Automatic Speech Emotion Recognition using Auditory Models with Binary Decision Tree and SVM

Enes Yüncü*, Hüseyin Hacıhabiboğlu†, Cem Bozşahin*
*Cognitive Science, Middle East Technical University, Ankara, Turkey
†Department of Modelling and Simulation, Middle East Technical University, Ankara, Turkey

***Abstract*—Affective computing is a term for the design and development of algorithms that enable computers to recognize the emotions of their users and respond in a natural way. Speech, along with facial gestures, is one of the primary modalities with which humans express their emotions. While emotional cues in speech are available to an interlocutor in a dyadic conversation setting, their subjective recognition is far from accurate. This is due to the human auditory system which is primarily non-linear and adaptive. An automatic speech emotion recognition algorithm based on a computational model of the human auditory system is described in this paper. The devised system is tested on three emotional speech datasets. The results of a subjective recognition task is also reported. It is shown that the proposed algorithm provides recognition rates that are comparable to those of human raters.**

I. Introduction

The advent of technology made it necessary and possible to design more natural human-computer interaction mechanisms. A good example to such mechanisms is computers that can respond to their users' emotional states. The field of study with designing such systems is known as affective computing[1]. Affective computing aims to improve human computer interaction in a way that computers could adapt their responses according to human needs. As such, affective computing aims to recognize emotions by capturing cues from visual, auditory and physiological signals recorded from humans.

Speech is the fundamental mode of communication between humans. Such communication involves a speaker and a listener. In a dyadic conversation, participants are both the source and the receiver in turns. On the source side, physiological correlates of emotional state that are manifested in the vocal tract results in speech production. On the receiver side, the speech signal received through the ears is exposed to a series of transformations in the auditory system. Human auditory system is a non-linear and adaptive mechanism which involves frequency-dependent filtering as well as temporal and simultaneous masking. A human listener has access only to cues which are available via the auditory system. Therefore, in a natural interaction scenario, emotion recognition is not perfect.

In this work, speech emotion recognition algorithm based on a model of the human auditory system is developed and its accuracy is evaluated. A state-of-the-art human auditory model is used to process speech signals simulating the process from pinna through the auditory nerve. Features are extracted from the output of the auditory model. They are used to train a binary decision tree for seven different classes (anger, fear, happiness, sadness, disgust, boredom and neutral) of emotions. The binary classifier at each level of the binary decision tree is a support vector machine (SVM). The classifiers are then tested using a validation set to assess the recognition performance.

In the scope of this study, German, Polish and English databases are selected and used to train and test the developed speech emotion recognition system. In order to compare the algorithm with emotion recognition performance of human listeners, a subjective emotion recognition task was carried out with non-German speakers. The correlation between the objective and subjective results was analysed.

The paper is organized as follows: Sec. 2 presents a brief overview of speech emotion recognition and computational models of the human auditory system. Sec. 3 introduces the proposed automatic speech emotion recognition algorithm based on a computational auditory model. In Sec. 4, accuracy of the proposed method and its comparison with a subjective emotion recognition task are reported.

II. Background

A. Automatic Speech Emotion Recognition

Speech emotion recognition has been a popular research topic during the past two decades within the context of affective computing. Automatic recognition of emotion from speech is primarily a pattern recognition task. A feature set based on spectro-temporal features, trajectories of pitch and intensity of the speech has developed in [2]. Selected features were classified into seven emotion labels using support vector machines. An accuracy of 88.6% was obtained on the Berlin emotional speech database [3]. Logarithmic frequency power coefficients were used to generate a feature set in [4]. Using a hidden Markov model, extracted features were classified into six emotion labels. 720 utterances of Burmese and Mandarin from 6 speakers of each language are used as a corpus and 77.1% accuracy level was obtained. Low level contours were extracted in [5]. The corporal were generated in English and German from five speakers. They were then classified into seven emotion labels using hidden Markov models. The accuracy was reported to be 77.8%. A feature set of Mel frequency cepstrum coefficient (MFCC), delta coefficients, zeroth cepstral coefficient and the speech energy is extracted in [6]. Gaussian mixture vector on the autoregressive models used as a classification method. The reported accuracy was 76.0% on Berlin Emotional Database. A feature set comprising 133 features from pitch, MFCC, cepstral coefficients, energy and formants was used in [7]. In speaker dependent level, probabilistic neural network classified the Berlin emotional

database with the accuracy of 94%. In another study, a model is developed to describe the emotional states and their mutual influence in a dyadic conversation [8].
For recording the Danish Emotional Speech Database[9], 2 male and 2 female semi-professional actors expressed five different emotional states. In Berlin Corpus [3], same sentence was uttered by the same person emulating different emotions. Corpus was constituted by 5 male and 5 female actresses. Each actor uttered 5 long and 5 short daily German sentences. Speech Under Simulated and Actual Stress [10] was recorded in English by 32 speakers. The database contains both simulated speech under stress (simulated Domain) and actual speech under stress (*actual domain*). The VAM database [11] was recorded from natural human communication in a German TV talk show Vera am Mittag. Corpus consists of 12 hours of recordings. VAM database was an audio-visual speech corpus. It was recorded from unscripted, authentic discussions between the guests of the talk-show. Database of Polish Emotional Speech [12] consist of speech by 4 male and 4 female speakers. Each speaker uttered 5 different sentences in 6 different emotions in the Polish language. Surrey Audio-Visual Expressed Emotion (SAVEE) [13] database was an audio-visual emotional database which is formed from 4 male actors in 7 different emotions.

B. Computational Models of Human Auditory System

Research on human auditory system led to the development of the computational models of human auditory system. Such models are used frequently in speech recognition and speech emotion recognition tasks. The process of understanding speech is divided into two stages : (1) an auditory preprocessing stage and (2) pattern recognition [14]. In auditory preprocessing stage, speech signal is transformed into internal representations that the human brain has access to.
The computational auditory model employed in this study consists of 6 main stages [15] as given in the Fig. 1. In the first stage, the process in the outer and middle ear is simulated. In the second stage, a 31 channel auditory filterbank is applied to the signal to simulate the basilar membrane. At the end of this stage, 31 frequency dependent signals, each of which corresponds to a position on the BM are obtained. Each signal is subject to an envelope extraction, expansion and adaptation. Hair cells that transduce mechanical energy to neural impulses respond to positive displacements of the basilar membrane. Half wave rectification is the corresponding behavior of such excitation [16]. Presence of 2 stimuli will affect the hearing threshold and additional stimuli will be harder/impossible to hear. Each signal at the output of the second stage is process by the envelope, expansion and adaptation stages. Modulation filterbank is defined as overlapping band pass filters with different center frequencies similar to auditory filterbanks simulating the BM. Although no physiological correlates of modulation filterbank were found in the auditory system [16], it is suggested that there are some neurons in the brainstem which are sensitive to modulation frequencies. In the modulation filterbank, each of 31 signals is exposed to a 12 channel modulation filterbank.

III. Proposed Method

This section presents the method we propose for automatic speech emotion recognition.

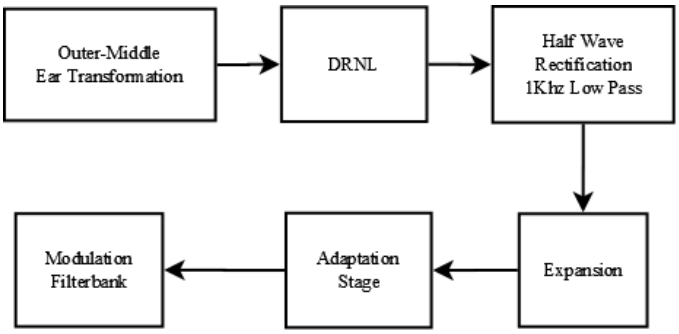

Fig. 1. Computational Model of Human Auditory System Stages

A. Auditory Features for Speech Emotion Recognition

At the output of the computational auditory model, there are a total of 283 modulation filtered signals. At the feature extraction part, simple statistical features such as mean and standard deviation are obtained from the output signals of the computational model. Mean of the signal is given in the following equation.

$$\overline{X}_{i,k} = \frac{1}{L}\sum_{n=1}^{L} x_{n,i,k} \tag{1}$$

x is the output signal of the auditory model where i and k represents the modulation filterbank channel and auditory filterbank channel respectively where $\mathbf{1} \leq \mathbf{i} \leq \mathbf{12}$ and $\mathbf{1} \leq \mathbf{k} \leq \mathbf{31}$. L represents the number of the samples in one speech sample. In the following equation, standard deviation of the signal is given.

$$S_{i,k} = \frac{1}{L-1}[\sum_{n=1}^{L}(\overline{X}_{i,k} - x_{n,i,k})^2]^{\frac{1}{2}} \tag{2}$$

Extracted vectors $\overline{X}_{i,k}$ and $S_{i,k}$ given in (1) and (2) constitute the feature vectors. Since there are total of 283 output signal, dimension of the feature vector became 566 for each speech sample. Fig. 2, 3, 4, 5 show the features represented as a matrix in auditory filterbank channel vs. modulation filterbank channel space. The examples shown in the figures are extracted from sentences in Berlin database. In subfigure (a), the record is taken from speaker 3 while uttering *Das will sie am Mittwoch abgeben* which means *She will hand it in on Wednesday*. In subfigure (b), the same sentence is uttered by speaker 8. In subfigure (c), speaker 3 uttered *Heute abend konnte ich es ihm sagen* which means *Tonight I could tell him*. In subfigure (d), same sentence in subfigure (c) is uttered by speaker 8. In the Figs. 2, 4 visualization of the anger and neutral emotions' mean of the auditory model outputs are visualized. For anger, the extracted feature in the modulation channels of 4 to 12, mean values deviation has a low level. In the Figs. 3, 5 visualization of the anger and neutral emotions' standard deviation of the auditory model outputs are visualized. For anger, the second modulation channel is high and do not change as the auditory filterbank channel changes. On the other hand, neutral emotions' second modulation channel drops as the auditory filterbank channel increases. Distinction between different emotions and similarity between same emotions make

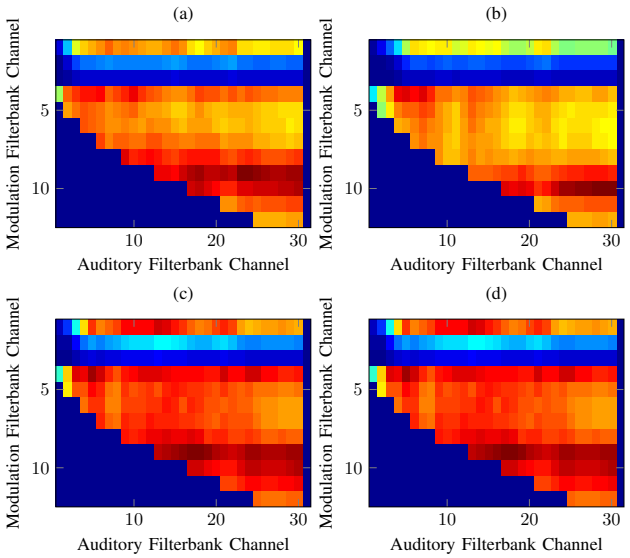

Fig. 2. Mean model output for 'Anger'

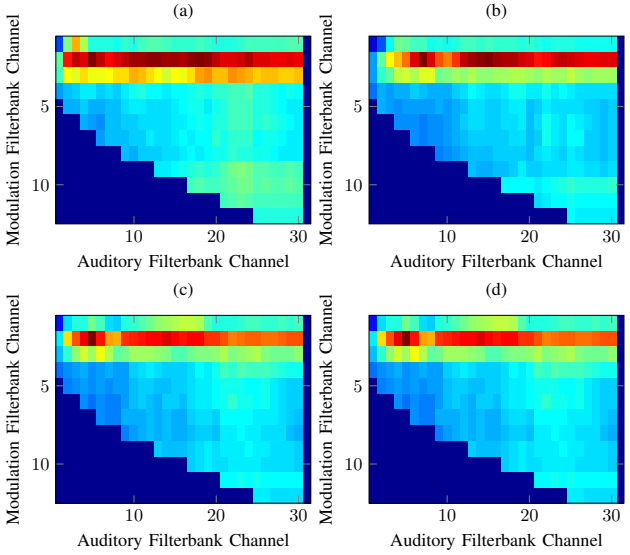

Fig. 3. Standard deviation model output for 'Anger'

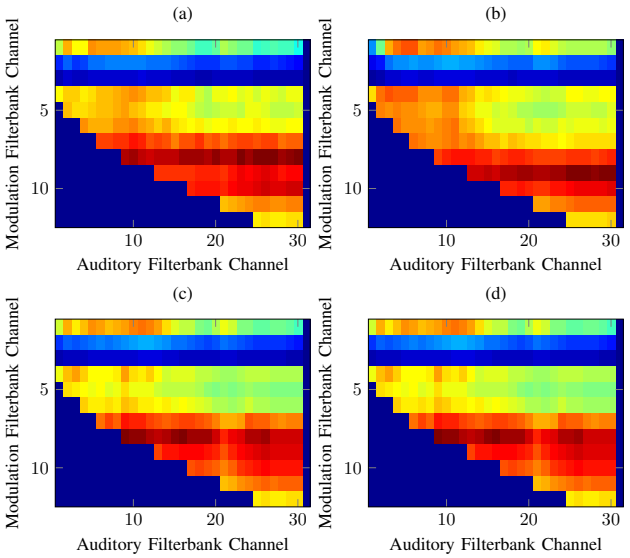

Fig. 4. Mean model output for 'Neutral'

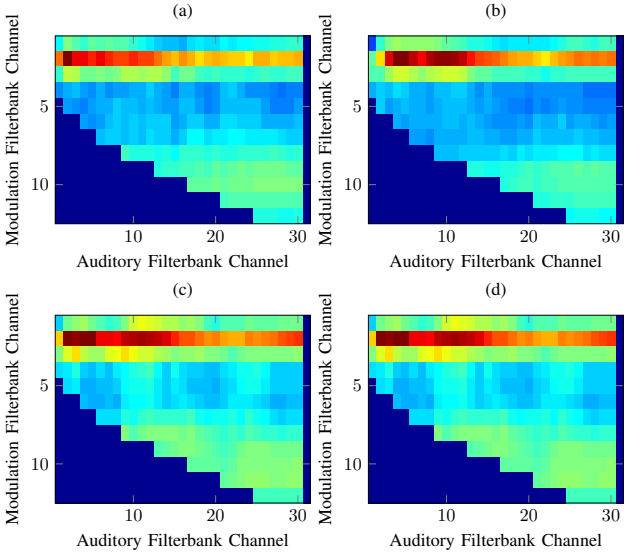

Fig. 5. Standard deviation model output for 'Neutral'

the extracted features appropriate for speech emotion recognition task.

B. Binary Decision Tree

Classification of seven emotion labels using binary classifiers such as SVM, corresponds to a multi-class classification problem. SVM-BDT (Support Vector Machines utilizing Binary Decision Tree), combines advantages of efficient computation of the tree architecture and the high classification accuracy of SVMs [17]. In order to generate a binary decision tree, at each branch of the tree, emotion labels are segmented into two groups until the binary tree reaches a single emotional label. To generate a binary decision tree, distance measurements between the extracted features of the emotion labels are used. In order to measure the distance between emotion labels, the high dimension of the feature set is reduced using principal component analysis. Principal component analysis is a variable reduction method, which transforms the variables that are highly correlated into a smaller number of uncorrelated variables. The projected feature vectors indicated that, the first five principal components carry most of the data. Since other components are very close to 0, they are omitted.
To segment seven emotion labels into two classes, k-means clustering algorithm with Euclidean distance is used. 40 samples from seven emotion labels are selected from the Berlin Emotional Speech Database. The following steps are carried out to design the BDT.

1) Extract the principal component of all the samples belonging to seven emotions labels.
2) Use only the first five principal components.
3) For each emotion label, extract one feature set getting the mean of vectors belonging to the same class. In that way, each emotion label has only one feature

vector with a size of five.

4) Using k-means clustering algorithm, segment seven emotion label into two groups.
5) Extract the principal component of the all samples belonging to the first group of emotion labels.
6) Recursively apply the same procedure until each branch is composed of a single emotion label.

First group of emotions are anger, fear, happiness and disgust. Second group of emotions are sadness, neutral and boredom. In Fig. 6 (a), mean of the projected vectors of excited and non-excited emotions are given. Excited emotion labels are anger, fear, happiness and disgust. Non-excited emotion labels are sadness, neutral and boredom. In (b), happy-anger and disgust-fear emotions' projected feature vectors are given. In (c) disgust and fear emotion labels are presented. Generated binary decision tree is given in Fig. 7. For the given binary decision tree, for each branch one SVM was trained. In that way, six different hyper planes were constructed. These six hyper planes divided the feature space into seven segments.

C. Binary Classification using SVMs

In proposed speech emotion recognition algorithm, a group of features is extracted using a state-of-art computational auditory model. The employed computational auditory model generated total of 283 channels signals. From each channel, 2 features are extracted which are signal's mean and standard deviation. When two features are concatenated, total length of the feature vector becomes 566. The binary tree given in Fig. 7 has 6 branches. For each branch, one SVM is trained. SVMs are trained with the feature vector obtained from auditory model output. Since the linear kernel provided the highest classification accuracy, it was used in all SVMs. In the initialization part, total of six different SVMs were trained. Emotional space is segmented into seven classes with 6 hyperplanes. In the first branch, emotional space is segmented into classes. In each branch, each class is re-segmented into two classes until each class belongs to one emotion label. Therefore, classification accuracy of the binary decision tree is the product of all classification accuracies.

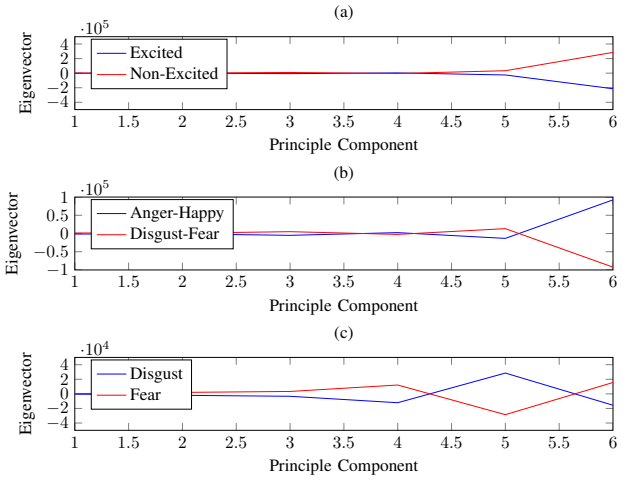

Fig. 6. Feature vectors first five principal components

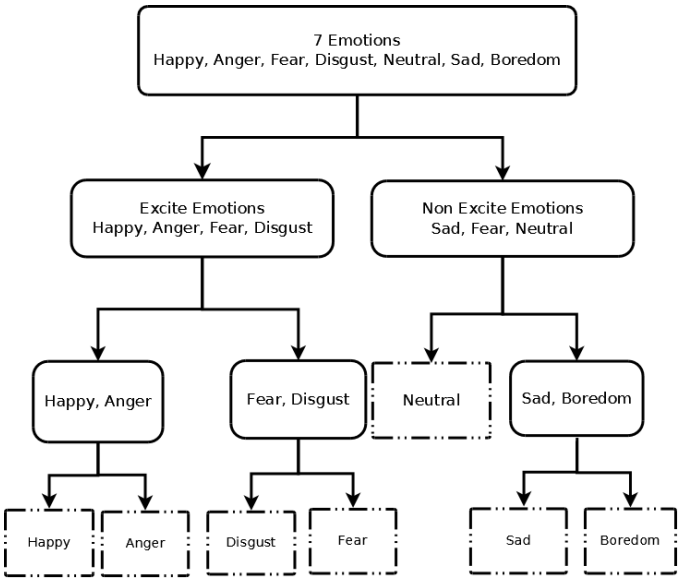

Fig. 7. Generated Binary Tree

IV. Evaluations

A. Subjective Emotion Recognition Performance

In subjective speech emotion recognition test, using only acoustic properties, human listeners' recognition performance is measured. Listening test speech samples are selected from Berlin Emotional Speech Database. In order to test the recognition using only the acoustic properties of speech, non-German speaking subjects were used. This way, it was ensured that the comparison between subjective and automatic recognition was made under equivalent conditions. Extracted recognition rates have provided a comparison for the automatic recognition rates. The test was carried out on 4 male and 6 female listeners. In the listening test, each subject was requested to detect the emotion of 10 speech samples from 7 emotions (A total of 70 sentences per subject and a total number of 700 responses.). Selected emotion labels are happiness, anger, fear, sad, disgust, neutral and boredom. The responses were elicited using a multiple alternative forced choice task. Besides emotion labels, "cannot decide" choice was included. Subjective recognition results are given in Table I(Anger(A), fear(F), happiness(H), disgust(D), neutral(N), sadness(S), boredom(B),can't decide(C), rate(R), precision(P))(leftmost column being the true emotions). Subjective tests have shown that, the overall recognition rate is 58.4%. Anger has the highest rate with 87.0 %. On the other hand, recognition of happiness has the lowest rate of 44.4%. Lowest overall recognition rate of a listener was 37.14%. Highest overall recognition rate of a listener was 78.57%.

B. Evaluation of the Proposed Algorithm

The proposed speech emotion recognition algorithm is applied on three emotional databases, Berlin Database of Emotional Speech [3], SAVEE [13] and Database of Polish Emotional Speech [12].

1) Details of the Evaluation: In the evaluation of the proposed algorithm, four different cases are tested using three

TABLE I. SUBJECTIVE TEST RESULTS

	A	F	H	D	S	N	B	C	R(%)
A	**87**	0	3	3	2	1	2	2	87.0
F	14	**52**	14	4	7	3	1	5	52.0
H	13	10	**40**	2	5	9	8	3	44.4
D	6	10	5	**47**	15	3	13	1	47.0
S	0	5	0	1	**67**	4	22	1	67.0
N	3	2	6	4	5	**58**	12	10	58.0
B	0	1	2	3	13	30	**58**	3	52.7
P(%)	70.3	65.0	57.14	73.44	58.77	53.7	50.0		**58.4**

TABLE II. BERLIN EMOTIONAL SPEECH DATABASE LEAVE ONE SAMPLE OUT CROSS VALIDATION RESULTS

	A	F	H	D	S	N	B	R(%)
A	**116**	2	9	0	0	0	0	91.3
F	1	**53**	8	2	1	3	1	76.8
H	15	8	**45**	1	0	1	1	63.4
D	1	2	2	**37**	0	0	4	80.4
S	0	0	0	0	**57**	3	2	91.9
N	0	1	0	2	2	**68**	6	86.1
B	0	0	0	1	3	9	**68**	83.9
P(%)	87.2	80.3	70.3	86.0	90.5	80.9	82.9	82.9

different databases. First case is the leave one sample out cross validation. This validation used all but one of the samples for training the classifier and the remaining samples for testing it. This validation allows observing the performance of the algorithm with language, speaker and semantic dependence. The second case is leave one speaker out cross-validation. This validation used sentences from all but one of the speakers for training the classifier and the remaining sentences (from a single speaker) for testing it. This validation tests the speaker independent performance of the classifier. The third case tests the performance of the classifier trained with sentences from a single speaker and tested with a sentence from the same speaker. In the fourth case a training set consisting of two of the different emotional speech databases and leave one sample out cross-validation was used.

2) Leave one sample out cross validation: In Berlin Emotional Speech Database, there are a total of 535 emotional speech samples. In order to test the algorithm, leave one sample out cross validation is applied. Each time, 534 speech samples are used for training and 1 speech sample was used for testing. In this task, 7 emotions were classified with an average recognition rate of 82.9%. Sadness has highest recognition rate of 91.9% and happiness has the lowest recognition rate of 63.4%. When automatic speech emotion recognition algorithm is tested on Polish Emotional Speech Database with leave one speech sample out cross validation, average recognition rate was measured as 71.3% as given in Table III. In the Polish emotion speech database, 6 emotions (anger, fear, happiness, boredom, sadness, neutral) are labeled. Anger and fear have the lowest recognition rate of 62.5%. Neutral emotion has the highest recognition rate of 82.5%. Recognition rates are given in Table III.

For the SAVEE database, results are given in TableIV. Overall recognition rate is measured as 73.81%. Six emotions are segmented which are anger, fear, happiness, sadness, neutral and disgust. The highest recognition rate is obtained on emotion label neutral and the lowest emotion recognition rate is obtained on disgust emotion with the given rates 87.5% and 60.0% respectively.

TABLE III. POLISH EMOTIONAL SPEECH DATABASE LEAVE ONE SAMPLE OUT CROSS VALIDATION RESULTS

	A	F	H	S	N	B	R(%)
A	**25**	3	8	2	2	0	62.5
F	7	**25**	1	2	3	2	62.5
H	11	1	**26**	0	2	0	65.0
S	0	5	0	**31**	0	4	77.5
N	1	2	0	4	**33**	0	82.5
B	0	2	2	4	1	**31**	77.5
P(%)	56.8	65.8	70.3	72.1	80.5	83.8	71.3

TABLE IV. SAVEE LEAVE ONE SAMPLE OUT CROSS VALIDATION RESULTS

	A	F	H	D	S	N	R(%)
A	**46**	3	7	4	0	0	76.6
F	6	**38**	11	0	4	1	63.3
H	6	10	**39**	5	0	0	65.0
D	5	2	2	**36**	8	7	60.0
S	0	4	0	5	**46**	5	76.6
N	2	1	0	7	5	**105**	87.5
P(%)	70.7	65.5	66.1	63.1	73.2	88.9	73.81

3) Speaker Independence: In order to measure the speaker independence, leave one speaker out method is used. Berlin Emotional Database consists of speech samples from 10 speakers. In the Polish database, there are eight speakers. In SAVEE database, there are four speakers. In leave one speaker out cross validation, training speech samples are selected from the speech samples which do not belong to the train samples speaker. In Table V, speaker independent results are provided. Recognition rate of Berlin Database is measured as 72.3%. Speaker independent results for Polish database is given in Table VI. Recognition rate is measured as 56.25% which is low compared to recognition rates given in Table III. Highest recognition rate is measured on boredom with a rate of 65.0% and lowest recognition rate is measured in the segmentation of fear with a rate of 47.4%.

4) Speaker Dependence: In Table VII, speaker dependent results are given for Berlin Emotional Speech Database. In

TABLE V. BERLIN EMOTIONAL SPEECH DATABASE SPEAKER INDEPENDENT RESULTS

	A	F	H	D	S	N	B	R(%)
A	**109**	2	15	0	0	0	0	85.8
F	2	**43**	13	0	2	8	1	62.3
H	27	10	**31**	1	0	2	0	43.6
D	1	3	4	**33**	1	1	3	71.7
S	0	1	0	0	**51**	7	3	82.2
N	0	2	0	3	2	**64**	8	81.0
B	0	1	0	8	5	11	**56**	69.14
P(%)	78.4	68.2	49.2	73.3	83.6	68.8	78.8	72.3

TABLE VI. POLISH EMOTIONAL SPEECH DATABASE SPEAKER INDEPENDENT RESULTS

	A	F	H	S	N	B	R(%)
A	**25**	2	9	2	2	0	62.5
F	9	**19**	2	6	1	3	47.5
H	14	3	**20**	0	3	0	50.0
S	1	4	0	**20**	6	9	50.0
N	1	4	1	6	**25**	3	62.5
B	0	4	1	3	6	**26**	65.0
P(%)	50.0	52.7	60.61	54.5	58.14	63.41	56.25

TABLE VII. BERLIN EMOTIONAL SPEECH DATABASE SPEAKER DEPENDENT RESULTS

	A	F	H	D	S	N	B	R(%)
A	**104**	4	15	2	0	2	0	81.8
F	4	**51**	6	6	0	1	1	73.9
H	17	6	**45**	1	0	1	1	63.3
D	3	9	5	**24**	0	1	4	52.17
S	0	0	0	0	**57**	0	5	91.9
N	0	2	0	2	1	**68**	6	86.0
B	0	2	1	1	8	10	**59**	72.8
P(%)	81.3	68.9	62.5	66.7	86.4	81.9	77.6	**76.26**

TABLE VIII. FUSION OF BERLIN AND POLISH EMOTIONAL SPEECH DATABASES LEAVE ONE SAMPLE OUT CROSS VALIDATION RESULTS

	A	F	H	D	S	N	B	R(%)
A	**129**	12	22	0	1	3	0	77.2
F	13	**72**	10	2	3	8	1	66.1
H	24	10	**71**	3	0	2	1	63.9
D	0	1	4	**33**	2	1	5	71.7
S	0	3	0	1	**83**	7	8	81.4
N	1	6	1	3	6	**90**	12	75.6
B	0	4	0	4	8	10	**95**	78.5
P(%)	77.7	66.6	65.7	71.7	80.6	74.4	77.9	73.4

speaker dependent test, training samples are selected from the same speaker of the test speech sample. Speaker dependent recognition rate is 76.26%. Speaker dependent test is applied on all 10 speakers of the Berlin Emotional Speech Database.

5) Language dependence and independence: In another recognition test, Berlin and Polish Emotional Speech Databases are fused and tested with leave one speech sample out cross validation. When two languages are mixed, recognition performance is measured 73.4% for seven emotions. Results are given in Table VIII.

C. Comparison of Subjective and Objective Recognition Performance

In order to measure the agreement between the subjective test results and proposed automatic speech emotion recognition algorithm results, Cohen's kappa was measured. The Kappa coefficient of agreement [18] measures the inter-rater reliability. In order to obtain Cohen's kappa, automatic recognition results for the Berlin Emotional Database using the method leave one sample out cross validation is used. Both subjective and automatic tests are made on the same database. In the subjective test, each subject was requested to detect the emotion of 70 speech samples. Same speech samples are used for the proposed automatic emotion recognition method. The Measured Cohen's kappa varies from 0.24 to 0.61 for ten subjects. Observed agreement score varies 0.36 to 0.67. These results have shown that there is a moderate agreement between the subjective and automatic emotion recognitions.

V. DISCUSSION AND CONCLUSIONS

This paper presents an automatic speech emotion recognition algorithm. The developed algorithm aims to classify seven emotions; anger, happiness, fear, sadness, disgust boredom and neutral. The output of the computational model of auditory system is used as a basis from which features are extracted. The extracted features are mean and standard deviations of the auditory model output signals. The classification is carried out by a binary decision tree consisting of SVM binary classifiers. Both subjective and objective classification results are presented. The results indicated that the proposed algorithm performs well for speaker independent conditions and for different languages. The highest accuracy is observed for Berlin Emotional Speech Database leave one sample out cross validation with a automatic recognition rate of 82.9%. The lowest accuracy is observed for Polish Emotional Speech Database speaker independent results with an accuracy of 56.25%. Automatic recognition results between the Polish and German databases show some differences. In Berlin Emotion Database, anger has the highest recognition rate and happiness has the lowest recognition rate. On the other hand, in the Polish database, anger and happiness has a low recognition rate compared to other emotions. Difference between the results may indicate that, emotions in the German database is exaggerated or emotions in the Polish database not well uttered. Kappa coefficient for the subjective emotion recognition task and the automatic emotion recognition algorithm indicates that there is a moderate agreement between the recognition rates.

REFERENCES

[1] T. T. Jianhua Tao, "Affective Computing: A Review," *Lecture Notes in Computer Science*, vol. 3784, pp. 981–995, 2005.

[2] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic recognition of speech emotion using long-term spectro-temporal features," *Digital Signal Processing*, pp. 1–6, 2009.

[3] F. Burkhardt, "A database of German emotional speech," *Proc Interspeech*, pp. 3–6, 2005.

[4] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, pp. 603–623, 2003.

[5] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model based speech emotion recognition," *ICASSP*, vol. 2, pp. 1–4, 2003.

[6] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," *ICASSP*, vol. 4, pp. 957–968, 2007.

[7] T. Iliou and C. N. Anagnostopoulos, "Classification on speech emotion recognition - a comparative study," *International Journal On Advances in Life Sciences*, vol. 2, pp. 18–28, 2010.

[8] C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," *Interspeech*, pp. 1983–1986, 2009.

[9] I. S. Engberg and A. V. Hansen, "Documentation of the Danish emotional speech database(des)," *Aalborg University, Denmark,*, 1996.

[10] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Getting started with susas: a speech under simulated and actual stress database," *EuroSpeech*, vol. 4, pp. 1743–1746, 1997.

[11] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio visual emotional speech database," *IEEE International Conference on Multimedia and Expo*, 2008.

[12] J. Cichosz and K. Slot, "Low dimensional feature space derivation for emotion recognition," *Interspeech*, 2005.

[13] S. Haq, P. Jackson, and J. Edge, "Audio visual feature selection and reduction for emotion classification," *AVSP*, pp. 185–190, 2008.

[14] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Springer, 2007.

[15] M. Jepsen, S. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *The Journal of the Acoustical Society of America*, vol. 124, p. 422, 2008.

[16] C. J. Plack, "Auditory perception," 2004.

[17] G. Madzarov and D. Gjorgjevikj, "Multi-class classification using support vector machines in decision tree architecture," *Informatica*, vol. 33, pp. 233–241, 2009.

[18] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.