Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm

Maxim Sidorov[1], Christina Brester[2], Wolfgang Minker[3], Eugene Semenkin[4]
[1,3] Institute for Communications Engineering, University of Ulm, Germany
[2,4] Insitute for System Analysis, Siberian State Aerospace University, Krasnoyarsk, Russia
[1,3]{maxim.sidorov, wolfgang.minker}@uni-ulm.de, [2]abahachy@mail.ru, [4]eugenesemenkin@yandex.ru

Abstract

Automated emotion recognition has a number of applications in Interactive Voice Response systems, call centers, etc. While employing existing feature sets and methods for automated emotion recognition has already achieved reasonable results, there is still a lot to do for improvement. Meanwhile, an optimal feature set, which should be used to represent speech signals for performing speech-based emotion recognition techniques, is still an open question. In our research, we tried to figure out the most essential features with self-adaptive multi-objective genetic algorithm as a feature selection technique and a probabilistic neural network as a classifier. The proposed approach was evaluated using a number of multi-languages databases (English, German), which were represented by 37- and 384-dimensional feature sets. According to the obtained results, the developed technique allows to increase the emotion recognition performance by up to 26.08% relative improvement in accuracy. Moreover, emotion recognition performance scores for all applied databases are improved.

Keywords: speech-based emotion recognition, feature selection techniques, multi-criteria genetic algorithm

1. Introduction

Machines are still quite bad at recognizing human emotions, meanwhile such an opportunity might be useful in various applications, including improvement of the spoken dialogue systems (SDSs) performance or call centers quality monitoring.

Speech-based emotion recognition (ER) is a classification problem which can be solved by various methods based on the supervised learning approach. One may extract a lot of numerical features out of speech waveforms. A feature selection procedure results in trade-off between time-consuming feature extraction and the accuracy of the model. However, some of features could be highly-correlated or their variability level could be dramatically low, therefore some attributes could not bring a beneficial impact to the system or they even could lower its performance. That is why, an effective feature set for the ER task (but also for the problems of speech-based speaker (SI) and gender (GI) identification) should be both representative and compact. Moreover, removing irrelevant attributes could significantly improve the performance of speech-based ER, SI and GI models.

Our proposal uses a multi-objective genetic algorithm (MOGA), which is a heuristic algorithm of pseudo-boolean optimization, in order to maximize the ER accuracy and minimize the number of items in feature sets simultaneously. We also propose here a self-adaptive scheme of MOGA, which exempts from the necessity of algorithm's parameters choosing.

In this study, two feature sets of speech signals (37- and 384-dimensional) have been used to represent 3 emotional speech databases of 2 languages (English, German), acted and non-acted recordings. It turns out, that the usage of a self-adaptive MOGA could improve the accuracy of the ER procedure up to 26.08% (for the 384-dimensional feature set) for some of the databases.

The rest of the paper is organized as follows: the 2. Section presents significant related work. The 3. Section describes the applied corpora and renders their differences. Our approach to automated emotion recognition improvement is proposed in the 4. Section having its results of numerical evaluations in the 5. Section. Conclusion and future work are described in the 6. Section.

2. Significant Related Work

One of the pilot experiments which deals with speech-based emotion recognition is in (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 the 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 the accuracy (70.1% and 42.3% on the databases, correspondingly) using Gaussian support vector machine.

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

Another approach for improving emotion recognition has been proposed by Polzehl et al. (2011) by adding linguistic information, e.g., Bag-of-Words or Self-Referential Information. Evaluation with three different databases showed that fusion at the decision level adding confidence scores slightly improves the overall scores. However, evaluating acoustic and linguistic models on separate levels showed the dominance of acoustic models.

Table 1: Databases description

Database	Language	Full length (min.)	Number of emotions	File level duration		Emotion level duration		Notes
				Mean(sec.)	Std. (sec.)	Mean (sec.)	Std. (sec.)	
Berlin	German	24.7	7	2.7	1.02	212.4	64.8	Acted
SAVEE	English	30.7	7	3.8	1.07	263.2	76.3	Acted
VAM	German	47.8	4	3.02	2.1	717.1	726.3	Non-acted

3. Corpora

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

Berlin 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 or disgust.

SAVEE The SAVEE (Surrey Audio-Visual Expressed Emotion) corpus (Haq and Jackson, 2010) 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).

VAM The VAM database (Grimm et al., 2008) was created at the 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 labeled by 6 annotators on the 3-dimensional emotional basis (valence, activation and dominance). To produce the labels for classification task we have used just valence (or evaluation) and arousal axis. The corresponding quadrant (counterclockwise, starting in positive quadrant, assuming arousal as abscissa) can also be assigned emotional labels: happy-exciting, angry-anxious, sad-bored and relaxed-serene (Schuller et al., 2009).

Two corpora (Berlin, SAVEE) consist of acted emotions, whereas VAM database comprises real ones. Acted and non-acted emotions have been considered for the German language, but only non-acted emotions are in English utterances. In comparison with Berlin and SAVEE corpora, VAM database is highly unbalanced (see Emotion level duration columns in Table 1).

Emotions themselves and their evaluations have subjective nature. That is why it is important to have at least several evaluators of emotional labels. Even for humans it is not always evident to make a decision about an emotional label. Each study, which proposes an emotional database, provides also an evaluators' confusion matrix and statistical description of their decisions.

There is a statistical description of the used corpora in Table 1.

4. Statistical Approach

The baseline set of the 37-dimensional feature vector, as well as the feature set used for the Interspeech 2009 Emotion Challenge (384 features) (Kockmann et al., 2009), which were extracted with the Praat (Boersma, 2002) and the openSMILE (Eyben et al., 2010) systems, were used in this study.

Average values of the following speech signal features were included into the baseline set of features: power, mean, root mean square, jitter, shimmer, 12 MFCCs and 5 formants. Mean, minimum, maximum, range and deviation of the following features have also been used: pitch, intensity and harmonicity (the 37-dimensional feature vector for one speech signal file, in total).

In this investigation a probabilistic neural network (PNN) was used as a classification algorithm (Specht, 1990) for building emotion models.

A genetic algorithm (GA), which is an effective pseudo-boolean optimization procedure, was chosen to solve the multi-criteria problem. A multi-criteria GA operates with a set of binary vectors coding the subsets of relevant features, where boolean *false* corresponds to non-essential attributes and *true* corresponds to essential ones.

In the Strength Pareto Evolutionary Algorithm (SPEA) (Zitzler and Thiele, 1999), non-dominated points are stored in the limited capacity archive named *an outer set*. The content of this set is upgraded through the algorithm execution and as a result we have an approximation of the Pareto set.

The investigated approach works as follows:

Input:

· the labeled database divided into training and validation sets;
· N: the population size;
· $\bar{N}$: the maximum number of non-dominated points stored in the outer set;
· M: the allowable number of generations.
Parameters of the self-adaptive crossover operator:
· *penalty*: a fee size for recombination types defeated in paired comparisons;
· *time of adaptation* T: the number of generations fulfilled before every resources allocation among recombination types;
· *social card*: the minimum allowable size of the subpopulation generated with a crossover operator type;
· available recombination types:
J={0| *single-point crossover*; 1| *two-point crossover*; 2| *uniform crossover*};

- n_j, $j \in J$: the amount of individuals in the current population generated by the j-th type of crossover.

Output:

- $PS = \bar{P} = \{\bar{x}_i\}$, $1 \leq i \leq \bar{N}$: the approximation of the Pareto set;
- PF: the approximation of the Pareto frontier.

Step 1. Initialization
Generate an initial population $P_t = \{x_i\}$, $t = 0$, $i = \overline{1, N}$, uniformly in the binary search space: probabilities of boolean *true* and *false* assignments are equal. Define initial values $n_j = \frac{N}{|J|}$.

Step 2. Evaluation of criteria values
For each individual from P_t, do:

2.1. Compile the feature subsystem from the database corresponding to the current binary string.
2.2. Perform classification on the obtained feature subsystem by PNN learned on the training data set.
2.3. Set the first individual's criterion value as the relative classification error on the validation sample.
2.4. Set the second individual's criterion value as the number of *true* genes in the binary string.

Step 3. Composing the outer set

3.1. Copy the individuals non-dominated over P_t into the intermediate outer set $\bar{P}'$.
3.2. Delete the individuals dominated over $\bar{P}'$ from the intermediate outer set.
3.3. If the capacity of the set $\bar{P}'$ is more than the fixed limit $\bar{N}$, apply the clustering algorithm (hierarchical agglomerative clustering).
3.4. Compile the outer set $\bar{P}_{t+1}$ with the individuals from $\bar{P}'$.

Step 4. Fitness-values determination
Calculate fitness-values for individuals both from the current population and from the outer set.

Step 5. Generation of new solutions
Set $j = 0$. For each j - realized recombination type, $j \in J$, do:

1) Set $k = 0$ and repeat:
2) Select two individuals from the united set $\widehat{P} = \bar{P}_{t+1} \bigcup P_t$ by 2-tournament selection.
3) Apply a current type of recombination to individuals chosen in step (2).
4) Perform a mutation operator: the probability p_m is determined according to the rule (Daridi et al., 2004):

$$p_m = \frac{1}{240} + \frac{0.11375}{2^t} \tag{1}$$

where t is the current generation number.

5) If $k = n_j$, then $j = j + 1$, otherwise $k = k + 1$.

Step 6. Stopping Criteria
If $t = M$, then stop with the outcome $PS = \bar{P}_{t+1}$, otherwise $t = t + 1$ and go to the next step.

Step 7. Resources allocation
If t is a multiple of T, do:

7.1. Determine *fitness*-values q_i for all $j \in J$:

$$q_j = \sum_{l=0}^{T-1} \frac{T - l}{l + 1} \cdot b_j, \tag{2}$$

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

$$b_j = \frac{p_j}{|\bar{P}|} \cdot \frac{N}{n_j}, \tag{3}$$

where p_j is the amount of individuals in the current outer set generated with the j-th type of recombination operator, $|\bar{P}|$ is the outer set size.

7.2. Compare all crossover operator types in pairs based on their *fitness*-values. Determine s_j is the size of a resource given by the j-th recombination type to those which won:

$$s_j = \begin{cases} 0, & \text{if } n_j \leq social\_card \\ int\left(\frac{n_j - social\_card}{n_j}\right), & \text{if } (n_j - h_j \cdot penalty) \\ & \quad \leq social\_card \\ penalty, & \text{otherwise} \end{cases} \tag{4}$$

where h_j is the number of losses of the j-th operator in paired comparisons.

7.3. Redistribute resources n_j based on s_j values, $j \in J$. Go to **Step 2**.

In **Steps 4** and **5** standard SPEA schemes of fitness assignment and selection are used.
The final solution is determined as a point from the Pareto set approximation PS with the lowest value of the relative classification error.

5. Evaluation and Results

To investigate the improvement of using the MOGA-based feature selection, the emotion recognition procedure has been conducted on both the baseline feature set (37-dimensional) and extended (384-dimensional) one. The first set consists of the most popular features for emotion recognition (cf. (Schmitt et al., 2009)) and has been chosen in order to perform baseline experiments. Obtained results have been compared with ones, which were achieved by performing speech-based emotion recognition procedure with and without proposed multi-objective genetic-algorithm-based feature selection technique on extended feature set. Furthermore, obtained results were compared with ones, which were achieved by applying state-of-the-art Principal Component Analysis (PCA) (Pearson, 1901) feature set reduction.
Dividing the data into the training and the testing sets, the training set was used to create and train a PNN-based emotion model (Experiment 1), as well as for the procedure of feature selection (Experiment 2). During the 2nd Experiment, the used feature set was coded with the boolean vector (*true* corresponds to the essential attribute, *false* to the

Table 2: Evaluation Result (mean / standard deviation) with the baseline and IS'09 feature sets: Accuracy of the baseline system (37 features), Accuracy of PNN (Without GA), Experiment with selected by proposed method features (GA), having the average number of selected features in parentheses, relative (comparing with Baseline results) improvement in per cent (Gain). Significant differences are marked with ** ($\alpha < 0.0001$) and * ($\alpha < 0.005$) using the T-test (Student, 1908).

Database	Baseline	Without GA	GA (Num.)	Gain
	37-dimensional	384-dimensional		
Berlin	56.68/3.37	58.90/4.89	71.46/5.16 (68.4)	26.08 **
SAVEE	41.64/2.88	47.32/2.76	48.41/2.74 (84.14)	16.26 **
VAM	68.01/1.71	67.07/2.63	70.63/1.35 (64.83)	3.85 *

unessential one), the corresponding representation of training samples was used in order to create a PNN-based emotion model for each individual of the self-adaptive MOGA. To assess the model's predictive ability the validation sample was generated as 20% per cent of the training data set. During experiments it was revealed that the PNN performance depended on the sample division significantly. Therefore, we produced the averaged estimations of the relative classification error for all individuals through multiple running (15 times) with the random data division. The testing set was used to get the final assessment of adjusted PNN-based emotion model (for both Experiments) with the feature set selected by MOGA (only for the 2nd Experiment).

In order to generate more statistically significant results, the complete classification process was run 10 times for each database. For each run, the databases were randomly divided into the training and testing sets (70–30% correspondingly). The final results are shown in Table 2. These results are calculated taking the average of all runs. The first two columns correspond to the PNN-based ER accuracy, which was achieved without feature selection procedure (Experiment 1, baseline for the 37-dimensional feature set, and the 384-dimensional extended one). In the third column, the accuracy of the emotion recognition system using the multi-criteria GA-based feature selection with the average number of selected features. The next column contains the relative accuracy improvement, comparing with the baseline system performance (the 37-dimensional feature set without feature selection procedure).

Table 3 shows the results comparison of PCA- and proposed MOGA-based feature selection/reduction methods. The 0.95 variance threshold was performed on each iteration of PCA in order to create a feature set for training and testing of PNN-based emotional models. The results clearly showed that the proposed method could produce fewer features and the higher ER accuracy simultaneously.

The results of baseline ER as well as ER with feature sets selected by MOGA- and PCA-based feature selection techniques have been examined for significance using the t-test (Student, 1908) for comparing the results of each of the 10 runs of the experiments. All differences are significant with at least $\alpha < 0.005$.

As a result, it can be concluded, that the proposed technique could significantly improve the accuracy of the ER procedure (up to 26.08% of the relative improvement) and essentially decrease the number of attributes in the feature set for the involved corpora (Berlin, VAM and SAVEE).

The realized method was run with the following values of parameters: an adaptive SPEA had 100 generations and 100 individuals, the size of the outer set was equal to 50, adaptation interval, penalty and social card were equal to 5, 10 and 10 correspondingly. As it was mentioned above, values of the classification error criteria were estimated with the PNN-based classifier for all individuals in each generation. For each run the Pareto set and frontier approximations were defined and as a result the final solution was determined as the point from the Pareto set with the highest value of the ER accuracy.

Table 3: Evaluation Result (mean / standard deviation) with the IS'09 feature set: Accuracy of the PNN-based ER with selected by PCA features (PCA-PNN), having an average number of selected features in parentheses (Num.), relative (comparing with proposed technique) improvement in per cent (Gain). Significant differences are marked with ** ($\alpha < 0.0001$) using the T-test (Student, 1908).

Database	PCA-PNN (Num.)	Gain
Berlin	43.66/2.95 (129.3)	-38.90 **
SAVEE	26.53/1.12 (123.6)	-45.19 **
VAM	59.41/4.49 (148.6)	-15.88 **

6. Conclusion and Future Work

An application of the PNN-MOGA hybrid system for selecting the most representative features and maximizing the accuracy of the supervised learning algorithm could decrease the number of features from 384 to 43 and increase the ER accuracy up to 26.08 % for some of the corpora. Moreover, the proposed method outperforms the PCA-based feature reduction technique in terms of the ER performance achieved by PNN-based models.

The next step is to analyze the frequency of chosen features to create the most appropriate set for the speech-based emotion recognition task. Please, note that the proposed approach also could be used for the speaker and gender identification.

While PNN already provides reasonable results for emotion recognition, we still examine its general appropriateness. The usage of more accurate classifiers might improve the performance of the system. Furthermore, dialogues do not only consist of speech, but also of a visual representation. Hence, an analysis of pictures or even video recordings might also improve the ER performance.

7. 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.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

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

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*.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Polzehl, T., Schmitt, A., Metze, F., and Wagner, M. (2011). Anger recognition in speech using acoustic and linguistic cues. *Speech Communication*, Special Issue: Sensing Emotion and Affect - Facing Realism in Speech Processing.

Schmitt, A., Heinroth, T., and Liscombe, J. (2009). On nomatchs, noinputs and bargeins: Do non-acoustic features support anger detection? In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, SigDial Conference 2009*, London (UK), September. Association for Computational Linguistics.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009). Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 552–557. IEEE.

Specht, D. F. (1990). Probabilistic neural networks. *Neural networks*, 3(1):109–118.

Student. (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

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.