VOL. 9, NO. 8, AUGUST 2014 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

SPEECH EMOTION RECOGNITION USING STATIONARY WAVELET TRANSFORM AND TIMBRAL TEXTURE FEATURES

M. Hariharan, Sazali Yaacob, M. N. Hasrul and Oung Qi Wei
School of Mechatronic Engineering, Universiti Malaysia Perlis (UniMAP), Campus Pauh Putra, Perlis, Malaysia
E-Mail: hari@unimap.edu.my

ABSTRACT

Nowadays, researchers are paying more attention to recognize the human emotions from various modalities such as facial features, physiological or biological signals and speech signals. In the last decade, a number of research works have been carried out on emotion recognition using speech signals. In this work, emotion recognition system was developed using the features extracted from the emotional speech signals (ES) and its glottal waveforms (GW). Inverse filtering and linear predictive analysis were used to derive GWs from the speech signals. ES and GW were decomposed into five levels using stationary wavelet transform (SWT) and timbral texture features were extracted from the decomposed SWT coefficients. A total of 288 features were extracted from ES and GW respectively. Two-stage feature reduction was proposed to reduce the number of features and also to enrich the discriminatory power of the extracted features. The raw and enhanced features were used as input for extreme learning machine (ELM) and k-nearest neighbor (kNN) based classifiers. Several experiments were carried out and the results shows that timbral texture features derived from the decomposed stationary wavelet coefficients can be used as possible cues for emotion differentiation.

Keywords: speech signals, glottal waveforms, timbral texture, feature selection/reduction, emotion recognition.

INTRODUCTION

In the field of human - machine interface (HMI) and human-to-human communication (HHC), emotion recognition plays an important role. Accurate recognition of human emotions may enhance the HMI and HHC [1-3]. Different characteristics of the speaker like gender, language used, socio-economic back ground, speaker health and emotional states can be observed from their speech. Generally, linguistic information is used for textual content recognition and paralinguistic information is used for emotion or speaker recognition. Emotion recognition from speech is nothing but a process of extraction of acoustic parameters from the speech signal and linguistic parameters from the transcriptions of the utterances of the corpus [2-5]. Researchers have proposed several parameterization methods in the field of emotion recognition from speech, however it is not clear which speech features are best in distinguishing between emotions. Researchers have developed different emotional speech databases for emotion recognition [2-4, 6]. Many feature extraction, feature selection or reduction and classification algorithms have been proposed for the past thirty years [2-4, 6]. However, few researchers have focused on the development of spoken emotion recognition system using glottal-based features [5, 7-13].

In this work, a new speech emotion recognition method was presented based on the features extracted from the speech signals and its glottal waveforms. Stationary wavelet transform based timbral texture features were proposed as the potential features for characterizing the emotions from the speech utterances. In all pattern recognition problems, curse of dimensionality is a challenging issue. In order to overcome the curse of dimensionality, two stage feature reduction was suggested to reduce the number of features and to enhance the discriminatory power the features as well. The raw and enriched features were used as input for ELM and kNN for improved speech emotion recognition.

EMOTIONAL SPEECH DATABASE

In this work, Berlin emotional speech database (BES) [14] was used for speech emotion recognition experiments. 10 professional actor/actresses were used to simulate 7 emotions (anger, boredom, disgust, anxiety, happiness, sadness and neutral) which include 5 male and 5 female. All the emotional speech samples were recorded with a sampling frequency of 16 kHz at 16 bit resolution. It consists of 535 emotional speech utterances in German, 233 of which were collected from male speakers, whereas the remaining 302 ones were collected from female speakers. In this work, all the emotional speech samples were downsampled into 8 kHz and used for further analysis.

FEATURE EXTRACTION USING SWT

Feature extraction is a process of extracting most discriminative features that can be used to characterize emotional speech signals. Several feature extraction methods have been proposed in speech emotion recognition applications. In the last decade, wavelet analysis has been successfully applied in speech emotion recognition and also in many other signal and image processing applications [15, 16]. The main advantage of wavelet analysis compared to Fourier analysis is its ability to analyze the both time and frequency information simultaneously within the signal. Wavelets transforms can be categorized into either continuous wavelet transform (CWT) or discrete wavelet transforms (DWT). Stationary wavelet transform is one of the types of discrete wavelet transforms. DWT is not shift-invariant and this is the main advantage of SWT [17, 18]. Shift-invariant can be achieved by removing the up and downsamplers in the

DWT analysis and synthesis process. The absence of up and downsamplers leads each subband contains the same number of samples as the input. Though there is redundant information at each decomposition level, the SWT is useful in various signal and image processing applications [17].

All the emotional speech samples from BES database were downsampled to 8 kHz and the unvoiced portions between words were removed by segmenting the downsampled emotional speech signals into frames with a length of 32 ms (256 samples) with non-overlap based on the energy of the frames. Frames with low energy were discarded and the rest of the voiced portions were concatenated and used for feature extraction. In this work, SWT and timbral texture based features were derived from speech signals and its glottal waveforms. To extract the glottal and vocal tract characteristics from the speech signals, several techniques have been proposed. In this work, we extracted the glottal waveforms from the emotional speech signals by using inverse filtering and linear predictive analysis [19, 20]. The emotional speech signals (only voiced portions) and its glottal waveforms were passed through a first order low pass filter. The purpose of this process is to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing [21]. The first order pre-emphasis filter is defined as

$$H(z)=1-a*z^{-1} \quad 0.9 \le a \le 1.0 \tag{1}$$

The commonly used a value is 15/16 = 0.9375 or 0.95 [21]. In this work, the value of a is set equal to 0.9375. Then the pre-emphasized emotional speech signals and glottal waveforms were decomposed into five levels using SWT with four different orders of Daubechies wavelets (db3, db6, db10 and db44). In this work, Daubechies wavelet has been chosen due to the following properties [22]: Time invariance, fast computation and sharp filter transition bands. Generally, timbral texture features were proposed for music-speech discrimination and speech recognition. But, here we investigated the efficacy of the timbral texture features for speech based emotion recognition applications. Energy entropy, short-time energy, zero-crossing rate, spectral rolloff, spectral centriod and spectral flux were extracted from the decomposed stationary wavelet coefficients (CA5, CD5, CD4, CD3, CD2 and CD1). After obtaining the timbral texture features for each decomposed stationary wavelet coefficients, the following statistical parameters were computed such as standard deviation of timbral texture features, maximum by standard deviation of timbral texture features, maximum by median of timbral texture features, square of standard deviation by square of mean of timbral texture features. A total of 288 features (6 timbral texture features x 4 statistical features x 6 subbands = 144 for each emotional speech signals + 6 timbral texture features x 4 statistical features x 6 subbands = 144 for each glottal waveforms) were derived for each emotional speech signals and its glottal waveforms after SWT decomposition. The detailed mathematical expressions for the proposed features can be found from [23, 24].

TWO STAGE FEATURE REDUCTION

In this work, 288 features were extracted from emotional speech signals and its glottal waveforms. To avoid over fitting during training of the classifiers and to improve their performance, feature selection/reduction should be applied to discard the redundant features and to select the most discriminative features. Several feature reduction/selection methods have been proposed in speech emotion recognition applications. In this work, two-stage feature selection/reduction was applied in which a filter method (information gain, IG) was used followed by a projection based feature reduction (linear discriminant analysis, LDA) was employed to select and reduce the original dimensionality of the extracted features.

Information gain (Stage-1)

IG is a feature ranking method and it measures the expected reduction in entropy [25, 26]. From the entropy measure, it can be observed that how the whole system is related to an attribute. The formula used to compute IG is:

$$IG(C|E) = H(C) - H(C|E) \tag{2}$$

where IG (C|E) is the information gain of the feature E, H(C) is the system's entropy and H (C|E) is the system's relative entropy when the value of the feature E is known. This measure IG value was calculated for each feature. After calculating IG values for all features, a threshold was established to select the most discerning features. In this work, a threshold 0 was used [25, 26]. The features with IG value more than 0 were selected and the remaining features were not used.

Linear discriminant analysis (Stage-2)

After selecting more discriminative features using IG, LDA was applied to reduce the dimension further. LDA utilizes supervised projection method and is to maximize the ratio of the between and within class scatters of the feature set as shown in the following equation [27-29].

$$Y_{opt} = \arg\max_{y} \frac{\left|Y^T S_b Y\right|}{\left|Y^T S_w Y\right|} \tag{3}$$

Where S_b is the between-class scatter matrix and S_w is the within-class scatter matrix. Using the transformation matrix Y, the between-class scattering was maximized whereas the within-class scattering was minimized and the main objective of LDA is to reduce the dimension while preserving as much of the class discriminatory information as possible. After applying LDA, only c-1 features which were projected with non-

zero eigenvalues were used as features based on the criterion given in the Equation (3).

CLASSIFIERS

K-Nearest neighbor

kNN classifier is a type of instance based learning technique and predicts the class of a new test data based on the closest training examples in the feature space [30]. Euclidean distance was used as distance measurement. We determined the suitable k-value as 6 based on the experimental investigations.

Extreme learning machine (ELM)

Extreme Learning Machine (ELM) was used in this study to perform multiclass emotion recognition. The weights between the input neurons and the hidden neurons in ELM were randomly assigned based on some continuous probability density function while the weights between the hidden layer and the output of the probability density function while the weights between the hidden layer and the output of the single layer feed forward network was determined analytically [31, 32]. In this study, a radial basis activation function was used.

RESULTS

A total 288 statistical features of timbral texture features were extracted from all the emotional speech signals and its glottal waveforms. Two-stage feature reduction using IG and LDA was applied to reduce the original dimensionality of the features. The dimensionality reduced features were fed to kNN and ELM classifiers for emotion recognition. Three different experiments were carried out. **Experiment-1:** emotion recognition using all the features (288 features), using the selected features after stage 1 and using the reduced features after stage-2. **Experiment-2:** emotion recognition using the features extracted from emotional speech signals (144 features), using the selected features after stage-1 and using the reduced features after stage-2. **Experiment-3:** emotion recognition using the features extracted from glottal waveforms (144 features), using the selected features after stage 1 and using the reduced features after stage-2. Table-1 shows the number of selected features after stage-1 and stage-2 in all the experiments.

Two performance measures were used and reported as the results of all the experiments such as average emotion recognition rates (AERR) and geometric mean (G-mean). G-mean was used as performance as it can clearly represent a tradeoff between the recognition rates on all the classes of emotion. 10-fold cross validation scheme was used in which, the proposed feature set was divided into 10 disjoint sets and training was repeated for 10 times. Average emotion recognition rates and G-mean were recorded in all the experiments and reported in Tables 2-4. Table-2 shows the results of all the extracted before two-stage feature reduction. Best average emotion recognition rate of 75.07±0.64 % (db10) and 66.60±0.55% (db44) were obtained using ELM and kNN classifiers respectively with all 288 features. Though, we obtained best average emotion recognition rate using all 288 features, all the features were not useful or relevant. Hence, two-stage feature reduction was applied to select the best features and to discard the irrelevant features. First, IG based feature selection method was used to select best features from all 288 features, 144 speech features and 144 glottal features. After stage-1, the average emotion recognition rates was recorded and tabulated in Table-3.

Table-1. Number of selected features after stage-1 and stage-2.

Different orders of daubechies wavelets	Speech + Glottal	Stage 1	Stage 2	Speech	Stage 1	Stage 2	Glottal	Stage 1	Stage 2
db3	288	185	6	144	97	6	144	88	6
db6	288	204	6	144	108	6	144	96	6
db10	288	213	6	144	117	6	144	96	6
db44	288	212	6	144	117	6	144	95	6

Table-2. Average emotion recognition rates using 10-fold cross validation.

Classifiers	Different order of daubechies wavelets	Statistics	Speech + Glottal		Speech features		Glottal features	
			AERR	G-mean	AERR	G-mean	AERR	G-mean
ELM	db3	Mean	72.32	67.48	71.44	67.65	65.10	60.92
		Std	0.55	0.76	1.01	1.18	0.77	0.78
	db6	Mean	72.67	68.07	72.99	69.66	66.82	61.68
		Std	0.86	1.06	0.95	1.22	1.00	1.44
	db10	Mean	75.07	72.12	73.93	71.26	65.18	60.45
		Std	0.64	0.64	1.07	1.11	0.73	0.76
	db44	Mean	74.90	70.78	74.47	72.04	67.23	62.43
		Std	0.76	0.81	0.94	0.99	1.00	1.14
kNN	db3	Mean	62.28	54.69	60.06	52.45	54.56	46.93
		Std	1.00	1.32	1.34	1.27	0.98	1.24
	db6	Mean	62.88	57.67	59.10	52.79	55.64	48.18
		Std	1.13	1.57	0.93	1.44	0.95	1.00
	db10	Mean	65.16	61.56	62.07	58.48	53.29	48.34
		Std	1.07	1.04	1.42	1.53	0.99	1.03
	db44	Mean	66.60	62.50	63.87	60.77	55.35	50.27
		Std	0.55	0.77	0.90	1.20	1.25	1.29

Table-3. Average emotion recognition rates using 10-fold cross validation after stage-1.

Classifiers	Different order of daubechies wavelets	Statistics	Speech + Glottal		Speech features		Glottal features	
			AERR	G-mean	AERR	G-mean	AERR	G-mean
ELM	db3	Mean	74.77	71.64	71.79	68.18	64.80	60.57
		Std	1.09	1.19	0.94	1.05	0.57	0.72
	db6	Mean	75.50	72.34	72.79	69.42	67.42	62.45
		Std	0.76	0.98	0.80	0.90	0.99	1.05
	db10	Mean	77.46	75.11	74.24	71.88	64.79	59.68
		Std	0.59	0.72	0.92	1.12	0.52	0.87
	db44	Mean	75.79	72.86	74.65	72.25	67.44	62.46
		Std	0.74	0.81	0.60	0.49	1.00	1.12
kNN	db3	Mean	62.79	54.64	60.28	53.01	53.93	46.07
		Std	1.08	1.27	0.63	0.76	0.88	1.04
	db6	Mean	66.49	61.34	59.14	52.71	55.91	48.41
		Std	0.56	0.97	0.99	1.05	0.79	1.22
	db10	Mean	65.53	62.85	62.36	58.75	53.21	47.81
		Std	0.75	0.61	0.68	0.82	1.05	1.35
	db44	Mean	68.50	65.95	63.79	60.74	55.16	49.94
		Std	0.96	1.08	1.22	1.29	0.95	1.12

From the Table-3, it can be seen that the average emotion recognition rate was improved to 77.46±0.59% (db10) using 213 features only with ELM classifier. Using 177 speech features, best average recognition rate of

74.65±0.60% was obtained with ELM classifier features. Features extracted from glottal waveforms always yielded below 70% of average emotion recognition rates before and after stage1 feature selection. The performance of emotion recognition system can still be improved by applying 2nd stage feature selection/reduction. In this work, LDA was employed after stage 1 feature selection to improve the emotion recognition further. After stage-2, average emotion recognition rates were recorded and tabulated in Table-4. From the Table-4, it can be observed that the average recognition rates of above 96% were obtained using both the classifiers (ELM and kNN) using only 6 features.

Table-4. Average emotion recognition rates using 10-fold cross validation after stage-2.

Classifiers	Different order of daubechies wavelets	Statistics	Speech + Glottal		Speech features		Glottal features	
			AERR	G-mean	AERR	G-mean	AERR	G-mean
ELM	db3	Mean	96.41	96.18	88.71	87.07	82.64	81.48
		Std	0.17	0.16	0.39	0.41	0.24	0.33
	db6	Mean	97.78	97.75	89.89	89.36	84.11	82.98
		Std	0.16	0.16	0.21	0.20	0.47	0.53
	db10	Mean	97.93	97.71	91.51	91.25	84.92	83.87
		Std	0.16	0.17	0.35	0.42	0.38	0.40
	db44	Mean	97.96	97.79	90.06	90.01	86.65	85.69
		Std	0.22	0.22	0.39	0.35	0.37	0.52
kNN	db3	Mean	96.22	96.10	88.60	87.43	81.36	80.88
		Std	0.17	0.16	0.51	0.56	0.69	0.71
	db6	Mean	98.04	98.06	88.22	87.24	81.48	79.87
		Std	0.24	0.26	0.70	0.62	1.06	1.19
	db10	Mean	98.15	98.08	89.87	89.20	82.93	81.76
		Std	0.26	0.26	0.36	0.44	0.38	0.42
	db44	Mean	97.61	97.49	89.08	89.05	84.21	82.64
		Std	0.21	0.26	0.43	0.44	0.58	0.66

DISCUSSIONS

Ling He *et al*., have proposed wavelet packet energy entropy features for emotion recognition from speech and glottal signals with GMM classifiers [5]. They achieved average emotion rates for BES database between 51% and 54%. Wavelet packet based adaptive filter-bank construction method was proposed in [33]. Additive Fisher ratio was used as wavelet packet tree pruning criterion. Most discriminative wavelet packet based normalized energy features were extracted and GMM was employed as classifier. They have obtained maximum recognition rate of 75.64% with Daubechies filter order of 40 and 21 normalized wavelet filter bank energies. Linear predictive cepstral coefficients (LPCCs), Mel-frequency cepstral coefficients (MFCCs) and formant features were used to represent the vocal tract system characteristics accurately [34]. These vocal tract features were used for speech emotion recognition and they have achieved a maximum emotion recognition rate of 68% with LPCCs+formant features. Ali Shahzadi *et al.* have proposed non-linear dynamics features (NLDs) for speech emotion recognition [35]. They have achieved overall recognition rates between 82% and 86% using NLDs+prosodic+spectral features. In [36], MPEG-7 low level audio descriptors were used to model the seven emotional categories available in the BES database. Support vector machine (SVM) with radial basis function kernel was used as classifier and achieved 77.88%. Margarita Kotti and Fabio Paterno [37] have proposed a psychologically-inspired binary cascade classification scheme for speech based emotion recognition using low level audio descriptors and high level perceptual descriptors with Linear SVM. The best emotion recognition accuracy of 87.7% was obtained using SVM with linear kernel. In [38], a new set of acoustic features based on the perceptual quality metrics are proposed for the binary arousal and valence discrimination which include partial loudness of the emotional difference, emotional difference-to-perceptual mask ratio, measures of alterations of temporal envelopes, measures of harmonics of the emotional difference etc. They had not reported the results for seven classes of emotions discrimination. In [39], prosodic features, spectral features, glottal flow features, AM-FM features were utilized and two-stage feature reduction was proposed for speech emotion recognition. The overall emotion recognition rates of 85.18% for gender dependent and 80.09% for gender independent was achieved using SVM classifier. From the previous work, we can observe that the several feature extraction, feature selection and classification algorithms have been proposed for speech based emotion recognition.

Although all the above works are novel contributions to the field of speech emotion recognition, it is difficult to compare them directly since the presentation and computation of the results are not consistent. In this work, we have proposed new acoustic features based on the SWT with timbral texture features for speech emotion recognition. We have conducted several experiments and

presented our results. From the simulation results, it can be observed that the proposed SWT based timbral texture features with two-stage reduction has yielded a maximum of accuracy between 96.10% and 98.15% for gender dependent speech based emotion recognition.

CONCLUSIONS

This paper addresses the speech emotion recognition problem by proposing SWT based timbral texture features with two-stage feature reduction. The speech signals from BES database were used. Glottal waveforms were derived using inverse filtering and linear prediction analysis. Features were extracted from both emotional speech signals and its glottal waveforms. Two-stage feature reduction in which information gain based feature selection was used to select the best features followed by linear discriminant analysis based feature reduction was utilized to reduce the dimension further. kNN and ELM kernel were employed for speech emotion recognition. The proposed method has provided a maximum of accuracy between 96.10% and 98.15% for gender dependent speech based emotion recognition with only 6 features. In the future, the proposed method will be tested with other emotional speech databases and also in speaker-independent emotion recognition environment.

ACKNOWLEDGEMENTS

This research is supported by Fundamental Research Grant Scheme (FRGS), Malaysia [Grant No: 9003-00297] and Journal Incentive Research Grant, UniMAP [Grant No: 9007-00071].

REFERENCES

[1] R. Cowie, *et al*. 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine. 18: 32-80.

[2] D. Ververidis and C. Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. Speech communication. 48: 1162-1181.

[3] M. El Ayadi, M. S. Kamel and F. Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 44: 572-587.

[4] S. G. Koolagudi and K. S. Rao. 2012. Emotion recognition from speech: a review. International journal of speech technology. 15: 99-117.

[5] L. He, *et al*. 2013. Study of wavelet packet energy entropy for emotion classification in speech and glottal signals. In: 5th International Conference on Digital Image Processing. pp. 887834-887834-6.

[6] C. M. Lee and S. S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing. 13: 293-303.

[7] K. E. Cummings and M. A. Clements. 1992. Improvements to and applications of analysis of stressed speech using glottal waveforms. In: Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on. pp. 25-28.

[8] K. E. Cummings and M. A. Clements. 1993. Application of the analysis of glottal excitation of stressed speech to speaking style modification. In: Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on. pp. 207-210.

[9] K. E. Cummings and M. A. Clements. 1995. Analysis of the glottal excitation of emotionally styled and stressed speech. The Journal of the Acoustical Society of America. 98: 88-98.

[10] E. Moore, M. Clements, J. Peifer and L. Weisser. 2003. Investigating the role of glottal features in classifying clinical depression. In: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 2849-2852.

[11] E. Moore, M. A. Clements, J. W. Peifer and L. Weisser. 2008. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Transactions on Biomedical Engineering. 55: 96-107.

[12] A. I. Iliev, M. S. Scordilis, J. P. Papa and A. X. Falcão. 2010. Spoken emotion recognition through optimum-path forest classification using glottal features. Computer Speech and Language. 24: 445-460.

[13] Alexander and S. Michael. 2011. Spoken Emotion Recognition Using Glottal Symmetry. EURASIP Journal on Advances in Signal Processing. pp. 1-11.

[14] F. Burkhardt, *et al*. 2005. A database of German emotional speech. In: Interspeech, Lisbon, Portugal. pp. 1517-1520.

[15] M. Hariharan, *et al*. 2013. Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy. Digital Signal Processing: A Review Journal. 23: 952-959.

[16] M. Hariharana, K. Polatb, R. Sindhuc and S. Yaacoba. 2013. A hybrid expert system approach for telemonitoring of vocal fold pathology. Applied Soft Computing Journal. 13: 4148-4161.

[17] H. Huang, N. Nguyen, S. Oraintara and A. Vo. 2008. Array CGH data modeling and smoothing in Stationary Wavelet Packet Transform domain. BMC genomics. 9: S17.

[18] H. Keskes, A. Braham and Z. Lachiri. 2013. Broken rotor bar diagnosis in induction machines through stationary wavelet packet transform and multiclass wavelet SVM. Electric Power Systems Research. 97: 151-157.

[19] D. E. Veeneman and S. BeMent. 1985. Automatic glottal inverse filtering from speech and electroglottographic signals. IEEE Transactions on Acoustics, Speech and Signal Processing. 33: 369-377.

[20] D. Wong, J. Markel and A. Gray Jr. 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Transactions on Acoustics, Speech and Signal Processing. 27: 350-355.

[21] L. R. Rabiner and B.-H. Juang. 1993. Fundamentals of speech recognition. vol. 14. PTR Prentice Hall Englewood Cliffs.

[22] A. Cohen, I. Daubechies and J. C. Feauveau. 1992. Biorthogonal bases of compactly supported wavelets. Communications on Pure and Applied Mathematics (Wiley Subscription Services, Inc., A Wiley Company New York). 45: 485-560.

[23] G. Tzanetakis and P. Cook. 2002. Musical genre classification of audio signals. Speech and Audio Processing, IEEE Transactions on. 10: 293-302.

[24] Y. Lavner and D. Ruinskiy. 2009. A decision-tree-based algorithm for speech/music classification and segmentation. EURASIP Journal on Audio, Speech, and Music Processing. 2009: 1-14.

[25] M. T. Martín-Valdivia, M. C. Díaz-Galiano, A. Montejo-Raez and L. Ureña-López. 2008. Using information gain to improve multi-modal information retrieval systems. Information processing and management. 44: 1146-1158.

[26] C.-H. Yang, L.-Y. Chuang and C.-H. Yang. 2010. IG-GA: a hybrid filter/wrapper method for feature selection of microarray data. Journal of Medical and Biological Engineering. 30: 23-28.

[27] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 19: 711-720.

[28] M. Hariharan, *et al*. 2012. A comparative study of wavelet families for classification of wrist motions. Computers and Electrical Engineering. 38: 1798-1807.

[29] F. Tang and H. Tao. 2007. Fast linear discriminant analysis using binary bases. Pattern recognition letters. 28: 2209-2218.

[30] R. O. Duda, P. E. Hart and D. G. 2012. Stork, Pattern classification. John Wiley and Sons.

[31] G.-B. Huang, H. Zhou, X. Ding and R. Zhang. 2012. Extreme learning machine for regression and multiclass classification. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on. 42: 513-529.

[32] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew. 2006. Extreme learning machine: theory and applications. Neurocomputing. 70: 489-501.

[33] Y. Li, G. Zhang and Y. Huang. 2013. Adaptive Wavelet Packet Filter-Bank Based Acoustic Feature for Speech Emotion Recognition. In: Proceedings of 2013 Chinese Intelligent Automation Conference. pp. 359-366.

[34] S. R. Krothapalli and S. G. Koolagudi. 2013. Emotion Recognition Using Vocal Tract Information. In: Emotion Recognition using Speech Features. Springer. pp. 67-78.

[35] A. Shahzadi, A. Ahmadyfard, A. Harimi And K. Yaghmaie. 2013. Speech emotion recognition using non-linear dynamics features. Turkish Journal of Electrical Engineering and Computer Sciences.

[36] S. Lampropoulos and G. A. Tsihrintzis. 2012. Evaluation of MPEG-7 Descriptors for Speech Emotional Recognition. In: 2012 8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). pp. 98-101.

[37] M. Kotti and F. Paternò. 2012. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. International journal of speech technology. 15: 131-150.

[38] M. C. Sezgin, B. Gunsel and G. K. Kurt. 2012. Perceptual audio features for emotion detection. EURASIP Journal on Audio, Speech, and Music Processing. 2012: 1-21.

[39] P. Giannoulis and G. Potamianos. 2012. A hierarchical approach with feature selection for emotion recognition from speech. In: LREC. pp. 1203-1206.