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