Speaker State Recognition with Neural Network-based Classification and Self-adaptive Heuristic Feature Selection

Maxim Sidorov[1], Christina Brester[2], Eugene Semenkin[2] and Wolfgang Minker[1]
[1]*Institute of Communication Engineering, Ulm University, Ulm, Germany*
[2]*Institute of System Analysis, Siberian State Aerospace University, Krasnoyarsk, Russia*
{*maxim.sidorov, wolfgang.minker*}*@uni-ulm.de, abahachy@mail.ru, eugenesemenkin@yandex.ru*

Keywords: Speech-based emotion recognition, speech-based speaker and gender identification, neural network, genetic algorithm-based feature selection, speech corpora analysis.

Abstract: While the implementation of existing feature sets and methods for automatic speaker state analysis has already achieved reasonable results, there is still much to be done for further improvement. In our research, we tried to carry out speech analysis with the self-adaptive multi-objective genetic algorithm as a feature selection technique and with a neural network as a classifier. The proposed approach was evaluated using a number of multi-language speech databases (English, German and Japanese). According to the obtained results, the developed technique allows an increase in emotion recognition performance by up to 6.2% relative improvement in average F-measure, up to 112.0% for the speaker identification task and up to 6.4% for the speech-based gender recognition, having approximately half as many features.

1 INTRODUCTION

Machines are still quite poor at analysing human-human dialogues, meanwhile such a possibility as the automatic understanding of emotional state, gender and speakers itself might be useful in various applications, including the improvement in performance of spoken dialogue systems (SDSs) or the quality monitoring of call-centres. Such tasks as emotion recognition, gender and speaker identification (or recognition) could be conditionally named as *speaker state recognition* or as a part of a *speech analysis* procedure.

Our proposal is a usage of the multi-objective genetic algorithm (MOGA), which is a baseline heuristic method of pseudo-boolean optimization, in order to select informative features based on two metrics characterising the relevancy of the reduced data sets. In fact, the efficiency of GA applications dramatically depends on its parameters and setting reasonable parameters requires the expert knowledge of an end-user. Therefore, we also proposed here a self-adaptive scheme of MOGA, which removes the necessity of choosing the algorithm's parameters.

It turns out that by using the proposed techniques we could significantly improve the performance of such procedures as emotion recognition, and speaker and gender recognition based on speech signals.

The rest of the paper is organized as follows: Significant related work is presented in Section 2, and Section 3 describes the applied corpora and renders their differences. Our approach to improving the automated dialogue analysis is proposed in Section 4, and the results of numerical evaluations are in Section 5. The conclusion and future work are described in the Section 6.

2 SIGNIFICANT RELATED WORK

One of the pilot experiments which dealt with speech-based emotion recognition has been presented by Kwon et al. (Kwon et al., 2003). The authors compared the emotion recognition performance of various classifiers: support vector machine, linear discriminant analysis, quadratic discriminant analysis and hidden Markov model on SUSAS (Hansen et al., 1997) and AIBO (Batliner et al., 2004) databases of emotional speech. The following set of speech 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, correspondingly) using

Database	Language	Full length (min.)	File level duration		Num. of Emotions	Num. of Speakers	Num. of Gender
			Mean(sec.)	Std. (sec.)			
Berlin	German	24.7	2.7	1.02	7	10	2
SAVEE	English	30.7	3.8	1.07	7	4	1
VAM	German	47.8	3.02	2.1	4	47	2
UUDB	Japanese	113.4	1.4	1.7	4	10	2
LEGO	English	118.2	1.6	1.4	5	291	2

Table 1: Databases description.

Gaussian support vector machine.

The authors in (Gharavian et al., 2012) highlighting the importance of feature selection for the ER used the fast correlation-based filter feature selection method. A fuzzy ARTMAP neural network (Carpenter et al., 1992) was used as an algorithm for emotion modelling. The authors have achieved an accuracy of over 87.52% for emotion recognition on the FARSDAT speech corpus (Bijankhan et al., 1994).

While our research of identifying cross-lingual salient features includes more different emotions, Polzehl et al. (Polzehl et al., 2011) only focused on the emotion anger. The authors analysed two different anger corpora of German and American English to determine the optimal feature set for anger recognition. The German database contains 21 hours of records from a German Interactive Voice Response (IVR) portal offering assistance troubleshooting. For each utterance, three annotators assigned one of the following labels: not angry, not sure, slightly angry, clear anger, clear rage, and garbage. Garbage-marked utterances are non-applicable, e.g., contain silence or critical noise. The English corpus originates from an US-American IVR portal which is capable of fixing Internet-related problems. Three labelers divided the corpus into angry, annoyed, and non-angry turns. A total of 1,450 acoustic features and their statistical description (e.g, means, moments of first to fourth order, the standard deviation) have been extracted from the speech signal. The features are divided into seven general groups: pitch, loudness, MFCC, spectrals, formants, intensity, and other (e.g., harmonics-to-noise). Analysing each feature group separately, the authors achieved in a baseline approach without further feature selection a maximal $f1$ score of 68.6 with 612 MFCC-based features of the German corpus and a maximal $f1$ score of 73.5 with 171 intensity-based features of the English corpus.

3 DATABASES

For the study, a number of speech databases have been applied for the dialogue analysis. In this Section, a brief description of each corpus is provided.

The *Berlin* emotional database (Burkhardt et al., 2005) was recorded at the Technical University of Berlin and consists of labeled 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, and disgust.

Haq and Jackson (Haq and Jackson, 2010) recorded the *SAVEE* (Surrey Audio-Visual Expressed Emotion) corpus for research on audio-visual emotion classification from four native English male speakers. The emotional label for each acted utterance is one of the standard emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral).

The *RadioS* database 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 done by only one 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 (Mori et al., 2011) 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 pleasantness (or evaluation) and the arousal axes are used.

Based on the popular German TV talk-show "Vera am Mittag" (Vera in the afternoon), the *VAM* database (Grimm et al., 2008) has been created at the Karlsruhe Institute of Technology. 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 six annotators.

The *LEGO* emotion database (Schmitt et al., 2012) 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 single class and *friendly* utterances have been deleted. This preprocessing results in a 3-classes emotion classification task.

A statistical description of the used corpora can be found in Table 1.

4 STATISTICAL APPROACH

The proposed approach combines two basic techniques.

The first one refers to the data preprocessing stage performed by MOGA with binary representation. It operates with two criteria which are the Intra-class distance (IA) and the Inter-class distance (IE):

$$IA = \frac{1}{n}\sum_{r=1}^{k}\sum_{j=1}^{n_r} d({p_j}^r, p_r) \rightarrow min, \quad (1)$$

$$IE = \frac{1}{n}\sum_{r=1}^{k} n_r d(p_r, p) \rightarrow max, \quad (2)$$

where ${p_j}^r$ is the j-th example from the r-th class, p is the central example of the data set, $d(.,.)$ denotes the Euclidian distance, p_r and n_r represent the central example and the number of examples in the r-th class.

Each attribute corresponds to a particular gene in a binary string: boolean true codes the selected informative feature whereas boolean false means the unselected one. A search of non-dominated points is implemented via the self-adaptive modification of the Strength Pareto Evolutionary Algorithm (SPEA) (Zitzler and Thiele, 1999). As a result of this stage we get a set of binary candidate solutions which cannot be compared with each other. The second main procedure is an application of a supervised learning algorithm that is used to predict class values for test examples. In this research, the Multilayer Perceptron (MLP) is applied as a classifier.

The following subsections describe the approaches used in more details.

4.1 Feature selection

The feature selection component works as follows:

1. Generate an initial population $P_t, t = 0$, uniformly in the binary search space.
2. Evaluate criteria values for each individual from P_t.
3. Compose the outer set: copy the individuals non-dominated over P_t into the intermediate outer set $\bar{P}'$, delete the individuals dominated over $\bar{P}'$ from the intermediate outer set. If the capacity of the set $\bar{P}'$ is more than the fixed limit $\bar{N}$, then apply the clustering algorithm (hierarchical agglomerative clustering). Compile the outer set $\bar{P}_{t+1}$ with the individuals from $\bar{P}'$.
4. Apply genetic operators (selection, crossover, mutation) in order to generate new solutions.
5. Check the stop-criterion: if it is *true*, then complete the working of MOGA otherwise continue from the second step.

In Step 4 the self-adaptive crossover and mutation operators were implemented. Recombination is based on the *co-evolution* idea (Potter and De Jong, 1994): one-point, two-point and uniform crossover operators compete for resources, i.e. for the amount of individuals in the current population generated by the particular type of recombination. The allocation of resources occurred in every T-th generation called *time of adaptation* via paired comparisons of operator's *fitness-values* $q_j, j = \overline{1,n}$:

$$q_j = \sum_{l=0}^{T-1} \frac{T-l}{l+1} \cdot b_j, \quad (3)$$

where $l = 0$ states the latest generation in the adaptation interval, $l = 1$ corresponds to the previous generation, etc. b_j is defined as following:

$$b_j = \frac{p_j}{|\bar{P}|} \cdot \frac{N}{n_j}, \quad (4)$$

where p_j is a number of individuals in the current outer set generated with the j-th type of recombination operator, $|\bar{P}|$ is the outer set size, n_j is the amount of individuals in the current population generated with the j-th type of crossover, N is the population size. A parameter *penalty* denotes the amount of resources that is reallocated from the genetic operator with lower *fitness* to the genetic operator with higher *fitness*. Furthermore, the decreasing of resources must be limited with a parameter named *social card* that allows the genetic operator's diversity to be maintained. Initially, all types of genetic operators have an equal amount of resources.

The probability of self-adaptive mutation is assigned according to the rule (Daridi et al., 2004):

$$p_m = \frac{1}{240} + \frac{0.11375}{2^t} \quad (5)$$

where t is the current generation number. Eventually, the final solution is determined as a point from the Pareto's set approximation with the lowest value of the relative classification error.

4.2 Classification

The MLP-model is used to define class values for test examples through involving reduced datasets which correspond to binary strings obtained at the previous stage. First, each non-dominated binary solution is decoded into the set of selected features. Then a number of neural networks are trained using reduced datasets. Finally, the class value for any test example is determined as a collective decision based on the majority rule engaging predictions from all trained classifiers.

5 EVALUATION AND RESULTS

To estimate the performance of the MOGA usage in speech-based recognition problems, a number of experiments were conducted. The proposed approach was applied to all databases described above, which were parametrised by the baseline (384-dimensional) feature sets.

The baseline sets used for the Interspeech 2009 Emotion Challenge (384 features) (Kockmann et al., 2009), which were extracted with the openSMILE (Eyben et al., 2010) system, were involved in this study.

In order to provide statistical comparison of the proposed methods, the classification procedure was conducted several times. Firstly, it was fulfilled by MLP (without feature selection) and secondly by the proposed approach. The following baseline methods of feature selection have also been applied in this study: feature selection with Principle Component Analysis (PCA), Information Gain Ratio (IGR) as it was done in (Polzehl et al., 2011) and the conventional genetic algorithm-based feature selection used in the *wrapper* mode. To reduce the dimensionality of data sets by PCA, α equal to 0.99 was set as a variance threshold. In order to determine the number of selected features using the IGR method, a grid-optimization technique with 10 steps has been applied, i.e. first 39, 78, 116, ..., 384 features ordered by the IGR procedure have been included to conduct classification experiments with MLP.

To obtain more statistically significant results, the validation procedure has been used with the number of folds equal to 6 for all considered problems on each database. There are average results over all runs in Tables 2, 3 and 4, with the average number of features in parentheses. The F-measure has been chosen as the main criterion of the classification performance. The columns entitled *MLP Baseline* contain results, which were achieved with the baseline 384-dimensional feature set without feature selection. Similarly, the columns titled *PCA MLP* and *IGR MLP* contain results obtained with PCA and IGR feature selection procedures correspondingly. The columns titled *GA MLP* show the classification performance attained with the conventional genetic algorithm-based feature selection (wrapper approach). Finally, the proposed MOGA-based feature selection is shown in the *SPEA MLP* columns.

6 CONCLUSION AND FUTURE WORK

An application of the proposed hybrid system in order to select the most representative features and maximize the accuracy of particular tasks could decrease the number of features and increase the accuracy of the system simultaneously. In most of the cases, the MOGA-based technique outperforms baseline results. It should be noted that the number of selected features using the IGR method is quite high. It means that in some cases the number of features was equal to 384, i.e. an optimal modelling procedure has been conducted without feature selection at all.

While MLP has already provided reasonable results for automated dialogue analysis, we are still examining its general appropriateness. The usage of other possibly more accurate classifiers may improve the performance of this system. Furthermore, dialogues do not only consist of speech, but also of a visual representation. Hence, an analysis of pictures or even video recordings may also improve the performance of dialogue analysis.

REFERENCES

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M. J., and Wong, M. (2004). " you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *LREC*.

Bijankhan, M., Sheikhzadegan, J., Roohani, M., Samareh, Y., Lucas, C., and Tebyani, M. (1994). Farsdat-the speech database of farsi spoken language. In *the Proceedings of the Australian Conference on Speech Science and Technology*, volume 2, pages 826–830.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F.,

Database	Baseline MLP	PCA MLP	IGR MLP	GA MLP	SPEA MLP
SAVEE	61.72	20.19	62.66 (262.8)	57.58 (191.8)	**63.58 (178.3)**
Berlin	**82.87**	30.08	82.60 (294.7)	79.88 (183)	82.26 (182.2)
VAM	41.08	29.49	40.75 (218)	**41.89 (194.2)**	43.05 (178.7)
RadioS	34.81	25.84	32.11 (345.8)	33.79 (180.8)	**35.23 (184.6)**
LEGO	67.53	39.78	69.86 (192.7)	69.82 (188)	**71.70 (180.9)**
UUDB	25.48	23.61	**36.78 (218.2)**	33.34 (196.7)	34.58 (179.2)

Table 2: Average F-measure for emotion recognition.

Database	Baseline MLP	PCA MLP	IGR MLP	GA MLP	SPEA MLP
SAVEE	100	57.72	99.80 (218.2)	99.80 (192.5)	**100 (178.7)**
Berlin	88.41	16.81	**88.97 (339.3)**	86.92 (194.5)	87.39 (182.4)
VAM	78.09	8.04	**78.40 (345.8)**	73.15 (195.2)	78.11 (185.3)
RadioS	**95.34**	6.50	95.14 (288.2)	93.62 (194.9)	94.56 (194.4)
UUDB	41.12	14.71	85.94 (301)	82.33 (185)	**87.18 (183.2)**

Table 3: Average F-measure for speaker identification.

Database	Baseline MLP	PCA MLP	IGR MLP	GA MLP	SPEA MLP
Berlin	97.56	75.45	97.18 (256.3)	**97.74 (189.8)**	96.20 (169.8)
VAM	93.33	78.97	93.13 (262.7)	93.34 (187.8)	**93.64 (170.1)**
RadioS	**97.15**	72.21	97.03 (275.7)	96.51 (195.7)	97,11 (173.7)
LEGO	80.86	64.90	84.94 (205.5)	83.06 (192.7)	**86.07 (179.5)**
UUDB	97.32	78.23	**98.13 (269)**	97.76 (187.5)	98.04 (178.4)

Table 4: Average F-measure for gender recognition.

and Weiss, B. (2005). A database of german emotional speech. In *Interspeech*, pages 1517–1520.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., and Rosen, D. B. (1992). Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. *Neural Networks, IEEE Transactions on*, 3(5):698–713.

Daridi, F., Kharma, N., and Salik, J. (2004). Parameterless genetic algorithms: review and innovation. *IEEE Canadian Review*, (47):19–23.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM.

Gharavian, D., Sheikhan, M., Nazerieh, A., and Garoucy, S. (2012). Speech emotion recognition using fcbf feature selection method and ga-optimized fuzzy artmap neural network. *Neural Computing and Applications*, 21(8):2115–2126.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The vera am mittag german audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 865–868. IEEE.

Hansen, J. H., Bou-Ghazale, S. E., Sarikaya, R., and Pellom, B. (1997). Getting started with susas: a speech under simulated and actual stress database. In *EUROSPEECH*, volume 97, pages 1743–46.

Haq, S. and Jackson, P. (2010). *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pages 398–423. IGI Global, Hershey PA.

Kockmann, M., Burget, L., and Černockỳ, J. (2009). Brno university of technology system for interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (2003). Emotion recognition by speech signals. In *INTERSPEECH*.

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

Polzehl, T., Schmitt, A., and Metze, F. (2011). Salient features for anger recognition in german and english ivr portals. In Minker, W., Lee, G. G., Nakamura, S., and Mariani, J., editors, *Spoken Dialogue Systems Technology and Design*, pages 83–105. Springer New York. 10.1007/978-1-4419-7934-6_4.

Potter, M. A. and De Jong, K. A. (1994). A cooperative coevolutionary approach to function optimization. In *Parallel Problem Solving from Nature–PPSN III*, pages 249–257. Springer.

Schmitt, A., Ultes, S., and Minker, W. (2012). A parameterized and annotated corpus of the cmu let's go bus information system. In *International Conference on Language Resources and Evaluation (LREC)*.

Zitzler, E. and Thiele, L. (1999). Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. *Evolutionary Computation, IEEE Transactions on*, 3(4):257–271.