6 ТЕХНИЧЕСКИЕ СИСТЕМЫ И ТЕХНОЛОГИИ ENGINEERING AND TECHNOLOGIES

UDC 004.9

EXTENDED SPEECH EMOTION RECOGNITION AND PREDICTION

T. Anagnostopoulos[a], S.E. Khoruzhnikov[a], V.A. Grudinin[a], C. Skourlas[b]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation, thanag@mail.ifmo.ru
[b] Technological Educational Institute of Athens, Athens, 12243, Greece, cskourlas@teiath.gr

Abstract. Humans are considered to reason and act rationally and that is believed to be their fundamental difference from the rest of the living entities. Furthermore, modern approaches in the science of psychology underline that humans as a thinking creatures are also sentimental and emotional organisms. There are fifteen universal extended emotions plus neutral emotion: hot anger, cold anger, panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt and neutral position. The scope of the current research is to understand the emotional state of a human being by capturing the speech utterances that one uses during a common conversation. It is proved that having enough acoustic evidence available the emotional state of a person can be classified by a set of majority voting classifiers. The proposed set of classifiers is based on three main classifiers: *k*NN, C4.5 and SVM RBF Kernel. This set achieves better performance than each basic classifier taken separately. It is compared with two other sets of classifiers: one-against-all (OAA) multiclass SVM with Hybrid kernels and the set of classifiers which consists of the following two basic classifiers: C5.0 and Neural Network. The proposed variant achieves better performance than the other two sets of classifiers. The paper deals with emotion classification by a set of majority voting classifiers that combines three certain types of basic classifiers with low computational complexity. The basic classifiers stem from different theoretical background in order to avoid bias and redundancy which gives the proposed set of classifiers the ability to generalize in the emotion domain space.

Keywords: speech emotion recognition, affective computing, machine learning.

Acknowledgements. The research was carried out with the financial support of the Ministry of Education and Science of the Russian Federation under grant agreement №14.575.21.0058.

РАСПОЗНАВАНИЕ И ПРОГНОЗИРОВАНИЕ ДЛИТЕЛЬНЫХ ЭМОЦИЙ В РЕЧИ

Т. Анагностопулос[a], С.Э. Хоружников[a], В.А. Грудинин[a], К. Скоурлас[b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, vlad@digiton.ru
[b] Технический образовательный институт Афин, Афины, 12243, Греция, cskourlas@teiath.gr

Аннотация. Люди действуют рационально, и это их фундаментальное отличие от других видов жизни. Кроме того, в современной психологии подчеркивается, что люди как разумные создания отличаются чувствами и эмоциями. Существует пятнадцать видов универсальных длительных эмоций, плюс нейтральное эмоциональное состояние, такие как гнев, злость, паника, страх, тревога, отчаяние, грусть, восторг, радость, интерес, скука, стыд, гордость, отвращение, презрение и нейтральное отношение. В данном исследовании рассматривается понимание эмоционального состояния человека по анализу речи в процессе общения. Доказано, что на основе достаточного объема акустических данных эмоциональное состояние человека может быть классифицировано набором мажоритарных классификаторов. Предложенный набор классификаторов построен на основе трех базовых классификаторов: kNN, C4.5 и SVMRBFKernel. Этот набор обеспечивает лучшую обработку классификаций эмоций, чем каждый из базовых классификаторов в отдельности. Он сравнивается с двумя другими наборами классификаторов: один-против-всех (OAA) мультиклассовый SVM с гибридными ядрами и с набором классификаторов, состоящим из двух базовых классификаторов C5.0, и нейронная сеть (NeuralNetwork). Предложенный вариант достигает лучшего результата, чем два других набора классификаторов. В настоящей статье осуществляется классификация эмоций набором мажоритарных классификаторов, который состоит из трех определённых базовых классификаторов, имеющих низкую вычислительную сложность. Базовые классификаторы базируются на различных теоретических данных с целью избегания отклонений и избыточности, что дает предложенному набору классификаторов возможность обобщиться в пространство определений эмоций.

Ключевые слова: распознавание эмоций в речи, расчет эмоций, машинное обучение

Благодарности. Исследования проводились при финансовой поддержке Министерства образования и науки Российской Федерации в рамках Соглашения о предоставлении субсидии №14.575.21.0058.

Introduction

Humans are considered to reason and act rationally and that is believed to be their fundamental factor that differentiates them from the rest of living entities. Although, modern approaches in the science of psychology underline that human except of thinking creatures are also sentimental and emotional organisms. The field of psychology that studies this aspect of human nature is Emotion Intelligence [1]. Emotion is a subjective, conscious experience characterized primarily by psycho physiological expressions, biological reactions and mental states. It is often associated and considered reciprocally influential with mood, temperament, personality, disposition and motivation [2]. Emotion is often the driving force behind motivation, positive or negative [3]. There are fifteen universal extended emotions plus neutral emotion, that is: hot anger, cold anger, panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt and neutral [4].

The experience of emotion is referenced as affect and it is a key part of the process of an organism's interaction with stimuli [5]. Affect also refers to affect display [6], which is the facial, speech or gestural behavior that serves as an indicator of affect. Affective computing is the study and development of systems and devices that can recognize, interpret, process and simulate human affects. It is an interdisciplinary field spanning from informatics, psychology and cognitive science [7]. One field of informatics that could be used in order to classify affects and exploit their fundamental emotional state is machine learning. Thus we expand the wide area of the affective computing with this of machine learning algorithms and classification models [8].

In the current paper the problem of speech emotion recognition will be treated as an ensemble classification and prediction issue. First, a number of base classifiers are going to be used for speech emotion classification. Then an ensemble majority voting classifier will expand the dynamics of the base classifiers in order to create a concrete classification model. The proposed model is evaluated with other state-of-the-art models and it is proven that it achieves higher classification scores over the other models.

The paper is organized as follows. In Section "Related Work", the related work of the state-of-the-art speech classification models is presented. In Section "Data Model", the data model which is used in the current study is described. In Section "Ensemble Classification", it is described how an ensemble classifier is built from a set of base classifiers. In Section "Emotion Prediction", it is presented how speech emotion can be predicted. In Section "Performance Evaluation", the evaluation of the proposed model with the other state-of-the-art models is performed. In Section "Discussion and Conclusion", a discussion is done in order to explain the effect of the proposed classification model. The paper concludes with Section "References", where future work and trends are outlined.

Related Work

A vast amount of work has been done in the area of speech emotion recognition. Among all we can distinguish [9] where an automatic feature selector which combined the random forest RF2TREE ensemble algorithm and the simple decision tree C4.5 algorithm is developed. In [10] a Hidden Markov Model (HMM) is proposed for joint speech and emotion recognition in order to include multiple versions of each emotion. Then emotion classification was performed using ensemble majority voting between emotion labels. The authors in [11] demonstrate commonly used k Nearest Neighbors (kNN) classifier for segment-based speech emotion recognition and classification. In [12] frame-wise emotion classification is used based on vector quantization techniques. Within this scheme in order to classify an input utterance an emotion was classified using an ensemble majority voting scheme between frame-level emotion labels. The authors in [13] used Fuzzy Logic classification in order to combine categorical and primitives-based speech emotion recognition.

The authors in [14] implemented a real-time system for discriminating between neutral and angry speech which used Gaussian Mixture Models (GMMs) for Mel-Frequency CepstralCoefficients (MFCC) features in combination with a prosody-based classifier. In [15] it is demonstrated that emotion can be better differentiated by specific phonemes than others using phoneme-specific GMM. The authors in [16] investigate combination of features at different levels of granularity by integrating GMM log-likelihood score with commonly-used suprasegmental prosody-based emotion classifiers. In [17] GMMs are applied to emotion recognition using a combined feature set which was obtained by concatenating MFCC and prosodic features.

The authors in [18] demonstrated Support Vector Machine (SVM) classification with manifold learning methods using covariance matrices of prosodic and spectral measures evaluated over the entire utterance. In [19] an SVM Multiple Kernel Learning (MKL) is proposed, where the decision rule is a weighted linear combination of multiple single kernel function outputs. The authors in [20] introduce SVM classification with Radial Basis Function (RBF) kernels and MFCC statistics over phoneme type classes in the utterance. In [21] the authors use an ensemble mixture model of base SVM classifiers where the outputs of the classifiers are normalized and combined using a thresholding fusion function in order to classify the speech emotion. The authors in [22] use an ensemble mixture model of which combines C5.0 and Neural Network (NN) base classifiers in order to achieve speech emotion classification.

Finally, in [23] the authors use an ensemble majority voting classifier which combines kNN, C4.5 and SVM Polynomial Kernel. The results were apparently better than the previous approaches. However, the emotions classified were limited to the six basic emotions with regards to the Ekman's emotion taxonomy [24]. In this paper we propose a model which is designed to extend the classification to sixteen emotions. The proposed model is compared with the model in [21] and the model in [22], given the same speech emotion database [25]. The results show that the proposed model achieves better performance than those models.

Data Model

HUMAINE [25] database is used in order to perform emotion classification from speech utterances. The speech utterances ranged from positive to negative emotions. We used fifteen universal extended emotions plus neutral emotion, that is: hot anger, cold anger, panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt and neutral. A set of acoustic parameters which are related to the aforementioned emotional [4] are employed.

The acoustic parameters are:
– F0:
1. Perturbation,
2. Mean,
3. Range,
4. Variability,
5. Contour,
6. Shift Regularity.

– Formants:
7. F1 Mean,
8. F2 Mean,
9. F1 Bandwidth,
10. Formant Precision.

– Intensity:
11. Mean,
12. Range,
13. Variability.

– Spectral Parameters:
14. Frequency range,
15. High-frequency energy,
16. Spectral noise.

– Duration:
17. Speech rate,
18. Transition time.

We perform z-transformation [26] to these eighteen acoustic parameters and we feed them to the base classifiers, as it is discussed in the next section.

Ensemble Classification

The proposed model is based on ensemble classification majority voting scheme over certain types of base classifiers which are of low computational complexity [27]. Three base classifiers are used from different theoretical background in order to avoid bias and redundancy [8].

The three base classifiers are:
1. *k*NN, which is a nonparametric classifier,
2. C4.5, which is a nonmetric classifier, and
3. SVM with RBF Kernel, which is a linear discriminant function classifier.

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0.1	0	0	0	0	0	0	0	0
	E5	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**0.8**	0	0	0.2	0	0	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0	0	0	0	0	0	**0.9**	0	0	0	0	0.1	0	0	0
	E9	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0	0	**0.8**	0.2	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0.1	0	0	0	**0.9**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 1. Confusion matrix of *k*NN for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.95$

Classification problem

A number of observation pairs (x_i, y_i) $i = 1, \dots, n$ where $x \in X \subset \mathbb{R}^p$ and $y \in Y = \{$hot anger, cold anger, panic, fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt, neutral$\}$ is observed. $X$is known as the predictor space (or attributes) and Y is the response space (or class). In this case the

number of attributes is eighteen like the number of the acoustic parameters. The objective is to use these observations in order to estimate the relationship between $X$and Y, thus predict Y from X. Usually the relationship is denoted as a classification rule,

$$h_j(X) = \arg\max P(y|X, \theta_j) \quad \text{(Eq. 1)}$$

where, $j = 1, \ldots, 3$, and $P(.,.)$ is the probability distribution of the observed pairs, θ is the parameter vector for each base classifier, and j is the number of the base classifiers. In this case, we have three classification rules, one for each base classifier.

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0
	E5	0	0	0	0	**0.9**	0	0	0	0	0	0	0	0	0	0.1	0
	E6	0	0	0	0	0	**0.8**	0	0	0	0	0.2	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0	0	**0.9**	0.1	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0.1	0	0	0	**0.9**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	**0.9**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 2. Confusion matrix of C4.5 for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.96$

Ensemble majority voting classification

Each of the three base classifiers is an expert in a different region of the predictor space because they treat the attribute space under different theoretical basis [28]. The three classifiers could be combined in such a way in order to produce an ensemble majority voting classifier that is superior to any of the individual rules. A popular way to combine these three base classification rules is to let an ensemble classifier,

$$C(X) = mode\,\{h_1(X), h_2(X), h_3(X)\} \quad \text{(Eq. 2)}$$

to classify X to the class that receives the largest number of classifications (or votes) [29]. In the next section the three base classifiers and the ensemble classifier are built. It is shown that the ensemble majority voting classifier achieves better accuracy as it is analyzed in the relative confusion matrices.

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0.1	0
	E5	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**0.9**	0	0	0	0.1	0	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0.1	0	0	0	0	0	**0.9**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**0.9**	0	0.1	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0.1	0	**0.9**	0	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0.1	0	0	**0.9**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	**0.9**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 3. Confusion matrix of SVM RBF Kernel for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.95$

Prediction accuracy

In order to measure the accuracy of the classification, we define the metric of classification accuracy. In the case of a separate emotional class, we define the emotion class accuracy,

$$e_k = \frac{tp_k + tn_k}{tp_k + tn_k + fp_k + fn_k} \quad \text{(Eq. 3)}$$

here, $k = 1, \dots, 16$, denotes the number of the emotional classes, and tp_k, tn_k, fp_k, fn_k denote the emotion class true positive, true negative, false positive and false negative classified utterances, respectively. In the case of all emotional classes in average, we define the prediction accuracy,

$$p = \frac{\sum_{k=1}^{16} e_k}{16} \quad \text{(Eq. 4)}$$

which denotes the overall accuracy of a classifier given a specific observed number of observation pairs (x_i, y_i) $i = 1, \dots, n$ for the fifteen universal extended emotions plus neutral emotion.

Emotion Prediction

10-*fold-cross-validation* technique [30] is used, provided by WEKA data mining open source workbench [31], in order to measure the emotion class accuracy e_k and the prediction accuracy p for the proposed classification scheme. HUMAINE [25] database is used in order to perform emotion classification from the same speech utterances. Specifically, English language speech information of 48 persons (26 males and 22 females) is exploited. Every person has expressed speech utterances of the fifteen universal extended emotions plus neutral emotion, thus the total number of the observed pairs is $n = 768$. Because of space limitations in visualizing the results in tabular format a label is assigned to each emotion, that is E1: hot anger, E2: cold anger, E3: panic, E4: fear, E5: anxiety, E6: despair, E7: sadness, E8: elation, E9: happiness, E10: interest, E11: boredom, E12: shame, E13: pride, E14: disgust, E15: contempt and E16: neutral.

The confusion matrix [32] is presented in Table 1 for the emotion class accuracy e_k and the prediction accuracy p of the *k*NN nonparametric classifier. Table 2 presents the confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the C4.5 nonmetric classifier. Table 3 presents the confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the SVM RBF Kernel classifier. The confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the proposed ensemble majority voting classifier are presented in Table 4.

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0.1	0	0	0	0	0	0	0	0
	E5	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0.1	0	0	0	0	0	**0.9**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0	0	**0.9**	0.1	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0.1	0	0	0	**0.9**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 4. Confusion matrix of Ensemble Majority Voting Classifier for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.97$

Confusion matrices were obtained from WEKA. Specifically, 10 fold cross validation is used where the sample is divided into 10 equal length parts. There is no resampling in the classification process. For 10 consecutive repetitions the classifier is trained with 9 parts and tested with the remaining 1 part. During the repetitions each of the 10 parts is considered only one time for testing. For each repetition classification results are computed which are summarized to true positives, true negatives, false positives and false negatives. After the 10 repetitions the classification results are averaged and presented in the confusion matrices. Confusion matrices have more information than the presented classification schema in (Eq. 2) because expect of the true positives and true negatives, described in (Eq. 2), they also incorporate the false positives and false negatives as well. In WEKA, confidence interval for each acoustic parameter value is set to 95 percent.

Table 5 depicts the overall emotion class accuracy e_k for the three base classifiers and the proposed ensemble majority voting classifier. Table 6 depicts the overall prediction accuracy p for the three base classifiers and the proposed ensemble majority voting classifier. As it is proved, the emotion class accuracy e_k and the prediction accuracy p of the ensemble majority voting classifier is greater than these of the three base classifiers. In the discussion Section "Discussion and Conclusion" it is explained why these experimental results are observed.

Performance Evaluation

The proposed model is compared with other two classification models [21] and [22], in literature and it is proved to achieve better results by means of emotion class accuracy e_k and prediction accuracy p, given the same speech emotion HUMAINE database [25]. The same experimental setup is used as in Section "Emotion Prediction". 10-*fold-cross-validation* technique is used in order to measure the emotion class accuracy e_k and the prediction accuracy p for the compared classification schemes. The model in [21] uses a one-against-all (OAA) multiclass SVM classification scheme with Hybrid kernel functions, which constitutes an ensemble classifier. The core of OAA for multiclass SVM classifiers, as it is introduced in [33], is that the observed pair (x_i, y_i) $i = 1, \dots, n$ can be classified only if one of the SVM classes accepts the observed pair while all other SVMs reject it at the same time, thus making a unanimous decision. The model in [22] uses an ensemble classifier which constitutes of a combination of C5.0 and NN base classifiers. The core of the combined ensemble classifier is that it classifies an observed pair (x_i, y_i) $i = 1, \dots, n$ to the class with the higher probability density function (PDF) among the two base classifiers.

		Emotion Class Accuracy e_k															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Classifier	*k*NN	1.0	0.9	1.0	0.9	1.0	0.8	1.0	0.9	1.0	0.8	1.0	0.9	1.0	1.0	1.0	1.0
	C4.5	1.0	1.0	1.0	1.0	0.9	0.8	1.0	1.0	1.0	0.9	1.0	1.0	0.9	1.0	0.9	1.0
	SVM RBF	1.0	0.9	1.0	0.9	1.0	0.9	1.0	0.9	0.9	0.9	1.0	1.0	0.9	1.0	0.9	1.0
	Majority	**1.0**	**1.0**	**1.0**	**0.9**	**1.0**	**1.0**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	**1.0**	**1.0**	**1.0**

Table 5. Emotion Class Accuracy e_k for the three base classifiers and the proposed Ensemble Majority Voting Classifier

		Prediction Accuracy p
Classifier	*k*NN	0.95
	C4.5	0.96
	SVM RBF	0.95
	Majority	**0.97**

Table 6. Prediction Accuracy p for the three base classifiers and the proposed Ensemble Majority Voting Classifier

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0.1	0
	E5	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**0.9**	0	0	0	0.1	0	0	0	0	0	0
	E7	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0
	E8	0	0.1	0	0	0	0	0	**0.9**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**0.9**	0	0.1	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0.1	0	**0.9**	0	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**0.9**	0	0	0	0.1	0
	E12	0	0	0	0	0	0	0	0.1	0	0	0	**0.9**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0.1	0	0	**0.9**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0
	E15	0	0	0	0	0	0	0	0.2	0	0	0	0	0	0	**0.8**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 7. Confusion matrix of OAA multiclass SVM classifier with Hybrid kernel functions for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.93$

		Predicted Emotion															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Actual Emotion	E1	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E2	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	E3	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0	0	0	0
	E4	0	0	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0.1	0
	E5	0	0	0.1	0	**0.9**	0	0	0	0	0	0	0	0	0	0	0
	E6	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0	0	0
	E7	0.1	0	0	0	0.1	0	**0.8**	0	0	0	0	0	0	0	0	0
	E8	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0	0
	E9	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0	0	0
	E10	0	0	0	0	0	0	0	0	0	**0.9**	0.1	0	0	0	0	0
	E11	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0	0	0
	E12	0	0	0	0	0	0	0	0.2	0	0	0	**0.8**	0	0	0	0
	E13	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**	0	0	0
	E14	0	0	0	0	0	0	0	0	0	0	0	0	0.1	**0.9**	0	0
	E15	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	**0.9**	0
	E16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1.0**

Table 8. Confusion matrix of Combined C5.0 and NN classifier for Emotion Class Accuracy e_k. Prediction Accuracy $p = 0.94$

Table 7 presents the confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the OAA multiclass SVM with hybrid kernel functions. Table 8 presents the confusion matrix for the emotion class accuracy e_k and the prediction accuracy p of the combined C5.0 and NN classifier. Table 9 depicts the overall emotion class accuracy e_k for these two models and our ensemble majority voting classifier. Table 10 depicts the overall prediction accuracy p for these two models and our ensemble majority voting classifier.

As it is proved the emotion class accuracy e_k and the prediction accuracy p of the ensemble majority voting classifier is greater than these of the other two compared classifiers. In the discussion Section "Discussion and Conclusion" it is explained why it is observed these experimental results.

		Emotion Class Accuracy e_k															
		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16
Classifier	SVM Hybrid	1.0	0.9	1.0	0.9	1.0	0.9	1.0	0.9	0.9	0.9	0.9	0.9	0.9	1.0	0.8	1.0
	C5.0 – NN	1.0	1.0	1.0	0.9	0.9	1.0	0.8	1.0	1.0	0.9	1.0	0.8	1.0	0.9	0.9	1.0
	Majority	**1.0**	**1.0**	**1.0**	**0.9**	**1.0**	**1.0**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**	**1.0**	**1.0**	**1.0**

Table 9. Emotion Class Accuracy e_k for the two compared classifiers and the proposed Ensemble Majority Voting Classifier

Discussion and Conclusion

A discussion is performed in order to explain why these experimental results are observed in the two previous Sections "Emotion Prediction" and "Performance Evaluation". In Section "Emotion Prediction" it is proved that the ensemble majority voting classifier achieves better scores than the three base classifiers. This is explained because each base classifier is biased in a specific domain of the emotion classification problem, thus the advantages of one classifier might be disadvantages for the other two classifiers and vice versa. The overall superiority of the ensemble classifier is its ability to combine the redundant information of the base classifiers in order to create a more sound classification scheme.

		Prediction Accuracy p
Classifier	SVM Hybrid	0.93
	C5.0 – NN	0.94
	Majority	**0.97**

Table 10. Prediction Accuracy p for the two compared classifiers and the proposed Ensemble Majority Voting Classifier

In Section "Performance Evaluation" it is also proved that the ensemble majority voting classifier achieves better scores than the two compared ensemble classifiers. In the case of [21] model this is explained because the base classifiers are of the same general SVM linear discriminant functions bias. In the case of [22] model this is explained because the base classifiers where too few (i.e., only two) in order their union not to be able to generalize to the whole set of pairs $(x_i, y_i)\ i = 1, \ldots, n$. Both [21] and [22] models do not take into

consideration the majority votes, of the whole set of the classifiers (see Eq. 2, Section "Ensemble Classification"), which is used by the proposed model.

It is proved that the proposed ensemble majority voting classifier achieves better performance in classifying the fifteen universal extended emotions plus neutral emotion than the three base classifiers and the other two compared ensemble classifiers. Future work is intended by exploiting other context (i.e., facial expressions) in order to design multimodal models.

References

1. Matthews G., Zeidner M., Roberts R.D. *Emotional Intelligence: Science and Myth*. Cambridge, MIT Press, 2003, 697 p.
2. Schacter D.L. *Psychology*. 2nd ed. NY, Worth Publishers, 2011, 624 p.
3. Gaulin S.J.C., McBurney D.H. *Psychology: An Evolutionary Approach*. Upper Saddle River, Prentice Hall, 2003.
4. Scherer K.R. Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 2003, vol. 40, no. 1–2, pp. 227–256. doi: 10.1016/S0167-6393(02)00084-5
5. Thompson E.R. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology*, 2007, vol. 38, no. 2, pp. 227–242. doi: 10.1177/0022022106297301
6. Parkinson B., Simons G. Worry spreads: interpersonal transfer of problem-related anxiety. *Cognition and Emotion*, 2012, vol. 26, no. 3, pp. 462–479. doi: 10.1080/02699931.2011.651101
7. Picard R.W. *Affective Computing*. Cambridge, MIT Press, 2000, 304 p.
8. Duda R.O., Hart P.E., Stork D.G. *Pattern Classification*. NY, John Wiley and Sons, 2000, 735 p.
9. Rong J., Chen Y.-P.P. Chowdhury M., Li G. Acoustic features extraction for emotion recognition. *Proc. 6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007*, 2007, art. 4276418, pp. 419–424. doi: 10.1109/ICIS.2007.48
10. Meng H., Pittermann J., Pittermann A., Minker W. Combined speech-emotion recognition for spoken human-computer interfaces. *Proc. IEEE International Conference on Signal Processing and Communications*, 2007, art. 4728535, pp. 1179–1182. doi: 10.1109/ICSPC.2007.4728535
11. Shami M.T., Kamel M.S. Segment-based approach to the recognition of emotions in speech. *Proc. IEEE International Conference on Multimedia and Expo, ICME 2005,* 2005, vol. 2005, art. 1521436, pp. 366–369. doi: 10.1109/ICME.2005.1521436
12. Sato N., Obuchi Y. Emotion recognition using mel-frequency cepstral coefficients. *Journal of Natural Language Processing*, 2007, vol. 14, no. 4, pp. 83–96. doi: 10.5715/jnlp.14.4_83
13. Grimm M., Mower E. Kroschel K., Narayanan S. Combining categorical and primitives-based emotion recognition. *Proc. 14th European Signal Processing Conference*. Florence, Italy, 2006, pp. 345–357.
14. Kim S., Georgiou P.G., Lee S., Narayanan S. Real-time emotion detection system using speech: multi-modal fusion of different timescale features. *Proc. 9th IEEE International Workshop on Multimedia Signal Processing, MMSP 2007*. Chania, Crete, 2007, art. 4412815, pp. 48–51. doi: 10.1109/MMSP.2007.4412815
15. Sethu V., Ambikairaja E., Epps J. Phonetic and speaker variations in automatic emotion classification. *Proc. Annual Conference of the International Speech Communication Association, Interspeech*. Brisbane, Australia, 2008, pp. 617–620.
16. Vlasenko B., Schuller B., Wendemuth A., Rigoll G. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. *Affective Computing and Intelligent Interaction*, 2007, vol. 4738 LNCS, pp. 139–147.
17. Vondra M., Vich R. Recognition of emotions in german speech using gaussian mixture models. *Multimodal Signals: Cognitive and Algorithmic Issues*, 2009, vol. 5398 LNAI, pp. 256–263. doi: 10.1007/978-3-642-00525-1_26
18. Ye C., Liu J., Chen C., Song M., Bu J. Speech emotion classification on a riemannian manifold. *Advances in Multimedia Information Processing – PCM 2008*, 2008, vol. 5353 LNCS, pp. 61–69. doi: 10.1007/978-3-540-89796-5_7
19. Gonen M., Alpaydin E. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 2011, vol. 12, pp. 2211–2268.
20. Bitouk D., Verma R., Nenkova A. Class-level spectral features for emotion recognition. *Speech Communication*, 2010, vol. 52, no. 7–8, pp. 613–625. doi: 10.1016/j.specom.2010.02.010
21. Yang N., Muraleedharan R., Kohl J., Demirkol I., Heinzelman W., Sturge-Apple M. Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion. *Proc. 4th IEEE Workshop on Spoken Language Technology, SLT 2012*. Miami, Florida, 2012, art. 6424267, pp. 455–460. doi: 10.1109/SLT.2012.6424267

22. Javidi M.M., Roshan E.F. Speech emotion recognition by using combinations of C5.0, neural network (NN), and support vectors machines (SVM) classification methods. *Journal of Mathematics and Computer Science*, 2013, vol. 6, no. 3, pp. 191–200.
23. Anagnostopoulos T., Skourlas C. Ensemble majority voting classifier for speech emotion recognition and prediction. *Journal of Systems and Information Technology*, 2014, vol. 16, no. 3, pp. 222–232. doi: 10.1108/JSIT-01-2014-0009
24. Ekman P. An argument for basic emotions. *Cognition and Emotion*, 1992, pp. 169–200.
25. Douglas-Cowie E., Cowie R., Sneddon I., Cox C., Lowry O., McRorie M., Martin J.-C., Devillers L., Abrilian S., Batliner A., Amir N., Karpouzis K. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. *Proc. 2nd International Conference on Affective Computing and Intelligent Interaction, ASCII 2007*. Lisbon, Portugal, 2007, vol. 4738 LNCS, pp. 488–500.
26. Jury E.I. *Theory and Application of the Z-Transform Method*. Malabar, Krieger Pub Co, 1973, 330 p.
27. Friedman J., Hastie T., Tibshirani R. *The Elements of Statistical Learning*. NY, Springer, 2001, 524 p.
28. Alpaydin E. *Introduction to Machine Learning*. 2nd ed. Cambridge, MIT Press, 2010, 581 p.
29. Basu S., Dasgupta A. The mean, median, and mode of unimodal distributions: a characterization. *Theory of Probability and its Applications*, 1997, vol. 41, no. 2, pp. 210–223. doi: 10.1137/S0040585X97975447
30. Seymour G. *Predictive Inference*. NY, Chapman and Hall, 1993, 240 p.
31. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. The WEKA data mining software: an update. *SIGKDD Explorations*, 2009, vol. 11, no. 1, pp. 10–18. doi: 10.1145/1656274.1656278
32. Stehman S.V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 1997, vol. 62, no. 1, pp. 77–89. doi: 10.1016/S0034-4257(97)00083-7
33. Vapnik V.N. *The Nature of Statistical Learning Theory*. 2nd ed. NY, Springer, 2000, 314 p.

Theodoros Anagnostopoulos – Lead Research Associate, Department of Infocommunication Technologies, ITMO University, Saint Petersburg, 197101, Russian Federation, thanag@mail.ifmo.ru

Sergei E. Khoruzhnicov – Dean of the Faculty, Department of Infocommunication Technologies, ITMO University, Saint Petersburg, 197101, Russian Federation, xse@mail.ifmo.ru

Vladimir A. Grudinin – Head of Department, Department of Infocommunication Technologies, ITMO University, Saint Petersburg, 197101, Russian Federation, grudinin@mail.ifmo.ru

Christos Skourlas – PhD, Professor, Professor, Department of Informatics, Technological Educational Institute of Athens, Athens, 12243, Greece, cskourlas@teiath.gr

Анагностопулос Теодорос – PhD, ведущий научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, thanag@mail.ifmo.ru

Хоружников Сергей Эдуардович – кандидат физ.-мат. наук, доцент, декан, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, xse@mail.ifmo.ru

Грудинин Владимир Алексеевич – кандидат технических наук, доцент, заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, vlad@digiton.ru

Скоурлас Кристос – PhD, профессор, профессор в департаменте информатики, Технический образовательный институт Афин, Афины, 12243, Греция, cskourlas@teiath.gr

Принято к печати 01.09.14
Accepted 01.09.14