Speech-based emotion recognition: Application of collective decision making concepts

Christina Brester[1,a], Eugene Semenkin[2,b] and Maxim Sidorov[3,c]
[1,2] Siberian State Aerospace University named after academician M. F. Reshetnev,
31 "Krasnoyarskiy Rabochiy" pr., Krasnoyarsk, 660014, Russian Federation
[3] Ulm University, Albert-Einstein-Allee, 43, Ulm, 89081, Germany
[a]christina.bre@mail.ru, [b]eugenesemenkin@yandex.ru, [c]maxim.sidorov@uni-ulm.de

***Abstract*—Nowadays "human-machine" interactions are not as vivid as they might be. Spoken dialogue systems use only uniform phrases to respond and do not take into consideration the personal qualities of the speaker (age, gender, educational level, and emotions) and adjust their answers accordingly. Therefore it is of utmost importance to elaborate effective procedures that succeed in speech signal processing and expose human features. In this research we propose some techniques based on collective decision making to improve the performance of existing methods applied for human emotion recognition. To prove the effectiveness of the proposed approaches, we engage a number of multilingual corpora (German, English and Japanese). According to the results obtained, a high level of emotion recognition was achieved (up to a 9.93% relative improvement compared with conventional models).**

Keywords—speech-based emotion recognition; classifier; performance; collective decision making.

I. INTRODUCTION

One of the obvious ways to improve the intellectual abilities of spoken dialogue systems is related to their personalization. While communicating, machines should perceive the qualities of the user (as people usually do) such as age, gender and emotions to adapt its answers for the particular speaker.

In this paper we consider one particular aspect of the personalization process that is *speech-based emotion recognition*. Generally, any approach used to solve this recognition problem consists of three main stages.

At first, it is necessary to extract acoustic characteristics from the collected utterances. At the «INTERSPEECH 2009 Emotion Challenge» an appropriate set of acoustic characteristics representing any speech signal was introduced. This set of features comprised attributes such as power, mean, root mean square, jitter, shimmer, 12 MFCCs and 5 formants. The mean, minimum, maximum, range and deviation of the following features have also been used: pitch, intensity and harmonicity. The number of characteristics is 384. To get the conventional feature set introduced at INTERSPEECH 2009, the Praat or OpenSMILE systems might be used [1] [2]. Secondly, all extracted attributes or the most relevant of them [3] should be involved in the supervised learning process to adjust a classifier. At the final stage, the signal that has to be analysed is transformed into an unlabelled feature vector (also with the usage of the Praat or OpenSMILE systems) and then the trained classification model receives it as the input data to make a prediction.

However, there is one crucial question related to the classification model providing high performance. Actually, it is almost impossible for the online dialogue systems to vary classifiers and determine the most effective one while interacting with a user.

Indeed, in other research it has been revealed that the recognition accuracy depends significantly on the classifier applied. In the study [4] the authors have compared the emotion recognition performance of various classifiers: a support vector machine, linear discriminant analysis, quadratic discriminant analysis and a hidden Markov model on the SUSAS [5] and AIBO [6] databases of emotional speech. The following set of signal features has been used in the study: pitch, log energy, formant, mel-band energies, and mel frequency cepstral coefficients (MFCCs). The authors have managed to achieve the highest value of accuracy (70.1% and 42.3% on the databases, respectively) using a Gaussian support vector machine.

Consequently, some general approaches based on involving different models should be elaborated. In this research we propose three schemes of taking into account predictions of different classifiers and producing the collective decision. The effectiveness of these algorithmic schemes is investigated on a set of multilanguage databases (English, German, and Japanese).

The rest of the paper is organized as follows: in Section II the corpora description is presented, Section III describes some conventional models and the approaches proposed. The experiment conducted, the results obtained, and the main inferences are introduced in Section IV. The conclusion and future work are described in Section V.

II. DATABASES DESCRIPTION

In the study a number of speech databases have been used and this section provides their brief description.

The *Berlin* emotional database (German) [7] was recorded at the Technical University of Berlin and consists of labelled emotional German utterances which were spoken

by 10 actors (5 female). Each utterance has one of the following emotional labels: neutral, anger, fear, joy, sadness, boredom or disgust.

The *SAVEE* (Surrey Audio-Visual Expressed Emotion) corpus (English) [8] was recorded as a part of an investigation into audio-visual emotion classification from four native English male speakers. The emotional label for each utterance
is one of the standard set of emotions (anger, disgust, fear, happiness, sadness, surprise and neutral).

The *LEGO* emotion database (English) [9] comprises non-acted American English utterances extracted from an automated bus information system of the Carnegie Mellon University in Pittsburgh, USA. The utterances are requests to the Interactive Voice Response system spoken by real users with real concerns. Each utterance is annotated with one of the following emotional labels: angry, slightly angry, very angry, neutral, friendly, and non-speech (critical noisy recordings or just silence). In this study different ranges of anger have been merged into a single class and friendly utterances have been deleted. This preprocessing results in a 3-class emotion classification task.

The *VAM* database (German) [10] was created at Karlsruhe University and consists of utterances extracted from the popular German talk-show "Vera am Mittag" (Vera in the afternoon). The emotional labels of the first part of the corpus (speakers 1-19) were given by 17 human evaluators and the rest of the utterances (speakers 20-47) were labelled by 6 annotators on a 3-dimensional emotional basis (valence, activation and dominance). To produce the labels for the classification task we have used just a valence (or evaluation) and an arousal axis. The corresponding quadrant (anticlockwise, starting in the positive quadrant, and assuming arousal as abscissa) can also be assigned emotional labels: happy-exciting, angry-anxious, sad-bored and relaxed-serene.

The *RadioS* database (German) consists of recordings from a popular German radio talk-show. Within this corpus 69 native German speakers talked about their personal troubles. Labelling has been performed by one single evaluator at Ulm University, Germany. One of the following emotional primitives has been set as a label for each utterance: happiness, anger, sadness and neutral.

The *UUDB* (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database (Japanese) [11] consists of spontaneous Japanese human-human speech. The task-oriented dialogue produced by seven pairs of speakers (12 female) resulted in 4,737 utterances in total. Emotional labels for each utterance were created by three annotators on a five-dimensional emotional basis (interest, credibility, dominance, arousal, and pleasantness). For this work, only the pleasantness and arousal axes are used.

III. Conventional and Proposed Approaches

At the beginning of this section there is a description of the conventional models which might be used to determine the most effective classifiers and to compare their performance on the set of corpora introduced earlier.

Then we propose three algorithmic schemes based on collective decision making. These approaches allow for a number of models to be involved in the classification process and to compose the final prediction.

A. Classification with Conventional Models

A list of the classifiers used is presented below [12]. Moreover, there are some details referring to these models.

Multilayer Perceptron (MLP). A feedforward neural network with one hidden layer containing *[(NumberOfFeatures+NumberOfClasses)/2+1]*-neurons is trained with the backpropagation algorithm (BP).

Support Vector Machine (SVM). To design a hyperplane separating sets of examples Sequential Minimal Optimization (SMO) is used for solving the large scale quadratic programming problem.

Linear Logistic Regression (Logit). This linear model describes the relationship between labels and independent variables using probability scores.

Radial Basis Function network (RBF). Gaussian radial basis functions are applied as activation functions in the neural network structure.

Naive Bayes. This supervised learning algorithm is based on Bayes' theorem and uses the "naive" assumption of independence between all pairs of attributes.

Decision trees (J48). Decision trees are generated with the J48 algorithm which is a version of the C4.5 procedure (J48 is an open source implementation of the C4.5 algorithm).

Random Forest. This approach combines the "bagging" idea and the random subspaces method to construct an ensemble (a forest) of random trees.

Bagging. This group of learning meta-algorithms is designed to reduce variance and to avoid overfitting.

Additive Logistic Regression (LogitBoost). The main concept of additive logistic regression pertains to the usage of the boosting approach to design a model.

One Rule (OneR). The idea of the OneR algorithm is related to finding one attribute (the best predictor) which provides the lowest classification error.

B. Collective decision making precedures

There are three concepts allowing the predictions of different classifiers to be taken into consideration while making the final decision [13].

Scheme 1. For each test example it is necessary to determine k-nearest neighbours from the training data set. The prediction of the model that classifies these k-nearest neighbours correctly is used as the final decision. (If several models demonstrate equal effectiveness, choose one of them

randomly). In the experiments conducted, the k-parameter was equal to 3.

Scheme 2. For each test example the engaged models vote for different classes according to their own predictions. The final decision is defined as a collective choice based on the majority rule.

Scheme 3. The previous scheme has one disadvantage: if the number of classes is greater than or equal to the number of classifiers involved (or the number of classifiers is even), a situation whereby several classes receive the majority of votes often occurs. Therefore we combine Schemes 1 and 2 in the following way:

- fulfil the voting procedure as it is described in Scheme 2;
- if several classes have the maximum number of votes, apply Scheme 1.

In all these schemes there is no limitation to the number of classifiers.

IV. Experiments and Results

Firstly, we applied a set of conventional classification models. For each classifier the *F-score* metric was evaluated to estimate the results of the 6-fold cross-validation procedure: the more effective the classifier that we used, the higher F-score value we obtained (Fig. 1-6).

Analysis of the results presented in Fig. 1-6 has exposed that there is no particular model that is equally effective for all of the databases. It can be seen that F-score values vary significantly for different classifiers. Even the best model for a certain corpus might be the worst for another one.

For instance, MLP shows the highest performance on the Berlin corpus, whereas for UUDB it demonstrates the worst results (and vice versa, for the One Rule classifier). Therefore it might be reasonable to engage a few classifiers in the decision making process in order to increase the reliability of the classification technique. Otherwise, the random choice of the classifier may lead to significant performance deterioration.

For the corpora used, the Multilayer Perceptron (MLP), the Support Vector Machine (SVM) and Linear Logistic Regression (Logit) demonstrated rather high performance. Therefore it was decided to involve these classifiers in the proposed schemes of collective decision making. We repeated the 6-fold cross-validation procedure for all of the corpora. The results obtained are presented in Table I.

On the Berlin database all schemes demonstrate high performance. The F-score values obtained with the usage of Schemes 2 and 3 even outperform the best results achieved by MLP (Fig. 1).

TABLE I. F-score values for Collective Decision Making Schemes, %

	Berlin	**SAVEE**	**LEGO**	**VAM**	**UUDB**	**RadioS**
Scheme 1	81.18	61.52	70.52	42.29	37.96	30.68
Scheme 2	84.01	64.33	71.19	50.19	36.41	26.39
Scheme 3	84.23	63.5	71.13	43.69	39.78	26.39

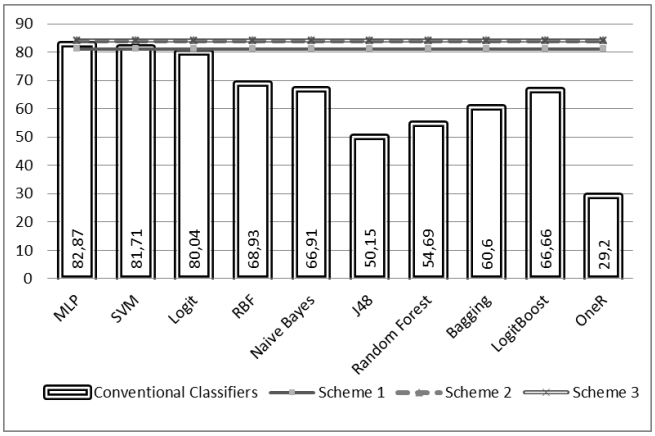

Figure 1. Classification results for Berlin

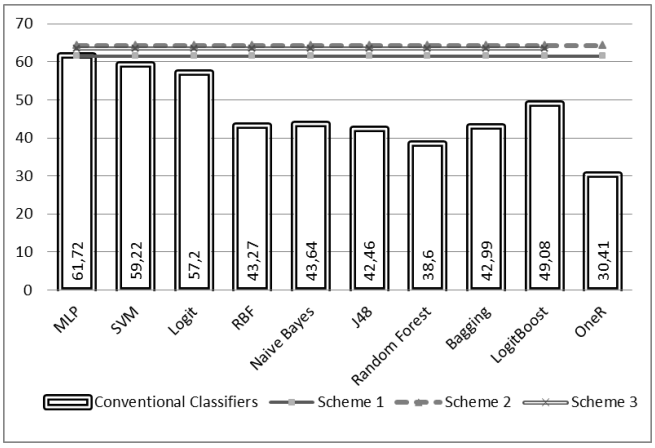

Figure 2. Classification results for SAVEE

The application of the collective decision making procedures to the SAVEE database also leads to increasing F-score values. It was found that due to the usage of Schemes 2 and 3 we managed to increase the F-score value by up to a 4.23% relative improvement (compared with the F-score value attained by MPL) (Fig. 2).

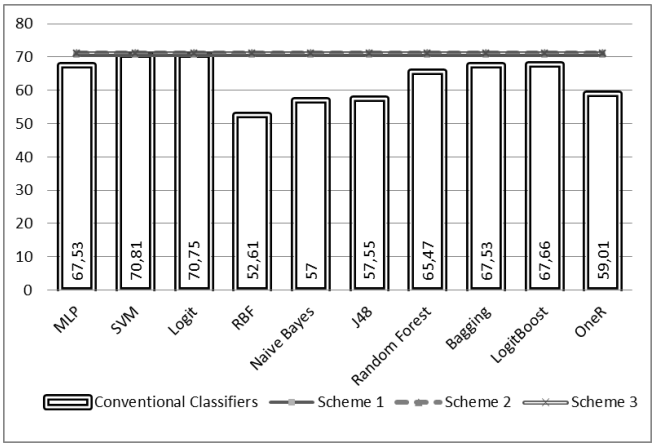

Figure 3. Classification results for LEGO

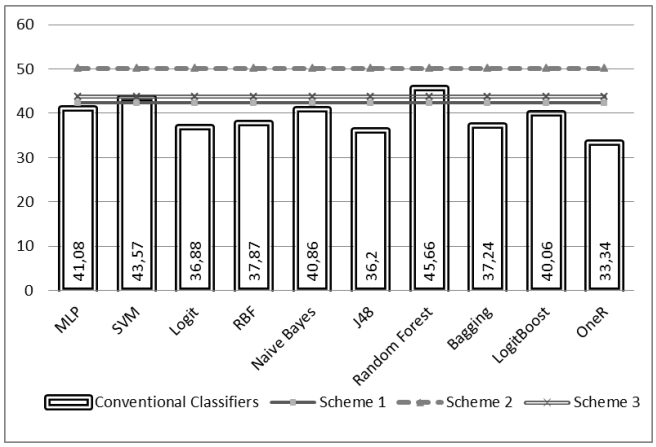

Figure 4. Classification results for VAM

The usage of Schemes 2 and 3 to classify the test examples from the LEGO database allowed us to outperform the results obtained with SVM (Fig. 3).

The best classification result on the VAM corpus provided by Random Forest was exceeded by the application of Scheme 2 (9.93% relative improvement). It is essential to take into account that in this case the most effective classification model (Random Forest) is not involved in the set of classifiers used in the framework of Scheme 2 (MLP, SVM, Linear Logistic Regression). Nevertheless, we attained a significantly better result with classifiers that demonstrated average effectiveness on this corpus (Fig. 4).

Even on the UUDB corpus we obtained rather high F-score values (especially with Scheme 1), although MLP, SVM and Linear Logistic Regression demonstrated the worst results separately (Fig. 5).

In most cases the F-score values achieved by any collective decision making scheme are comparable with the best results provided by the most effective models or, at least, higher than the average F-score value obtained by conventional classifiers (as a case in point, the RadioS database) (Fig. 6).

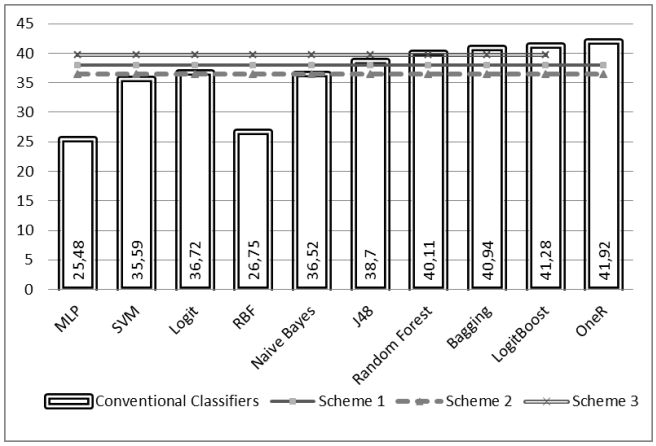

Figure 5. Classification results for UUDB

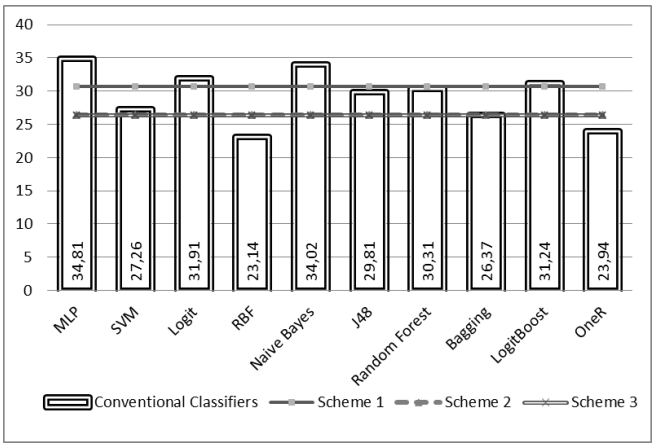

Figure 6. Classification results for RadioS

Based on the experimental results, it might be concluded that on the set of the presented databases Scheme 2 was the most effective for the collective classification process.

V. CONCLUSIONS

In this paper some effective approaches to the emotion recognition problem based on collective decision making are considered. Due to the usage of these techniques it became possible to improve the classification results for most of the corpora (in some cases even by up to a 9.93% relative improvement). Although we have managed to achieve some good results, there are a number of questions. *How many classifiers should we use to provide the most accurate scheme? What kind of models should it be compulsory to include in the ensemble of classifiers?* In other words, there is a necessity to choose appropriate classifiers which should be included in the ensemble automatically. The Genetic Algorithm with binary representation might be used for this purpose.

Moreover, there are some other aspects related to recognition of qualities of the user such as gender and speaker identification. Consequently, the proposed schemes might be applied to solve these problems.

REFERENCES

[1] O.W. Kwon, K. Chan, J. Hao, and T.W. Lee, “Emotion recognition by speech signals,” in INTERSPEECH, 2003.

[2] J. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, “Getting started with susas: a speech under simulated and actual stress database,” in EUROSPEECH, 1997, vol. 97, pp. 1743–1746.

[3] A. Batliner, Ch. Hacker, S. Steidl, E. N’oth, Sh. D’Arcy, M. J. Russell, and M. Wong, “” you stupid tin box”-children interacting with the aibo robot: A cross-linguistic emotional speech corpus,” in LREC, 2004.

[4] P. Boersma, “Praat, a system for doing phonetics by computer,” Glot international, vol. 5, no. 9/10, pp. 341–345, 2002.

[5] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast opensource audio feature extractor”, Proceedings of the international conference on Multimedia, 2010. ACM, pp. 1459–1462.

[6] Ch. Brester, M. Sidorov, E. Semenkin, “Acoustic Emotion Recognition: Two Ways of Features Selection Based on Self-Adaptive Multi-Objective Genetic Algorithm”, Proceedings of the

International Conference on Informatics in Control, Automation and Robotics (ICINCO), 2014, pp. 851-855.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech”, In Interspeech, 2005, pp. 1517–1520.

[8] S. Haq, P. Jackson, “Machine Audition: Principles, Algorithms and Systems, chapter Multimodal Emotion Recognition”, IGI Global, Hershey PA, Aug. 2010, pp. 398–423.

[9] A. Schmitt, S. Ultes, and W. Minker, “A parameterized and annotated corpus of the cmu let’s go bus information system”. Proceedings of International Conference on Language Resources and Evaluation (LREC), 2012.

[10] M. Grimm, K. Kroschel, and S. Narayanan, “The vera am mittag german audio-visual emotional speech database”, In Multimedia and Expo, IEEE International Conference on, IEEE, 2008, pp. 865–868.

[11] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, “Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics”, Speech Communication, 53, 2011.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA Data Mining Software: An Update”, SIGKDD Explorations, Volume 11, Issue 1, 2009.

[13] E.A. Popov, M.E. Semenkina, L.V. Lipinskiy, “Decision making with intelligent information technology ensemble”, Vestnik SibGAU. 2012, № 5 (45), pp. 95–99 (In Russ).