Speech Emotion Classification

Using SVM and MLP on Prosodic and Voice Quality Features

Inshirah Idris
Computer Science Department
Sudan University of Science and Technology
Khartoum, Sudan
inshirah15@hotmail.com

Md Sah Hj Salam
Software Engineering Department
Universiti Teknologi Malaysia(UTM)
Skudai, Johor Malasia
sah@utm.my

Mohd Shahrizal Sunar
UTM-IRDA Digital Media Center
Universiti Teknologi Malaysia(UTM)
shahrizal@utm.my

***Abstract*—In this paper comparisons of emotion classification between Support Vector Machine (SVM) and Multi Layer Perceptron (MLP) Neural Network using prosodic and voice quality features extracted from Berlin Emotional Database are reported. The features were extracted using PRAAT tools while WEKA tool was used for classification. Different parameters set up for both SVM and MLP were implemented in getting the optimized emotion classification. The results show that MLP overcomes SVM in overall emotion classification. Nevertheless, the training for SVM was much faster compared to MLP. The overall recognition rate was (76.82%) for SVM and (78.69%) for MLP. Sadness was the highest emotion recognized by MLP with recognition rate of (89.0%) while anger was the highest emotion recognized by SVM with recognition rate of (87.4%). The most confusing emotion using MLP classification were happiness and fear while for SVM, the most confusing emotions were disgust and fear.**

Keywords- Emotion Recognition; SMO;SVM; MLP; Prosodic Features;Voice Quality Featuers.

I. INTRODUCTION

There is a major different between how human and machine understands speech. Humans understand speech via perception of all action from the speaker, including hand gesture, eye movement and the speech emotions while this is the case for machine.

Speech emotion recognition (SER) is a technology aim to identify the emotional or physical state of a speaker from his speech signal. It has attracted many researchers at the present time due to its important in many applications such as: E-Learning, Security, Healthcare, Automatic Translation Systems, and Robotic.

Speech emotion recognition can be divided into three different approaches: Data-based, Feature-based and Classifier-based.

Data-based concentrate in creating or searching for the best speech emotional database that could be used in testing or investigating speech emotion recognition systems. Some researchers use standard databases that are publicly available as in [1], and others create their own dataset as in [2].

Feature-based approach aims to extract and select the best speech features that can optimize the speech emotion recognition performance. Based on literatures, many types of emotional speech features were used. Some researchers worked on extracting one type of speech features as in [3] and others use two or more types of features and proposed new features [4]. There were also researchers who cater issue in features selection [5].

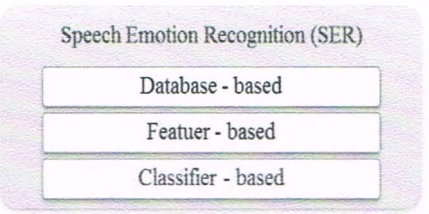

Figure 1: Speech Emotion Approaches

The classification-based approaches focus on selecting or designing a classifier that determine actual mapping between the emotions [6]. Finding appropriate classification algorithms is the most difficult problem in this area. Several types of classifiers were used such as Hidden Markov Model (HMM), k-nearest neighbors (KNN), Artificial Neural Network (ANN), Gaussian Mixtures Model (GMM) and Support Vector Machine (SVM).

The most popular classifiers in speech emotion recognition are Support Vector Machine (SVM) [7], and Artificial Neural Network (ANN) [8].

Artificial Neural network (ANN) can be categorized into their main basic types: multilayer perceptron (MLP), recurrent neural networks (RNN), and radial basis functions (RBF) networks. The latter is rarely used in speech emotion recognition [9].

A multilayer perceptron (MLP) is a feed forward artificial neural network of back-propagation learning rule [10]. It is commonly used in speech emotion recognition due to the simplicity of implementations [9], [11].

On the other hand Support Vector Machine (SVM) is binary classifier which is usually used for classifications and regression purposes [12], [13]. SVM basically can handle only two class problems [14], [15]. It shows good performance with limited data [16] that has many features [17]. SVM classifiers are widely used in many pattern recognition applications and shown to outperform other well-known classifiers [9].

There has been no agreement on which classifier is the most suitable for emotion classifications, because each

classifier has its own advantages and limitations. In this paper we compared between SVM and MLP classifiers in terms of emotion classification accuracy of speaker independent and the time to build the model, using prosodic and voice quality features.

This paper is organized as follows: Section 2 reviews the related works. Section 3 describes our experimental setup. Section 4 shows the classification results and discussion. Lastly, Section 7 is the conclusion.

II. RELATED WORKS

Recently, many studies in finding suitable classifier for speech emotion were conducted. Lee et. al investigated the accuracy rate of support vector machine classifiers using a prosodic feature related to pitch and speech rates. The accuracy rate was (55.68%) [18]. Similarly, using prosodic features extracted from the NATURAL data set, Morrison et al. used find the accuracy of different classifiers. The top performers are the SVM (RBF) (76.93%), then Multi-layer perceptron (74.25%), finally SVM (polynomial) (69.50%) [19].

Javidi and Roshan used SVM and NN with 68 features related to pitch, energy, ZCR, power, and MFCC from Berlins database to detect seven emotions (anger, happiness, fear, sadness, disgust, boredom, and neutral). The average recognition rates for NN is to (39.41%) and for SVM (53.22%) [20]. Using the same database, Ayadi et.al compared his proposed classifier, Gausian mixture autoregressive model with HMM, KNN and NN. The result showed that, the proposed technique provides a classification accuracy of (76%) versus (71%) for the hidden Markov model, (67%) for the k-nearest neighbors, and (55%) for feed-forward neural networks.

Fersini et al. investigates the accuracy of emotion recognition of different classifiers using different data sets. They found that best classifier was Support Vector Machines. Berlin German corpus with linear kernel (E=1) and complexity parameter C=2 give accuracy of (59.3%). Polish corpus with linear kernel (E=1) and complexity parameter C=2 give accuracy of (68.7%). Italian corpus (acted emotions) with linear kernel (E=1) and complexity parameter C=3 give accuracy of (56.5%). Italian corpus (real emotions) with linear kernel (E=1) and complexity parameter C=6 give accuracy of (82.9%).

It can be noticed from previous works that the emotion classification performance varies from each other depending on the features and classifier used. The classification rate varied from 40% to 80%. SVM seemed to surpass NN in emotion classification. Nevertheless, the result depended on experimental set up, database used and parameters chosen.

III. EXPERIMENTAL SETUP

A. Dataset

From the available literatures, there are three types of databases used for analytical study of speech emotions. They are acted, real and elicited emotional speech database. The three types of databases mentioned serve different purposes. The first type is suitable for theoretical researches while the second and third type can be a good baseline for creating real-life applications for a specific industry.

In this work, the Berlin Emotional Acted. Database (EMO-DB) was selected as the database for benchmarking. The database was is easily available and used by many researchers.

EMO-DB is acted German speech emotional database which recorded at the Department of Acoustic Technology of Technical University of Berlin in Germany (funded by the German Research Community). It was recorded using a Sennheiser microphone at a sampling frequency of 16 kHz. Ten professional actors (five male and five female) asked to simulate seven emotions (anger, boredom, disgust, fear, happiness, sadness and neutral) for ten utterances (five short and five longer sentences) that can be used in daily communication and can also be applied in all the emotions. The total utterances are about 800 (seven emotions * ten actors * ten sentences + some second versions).

Twenty judges listened to the utterances in random order in front of a computer monitor. They were allowed to listen to each sample only once before they had to decide in which emotional state the speaker had been. After selection, the database contained a total of 535 speech files [21].

A web interface was developed to present the database of emotional speech. All the available information of the speech database can be accessed via the internet: http://www.expressive-speech.net/emoDB/.

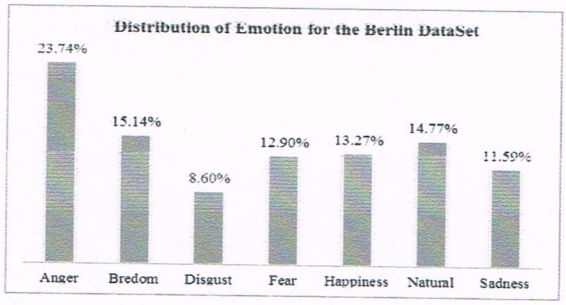

Figure 2: Distribution of Emotion for the Berlin Dataset

B. Features

To extract the features from the speech samples the data mining tools Praat was employed. Praat toolkit is a free scientific computer software package for the analysis of speech in phonetics. In this research Praat was used to extract a set of 80 prosodic and voice quality features.

The prosodic features are known as the primary indicator of the speakers emotional states [22]. These features are extracted at the utterance level. They include formants, energy, pitch, intensity, duration of speech segment [23] and their derivatives [13]. Prosodic features are the most commonly used features in speech emotion recognition, because they provide reliable indication of the emotion [9].

In contrast voice quality features are the less frequently used features [24]. However, studies proved that voice quality features complement prosodic features [4]. Voice quality

features are jitter, shimmer, harmonic to noise ratio (HNR), noise to harmonic ration (NHR), and autocorrelation. The above mentioned features are as follow:

1) Prosodic Features:

- Pitch: minimum, maximum, mean, median, standard deviation, time of minimum, time of maximum, first quartiles, third quartile, mean Slop.
- Energy: minimum, maximum, mean, standard deviation, variance, range.
- Intensity: minimum, time of minimum, maximum, time of maximum, first quartile, third quartile, mean, standard deviation.
- Duration: time of utterance.
- Formant: position of minimum of first formant, maximum of first formant, position of maximum first formant, mean of first formant, first quartile of first formant, third quartile of first formant, first formant bandwidth, minimum of second formant, position of minimum of second formant, maximum of second formant, position of maximum second formant, mean of second formant, first quartile of second formant, third quartile of second formant, second formant bandwidth, difference of mean of second and first formant, minimum of third formant, position of minimum of third formant, maximum of third formant, position of maximum third formant, mean of third formant, first quartile of third formant, third quartile of third formant, third formant bandwidth, difference of mean of third and second formant, minimum of fourth formant, position of minimum of fourth formant, maximum of fourth formant, position of maximum fourth formant, mean of fourth formant, first quartile of fourth formant, third quartile of fourth formant, difference of mean of fourth and third formant, minimum of fifth formant, position of minimum of fifth formant, maximum of fifth formant, position of maximum fifth formant, mean of fifth formant, first quartile of fifth formant, third quartile of fifth formant, difference of mean of fifth and fourth formant.

2) Voice Quality Feature:

- Jitter: local, absolute, rap, ppq5, ddp.
- Mean harmonic to noise ratio (HNR).
- Mean noise to harmonic ration (NHR).
- Mean Autocorrelation.

C. Classification

SVM classifiers are mainly based on the use of kernel functions to nonlinearly map the original features to a high dimensional space where data can be well classified using a linear classifier [9]. However, their treatment of non-separable cases is somewhat heuristic. In fact, there is no systematic way to choose the kernel functions, and hence, reparability of the transformed features is not guaranteed [9].

ANN is commonly consists of neurons that constitute the input layer, one or more hidden layers and an output layer of computational nodes. The learning rule typically used for the multilayer neural network is the back-propagation rule that allows the network to learn to classify.

Three experimentations on speaker independent were done, using Weka Tool with SVM and ANN. Prosodic and voice quality features set were extracted on all speech utterance in Berlin dataset. 10-fold cross validation was used in this experiment. This validation method was used in many other works on EMO-DB.

In WEKA SVMs are implemented as John Platts sequential minimal optimization (SMO) algorithm, ANN implement as Multilayer Perceptron (MLP).

Firstly CVParameterSelection was used to determine the best cost value (C) for three SMO kernels that is: Normalize Poly, Poly and RBF kernel. A comparison between the different kernels was done to determine the best recognition rate among them.

Secondly CVParameterSelection is used to determine the best number of neurons (H) in hidden layer for three different values of learning and momentum rate that is: 1) learning rate 0.3 and momentum rate 0.2 which is the default setting of WEKA. 2) learning rate 0.25 and momentum rate 0.5. 3) learning rate 0.1 and momentum rate 0.9. The pair of learning rate and momentum rate of value {0.25,0.5} and {0.1,0.9} are pairs succesfullt used in speech recognition [25]. The number of epochs was set to 500. Error back propagation was used as a training algorithm. A comparison between the different MLP topology was done to determine the best recognition rate among them.

Figure III show Multilayer Perceptron model from WEKA using the voice quality features.

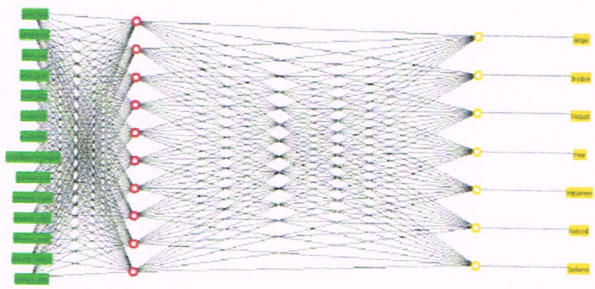

Figure3: Multilayer Perceptron model from WEKA showing the voice quality features

IV. RESULTS AND DISCUSSION

Three comparative experiments were designed. Firstly, different SMO kernels were compared. Table I shows that all parameter sets gave recognition rate of above 74% where the highest and lowest recognition rate different was slightly below 2%. The highest recognition rate was 76.83% using RBF kernel function with cost parameter of value 100.

In the second experiment, different MLP parameters were compared. Table II shows the comparison result. Similar to SMO, all the result gave classification rate of above 74%. Nevertheless, the different between the highest and lowest recognition rate is nearly 4%. The highest recognition rate was 78.69% using the pair of learning rate and momentum rate value of 0.3 and 0.2 respectively with number of hidden node is 100.

TABLE I. RECOGNITION RATE FOR DIFFERENT SMO KERNAL

Kernel	Cost Parameter	Recognition Rate
Normalized Poly	51	74.95%
Poly	2	75.89%
RBF	100	76.82 %

TABLE II. RECOGNITION RATE FOR DIFFERENT MLP TOPLOLGY

MLP Parameters		No. of Neurons	Recognition Rate
learning rate	*momentum rate*		
0.3	0.2	100	78.69%
0.25	0.5	100	77.57%
0.1	0.9	51	74.77

Overall comparison between SMO and MLP indicated that MLP surpassed SMO in recognition performance. However, SMO had much less time in training to build the model. Table III shows the comparison between the two classifiers.

TABLE III. RECOGNITION RATE AND TIME FOR SMO AND MLP

Classifiers	Recognition Rate	Time Taken to Build the Model in Seconds
SMO	76.82%	0.5
MLP	78.69%	60.48

TABLE IV. CONFUSION MATRIX FOR (SMO)

Emo.	Ang.	Bor.	Dis.	Fea.	Hap.	Nat.	Sad.	Rate
Ang.	111	0	2	6	8	0	0	87.40%
Bor.	0	62	2	3	0	13	1	77%
Dis.	6	1	28	4	4	2	1	61%
Fea.	2	3	2	48	8	5	1	70%
Hap.	11	0	3	9	48	0	0	68%
Nat.	0	13	1	3	0	62	0	78%
Sad.	0	3	3	0	1	3	52	84%

In term of the seven emotions classification, the performance varied significantly. Angry was the best emotion recognized by SMO with 87.40% while sadness was the best emotion recognized using MLP with 89.0%. Nevertheless, both emotions are among the highest recognized by both classifiers. On the other hands, the most confusing emotions for SMO were disgust and happy with recognition rate of 61% and 68% respectively. As for MLP, the most confusing emotions were fear and happy with recognition rate of 71% and 69%.

TABLE V. CONFUSION MATRIX FOR (MLP)

Emo.	Ang.	Bor.	Dis.	Fea.	Hap.	Nat.	Sad.	Rate
Ang.	107	0	1	3	16	0	0	84.25%
Bor.	0	66	2	0	0	10	3	81%
Dis.	2	1	33	2	4	1	3	72%
Fea.	2	3	1	49	9	5	0	71%
Hap.	10	0	6	6	49	0	0	69%
Nat.	0	13	1	2	0	62	1	78%
Sad.	0	3	1	1	0	2	55	89%

V. CONCLUSION

This paper compared the performance of two popular classifiers in speech emotion recognition. The Multilayer Neural Network (MLP), and the Support Vector Machine (SVM) using sequential minimal optimization algorithm (SMO).

The results obtained from the experiments show that MLP overcomes SMO in the overall recognition. However, the training for SMO had much lesser time compared to MLP.

Anger and sadness were the best emotions to be recognized for both of the classifiers while disgust fear and happy were the hardest emotions to be recognized.

ACKNOWLEDGMENT

UTM-IRDA Digital Media Centre, Faculty of Computing, Universiti Teknologi Malaysia using Fundamental Research Grant Scheme (FRGS) vot number R.J130000.7828.4F253 . Special thanks to Ministry of Education (MOE) and Research Management Centre (RMC) providing financial support of this research.

REFERENCES

[1.] . Sathe-Pathak and A. R. Panat, "Extraction of pitch and formants and its analysis to identify 3 different emotional states of a person," International Journal of Computer Science,vol. 9.

[2.] J. Sidorova, "Dea report: Speech emotion recognition," 2007.

[3.] A° . Abelin and J. Allwood, "Cross linguistic interpretation of emotional prosody," in ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, 2000.

[4.] M. Lugger and B. Yang, "Psychological motivated multistage emotion classification exploiting voice quality features," Speech Recognition, In-Tech, pp. 395–410, 2008

[5.] J. P. Arias, C. Busso, and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," Computer Speech & Language, vol. 28, no. 1, pp. 278–294, 2014.

[6.] N. A. Hendy and H. Farag, "Emotion recognition using neural network: A comparative study,"

[7.] T. Vogt, E. Andr´e, and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation," in Affect and emotion in human-computer interaction, pp. 75–91, Springer, 2008.

[8.] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and P.-J. Li, "Mandarin emotional speech recognition based on svm and nn," in Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 1, pp. 1096–1100, IEEE, 2006.

[9.] 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, 2011.

[10.] Tickle, S. Raghu, and M. Elshaw, "Emotional recognition from the speech signal for a virtual education agent," in Journal of Physics: Conference Series, vol. 450, p. 012053, IOP Publishing, 2013.