC+ZCR	1092	83.5
MFCC+$vpitch$+ZCR	1326	83.3
MFCC+energy+spectral	2028	83.1
MFCC+spectral+ZCR	2028	83
MFCC+energy	1092	83
MFCC+Mel-spectrum	3042	82.8
MFCC+Mel-spectrum+energy	3120	82.6
MFCC+$vpitch$	1248	82.6
MFCC+energy+$vpitch$	1326	82.4
MFCC+energy+ZCR	1170	82.2
Mel-spectrum+spectral+$vpitch$	3198	81.4
MFCC	1014	81.3
Mel-spectrum+spectral+energy	3042	80.5

TABLE V
THE TIMES OF EACH FEATURE GROUP SHOWING IN TABLE IV

Feature Group	# of in 18 Experiments
MFCC	16
Mel-spectrum	7
spectral	7
energy	6
$vpitch$	6
ZCR	5

TABLE VI
ORDER OF 5 LLD EXPERIMENTS

LLD	Accuracy
log energy	60.2
ZCR	57.6
voice probability	53.6
F0	52.5
F0env	52

TABLE VII
4 FEATURE SETS WITH DIFFERENT REDUCING STEPS

Baseline

Feature Group Set	# of Features	Accuracy
baseline	6552	85.2

Reducing Regression Features

Feature Group Set	# of Features	Accuracy
baseline without Δ^2	4368	86.1

Reducing Feature Groups and LLD
Some results in this table is different with TABLE IV because the $vpitch$ group in TABLE IV have not deleted F0 and F0env LLDs yet.

Feature Group Set	# of Features	Accuracy
MFCC+$vpitch$+ZCR	1170	83.7
MFCC+energy+$vpitch$+ZCR	1248	81.9
MFCC+energy	1092	83
MFCC+ZCR	1092	83.5

Reducing Functionals

Feature Group Set	# of Features	Accuracy
MFCC+$vpitch$+ZCR	180	81.1
MFCC+energy+$vpitch$+ZCR	192	80.7
MFCC+energy	168	79.1
MFCC+ZCR	168	78.7

Reduction via Principal Component Analysis

Feature Group Set	# of Features	Accuracy
MFCC+$vpitch$+ZCR	37	80.2
MFCC+energy+$vpitch$+ZCR	38	78.5
MFCC+energy	34	78.5
MFCC+ZCR	36	79.4

E. Reducing Functionals

Because we integrate quartile 0.25, 0.5, and 0.75 into one functional, and so does inter-quartile range 1-2, 2-3, 1-3, we just perform 35 experiments with each functional. TABLE VIII shows 35 experimental results which sort by accuracy.

We select the best 14 of functionals from TABLE VIII.

TABLE VIII
35 EXPERIMENTS OF FUNTIONALS SORT BY ACCURACY

No.	Functionals, etc.	#	Accuracy
1	quartile 0.25, 0.5, and 0.75	3	82.4
2	linear error between contour and linear regression line	1	79.6
3	arithmetic mean of absolute values	1	78.7
4	quadratic error between contour and linear regression line	1	78.7
5	95 percent percentiles to compute	1	78.5
6	98 percent percentiles to compute	1	78.5
7	standard deviation	1	78.5
8	arithmetic mean of absolute values	1	78.3
9	geometric mean	1	78.3
10	inter-quartile range 1-2, 2-3, 1-3	3	78.1
11	variance	1	77.6
12	quadratic mean	1	77
13	quadratic mean (of non-zero values only)	1	76.8
14	linear error between contour and quadratic regression line (parabola)	1	75.9
15	arithmetic mean of peaks	1	75.3
16	quadratic error between contour and quadratic regression line (parabola)	1	74.2
17	the arithmetic mean of the contour	1	72.7
18	The offset (t) of a linear approximation of the contour	1	71.5
19	arithmetic mean of peaks - arithmetic mean of all values	1	70.6
20	maximum value minus arithmetic mean	1	69.5
21	arithmetic mean - minimum value	1	68.7
22	max-min	1	64.1
23	The slope (m) of a linear approximation of the contour	1	62.8
24	zero-crossing rate	1	60.3
25	number of peaks	1	59.8
26	quadratic regression coefficient 3 (c = offset)	1	59.8
27	The kurtosis (4th order moment)	1	59.4
28	quadratic regression coefficient 2 (b)	1	56.4
29	the skewness (3rd order moment)	1	55.9
30	mean distance between peaks	1	52.7
31	quadratic regression coefficient 1 (a)	1	47.1
32	centroid of contour	1	42.1
33	number of non-zero values	1	36.6
34	the absolute position of the maximum value (in frames)	1	35.5
35	the absolute position of the minimum value (in frames)	1	35.3

After we try to perform different experiments with different functional combination, we choose 3rd, 4th, 5th, 7th, 8th, and 9th functionals to be our optimal functional set. We discard functional No.1 and No.2, which degrade the performance in the functional combination phase. This optimal functional set combine with our 4 feature group sets can get 78.7 to 81.7 accuracy, and the feature numbers are 168 to 192. The 3rd table of TABLE VII shows these 4 experiments.

TABLE IX
DETAIL OF REDUCING NUMBER OF FEATURE STEP BY STEP USING OUR OPTIMAL FEATURE GROUP SET. 'DD' - WITHOUT SECOND ORDER DELTA REGRESSION COEFFICIENT; 'FEA' - OPTIMAL FEATURE GROUP SET; 'FUN' - OPTIMAL FUNCTIONAL SET; 'PCA' - PRINCIPAL COMPONENT ANALYSIS

Feature Set	Accuracy	# of Features	% Reduction
baseline	85.2	6552	0%
dd	86.1	4368	33%
dd+fea	83.7	1170	82.1%
dd+fea+fun	81.1	180	97.2%
dd+fea+fun+PCA	80.2	37	99.4%

F. Reduction via Principal Component Analysis

The last step, we use PCA to select smaller number of components, which keeps at least 95 percent of the variance in the original data. When principle component number is 34, 36, 37 or 38, we get the best results with different feature sets. With PCA, MFCC+$vpitch$+ZCR set gets 80.2% accuracy with 37 features; MFCC+energy+$vpitch$+ZCR set gets 78.5% accuracy with 38 features; MFCC+energy gets 78.5% accuracy with 34 features; and MFCC+ZCR gets 79.4% accuracy with 36 features. Finally, we choose MFCC+$vpitch$+ZCR to be our optimal feature group set because it is the only set that keeps performance above 80% after reduction via PCA. The 4th table of TABLE VII shows 4 experiments. TABLE IX, TABLE X, Fig. 1 adopt our optimal feature group set. TABLE IX shows the accuracy and feature number each step. Fig. 1 shows the relationship between accuracy and feature number. TABLE X shows the confusion matrix. Comparing with our best reducing experiment, we get 77.2% accuracy with only using PCA reducing from 6552 to 37 features.

TABLE X
CONFUSION MATRIX WITH 37 FEATURES USING OUR OPTIMAL FEATURE GROUP SET. 'A' - ANGER; 'B' - BOREDOM; 'D' - DISGUST; 'F' - FEAR; 'H' - HAPPINESS; 'N' - NEUTRAL; 'S' - SADNESS

Reference \ Answer	A	B	D	F	H	S	N	Avg.
Anger(127)	113	0	0	3	10	0	1	88.98
Boredom(81)	1	68	1	0	0	7	4	83.95
Disgust(46)	0	3	41	2	0	0	0	89.13
Fear(69)	8	0	1	53	3	2	2	76.81
Happiness(71)	24	0	0	5	42	0	0	59.15
Sadness(62)	0	11	1	0	0	50	0	80.64
Neutral(79)	3	5	1	4	1	3	62	78.48

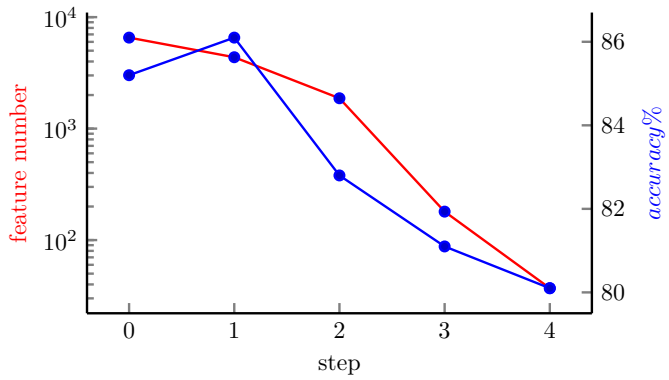

Fig. 1. This figure gives accuracy and feature number step by step. The x axis stands for step of reducing. The left y axis stands for feature number. The right y axis indicates the accuracy of each reducing step.

IV. CONCLUSIONS

In this paper, we reduce the feature set of a baseline speech emotion recognition system. This is achieved by a sequence of steps, including the removal of delta features, the selection of feature groups, the selection of functionals, and finally the application of principal component analysis. The resultant feature set consists of only 37 features, amounting to a size reduction of more than 99% from the original set of 6,552 features. Furthermore, the accuracy is kept above 80%, not severely degraded from the baseline of 85.2%.

In the future, we will experiment this optimal feature set on different corpus to verify that it is general enough on SER. Also, we will test other classifiers in combination of the feature set. From the results, it is curious why the emotion of happiness has the worst accuracy, and we will look into this issue. In summary, we hope to improve the fronted of speech emotion recognition.

REFERENCES

[1] R. W. Picard, *Affective computing*. MIT Press, 1997.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *in IEEE on Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *in IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[4] D. Amelynck, M. Grachten, L. van Noorden, and M. Leman, "Toward e-motion-based music retrieval a study of affective gesture recognition," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 250–259, 2012.

[5] M. Hoque, D. McDuff, and R. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *Affective Computing, IEEE Transactions on*, vol. 3, no. 3, pp. 323–334, 2012.

[6] D. McDuff, R. Kaliouby, and R. Picard, "Crowdsourcing facial responses to online videos," *Affective Computing, IEEE Transactions on*, vol. 3, no. 4, pp. 456–468, 2012.

[7] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[8] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of Interspeech 2009*. ISCA, 2009, pp. 312–315.

[9] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proceedings of IEEE Workshop on of Automatic Speech Recognition Understanding (ASRU)*, 2009, pp. 552–557.

[10] T. Iliou and C.-N. Anagnostopoulos, "SVM-MLP-PNN classifiers on speech emotion recognition field - a comparative study," in *Proceedings of International Conference on Digital Telecommunications (ICDT)*, 2010, pp. 1–6.

[11] A. Ivanov and G. Riccardi, "Kolmogorov-Smirnov test for feature selection in emotion recognition from speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5125–5128.

[12] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *in IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 116–125, 2012.

[13] S. Ananthakrishnan, A. Vembu, and R. Prasad, "Model-based parametric features for emotion recognition from speech," in *in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 529–534.

[14] C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5692–5695.

[15] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit - searching for the most important feature types signalling emotion-related user states in speech." *in Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, 2011.

[16] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[17] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of Interspeech 2005*. ISCA, 2005, pp. 1517–1520.

[18] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods — Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 185–208.

[19] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with java implementations," 1999.

[20] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR introducing the munich open-source emotion and affect recognition toolkit," in *Proceedings of ACII International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.