Using Adaptive Genetic Algorithms to Improve Speech Emotion Recognition

Mohammad H. Sedaaghi*, Constantine Kotropoulos†, and Dimitrios Ververidis†
*Dept. of Electrical Engineering, Sahand University of Technology, Tabriz, Iran Email: sedaaghi@sut.ac.ir
†Dept. of Informatics, Aristotle Univ. of Thessaloniki, Box 451, Thessaloniki 54124, Greece,
Email: {costas, jimver}@aiia.csd.auth.gr

***Abstract*— In this paper, adaptive genetic algorithms are employed to search for the worst performing features with respect to the probability of correct classification achieved by the Bayes classifier in a first stage. These features are subsequently excluded from sequential floating feature selection that employs the probability of correct classification of the Bayes classifier as criterion. In a second stage, adaptive genetic algorithms search for the worst performing utterances with respect to the same criterion. The sequential application of both stages is demonstrated to improve speech emotion recognition on the Danish Emotional Speech database.**

I. Introduction

Vocal emotions form an important part of multimodal human computer interaction [1]. Several recent surveys are devoted to the analysis and synthesis of speech emotions from the point of view of pattern recognition and machine learning as well as psychology [2], [3].

In this paper, we build on the earliest ‘discrete’ theories of emotion (stemming from Darwin’s work in 1872) that assumes the existence of a small number of emotions, such as happiness, sadness, fear, anger, surprise, and disgust [4]. These emotions are also terms as basic emotions, i.e. emotions that are universal and primitive. On the one hand, such a theory is in par with neurophysiological and neuroimaging evidence suggesting that the human brain contains facial expression recognition detectors specialized for specific discrete emotions [5]. Fear-specific responses within the amygdalae were reported for vocal emotional expressions as well [6]. However, it is unsettled to which extend exact localization of cerebral activation during comprehension of emotional prosody is linked to specific emotional categories [7]. On the other hand, behavioral evidence is consistent with some form of lower order dimensional representation of emotions that reflects subjective aspects of behavior such as positive vs. negative and active vs. passive [5]. The so-called dimensional approach is another early model for emotion proposed by Wundt in 1874 [8]. This dichotomy is evident in speech emotion classification literature, where researchers adopt either the discrete case [9]–[12] or work on the continuous arousal-valence space [13], [14], to mention a few.

Feature selection is essentially an optimization problem that involves searching the space of possible feature subsets to find one subset that is optimal (or near-optimal) with respect to a certain criterion [15], [16]. Every feature subset selection algorithm contains two main parts: (1) the search strategy employed to select the feature subsets and (2) the evaluation method applied to test their goodness and fitness based on some criteria. Search strategies can be classified into one of the following three categories: (1) optimal, (2) heuristic, and (3) randomized. Exhaustive search is the most straightforward approach to optimal feature selection. However, since the number of possible subsets grows exponentially, exhaustive search becomes impractical even for moderate feature numbers. The only optimal feature selection method, which avoids the exhaustive search is based on the branch and bound algorithm [17]. Sequential forward selection (SFS) and sequential backward selection (SBS) are two well-known heuristic suboptimal feature selection schemes. Combining SFS and SBS gives birth to plus l-take away r feature selection. Sequential forward floating search (SFFS) and sequential backward floating search (SBFS) are generalizations of the plus l-take away r method, where l and r are determined automatically and updated dynamically [18]. SFFS is found to dominate among 15 feature selection methods in terms of classification error and run time on a 2-class, 20-dimensional, multivariate Gaussian data set [16]. Feature selection can be performed with respect to properties, such as orthogonality, correlation, mutual information, etc.

Evolutionary algorithms are random search algorithms. Among them, genetic algorithms (GAs) comprise a subset of evolutionary algorithms focusing on the application of selection, mutation, and recombination to a population of competing problem solutions [19]. Obviously, GAs are prime candidates for random probabilistic search algorithms within the context of feature selection.

There are three reasons for subset feature selection in conjunction with classification. First, irrelevant, non informative features may result in a classifier which is not robust. This is due to the fact that classification error does not satisfy monotonicity. Second, a large number of features implies also a large number of observations to properly design a classifier. Finally, by eliminating irrelevant features, classification time and time for data collection can be reduced. Frequently, before proceeding to speech emotion recognition subset feature selection is performed [9], [11], [20]. GAs have also been employed for feature generation in speech emotion recognition [10].

In this paper, we employ adaptive GAs to further reduce the prediction error for speech emotion recognition reported in [9], [12]. Adaptive GAs change the probabilities of crossover and mutation during generations based on the diversity of population [21], [22]. They search for the worst performing features with respect to the probability of correct classification achieved by the Bayes classifier in a first stage. These features are subsequently excluded from sequential floating feature selection employing the probability of correct classification achieved by the Bayes classifier as criterion. In a second stage, adaptive GAs are employed to search for the worst performing utterances with respect to the same criterion. The sequential application of both stages is demonstrated to improve speech emotion recognition on the Danish Emotional Speech database [23].

In GA literature, a binary string codes the chromosomes (i.e. features or utterances here). In this binary coding, 1 implies that the feature/utterance is active and 0 implies the opposite. In this paper, another coding is employed that codes the location of active features/utterances. That is, integer values are used, which index the location of the worst features/utterances that should be excluded from further consideration. Definitely, the number of the worst features are much less than the best ones. Therefore, instead of having a lengthy binary stream, we have a very short integer stream that can easily be interpreted.

The outline of the paper is as follows. Section II briefly describes GAs. The proposed method is outlined in Section III. Experimental results are demonstrated in Section IV and conclusions are drawn in Section V.

II. Genetic algorithms

In this section, the operators of the adaptive GAs are briefly described. In the following, genes refer to integer-valued elements of chromosomes (i.e. strings of genes encoding individuals). Instead of searching for the best genes, we are interested in seeking the worst ones. An integer matrix $\mathbf{P}$ of dimensions $N_p \times N_w$ is defined whose element P_{ij} codes the feature index of the jth worst gene of the ith individual (chromosome). P_{ij} admits an integer value in the range $[1, N]$, where N is the number of features in the first stage or the number of utterances in the second stage. N_p and N_w are predefined.

Let us define the *population diversity* as the normalized square root of the sum of differences between any two distinct rows of the population matrix, i.e.

$$D = \frac{2}{N_p\,(N_p-1)} \sum_{i=1}^{N_p-1} \sum_{j=i+1}^{N_p} \sqrt{(\mathbf{p}_i - \mathbf{p}_j)(\mathbf{p}_i - \mathbf{p}_j)^T} \quad (1)$$

where $\mathbf{p}_i$ is a row vector that represents the ith chromosome. To avoid misunderstandings, inner products are employed in (1).

In general, the initial population is generated randomly. To do so, a uniform random number generator fills in $\mathbf{P}$ with integers in the desired range. P_{ij} are checked for uniqueness inside each chromosome. Typical values of N_p could be 50, 100, 200. The default value of N_p is 100. Experiments with $N_p = 50$, 200 did not yield any significant difference. Let N_{iter} denote the number of iterations. N_{iter} typically admits values 50, 100, and 200. However, the larger N_{iter} is, the higher the chance to find the optimal value is, but at the expense of more computational time. If adaptive GAs are not employed, it is more probable to get a null diversity, when N_{iter} is large. This is due to, it is most probable to have the dominant chromosome to fill all rows of $\mathbf{P}$ after some iterations.

The selection strategy is cross generational and differs from traditional selection. In traditional selection, the fittest genes have more chance to survive. However, in cross generational selection, additional random chromosomes are appended in $\mathbf{P}$. The number of new chromosomes could be N_p or a fraction of N_p. In our experiments another N_p chromosomes are randomly generated, and the N_p out of the $2N_p$ worst chromosomes with respect to the fitness criterion are given a chance to survive in the next generations. The evaluation procedure for the fitness of population is the repeated ψ-fold cross validated prediction error [24].

We apply a simple multi-point crossover operator [25]. The number of points and also their positions are determined randomly for any pair of candidate parents for crossover. The probability of the crossover is determined by the status of population diversity. We call it adaptive crossover.

A single-point binary mutation at point k (i.e., the kth bit is toggled) is performed [25] (integer-binary-integer conversion is considered). The probability of mutation is also determined by the status of population diversity. We call it adaptive mutation. The choice of the crossover rate is not critical compared to the mutation probability. A large value of mutation probability will not allow the search to focus on the better regions and the GA will perform a random search. However, a small value will not allow the search to escape from local minima. An optimal choice of the probability of mutation will allow GA to explore the more promising regions, while avoiding getting trapped into local minima.

III. The proposed method

The outline of the proposed method is as follows.

1) Generate the matrix $\mathbf{P}$ of size $N_p \times N_w$, for $N_p = 100$. For feature trimming, N_w may vary from 1 to N_f, where N_f denotes the number of the features. In the experiments reported in Section IV, $N_w = 1$ for feature trimming, while $N_w = 3$ for utterance trimming.
2) Assure that there are no repetitions inside each row as well as between rows.
3) Evaluate the fitness of the initial population.
4) Repeat the following steps, until all population chromosomes have been examined (i.e. the maximum generation is reached). Also control the diversity of the population. If it reaches 0, then quit the loop.
5) Start a loop. Generate another N_p chromosomes in the selection stage and attach them to the previous population. Then, evaluate their fitness. Select the worst N_p chromosomes.
6) Calculate the diversity of the population and select probabilities of the crossover and mutation operators. If the diversity is more than a threshold, then assign

a minimum value to both probabilities (e.g. 0.5 to crossover and 0.01 to mutation). Let $T_{\min}$ and $T_{\max}$ define two thresholds. If $D < T_{\min}$, then increase the probabilities of crossover and mutation. If $D > T_{\max}$, then decrease them. Otherwise, do not modify them. In our experiments, $T_{\min}$ and $T_{\max}$ were defined as 0.1 and 0.95, respectively.

7) Apply crossover to randomly selected parents pairs.
8) Apply mutation to randomly selected parents.
9) Repeat the loop (i.e., jump to step 4).
10) After the GA has converged, then remove the worst features/utterances from the dataset.
11) Evaluate the remaining features using the SFFS algorithm with criterion the probability of correct classification achieved by the Bayes classifier, when the features are modelled by a multivariate Gaussian probability density function. If some utterances are excluded, then SFFS is applied on the retained utterances and the probability of correct classification of the Bayes classifier is estimated by the repeated ψ-fold cross validation.

IV. Experimental results

Emotional speech data from Danish Emotion Speech (DES) [23] are employed. The recordings correspond to speech expressed by 2 male and 2 female actors under 5 emotional states such as anger, happiness, neutral, sadness, and surprise. The speech data consist of 2 words, 9 sentences, and 2 paragraphs. Overall, 1160 utterances have been used. Gender information has not been exploited. The basis for our experiments is the results reported in [9], [12].

The statistical features employed in this study are grouped in several classes as is explained in the sequel.

Formants features: The set of formants features indexed by 1-15 is comprised by the statistical properties of the 4 formant frequency contours. 1-4: Mean value of the first, second, third, and fourth formant. 5-7: Maximum value of the first, second and third formant. 8-11: Minimum value of the first, second, third, and fourth formant. 12-15. Variance of the first, second, third, and fourth formant.

Pitch features: The pitch features indexed by 16-39 are statistics of the pitch frequency contour. 16-20: Maximum, minimum, mean, median, interquartile range of pitch values. 21: Pitch existence in the utterance expressed in percentage (0-100%). 22-24: Maximum, mean, median of durations for the plateaux at maxima. 25-27: Mean, median, interquartile range of the pitch values within the plateaux at maxima. 28-30: Maximum, mean, median range of durations of the rising slopes of pitch contours. 31-33: Mean, median, interquartile range of the pitch values within the rising slopes of pitch contours. 34-36: Maximum, mean, median range of durations of the falling slopes of pitch contours. 37-39: Mean, median, interquartile range of the pitch values within the falling slopes of pitch contours.

Energy (intensity) features: The energy features indexed by 40-63 are statistics of the energy contour. 40-44: Maximum, minimum, mean, median, interquartile range of energy values. 45-48: Maximum, mean, interquartile range, upper limit (90%) of duration for the plateaux at maxima. 49-51: Mean, median, interquartile range of the energy values within the plateaux at maxima. 52-54: Maximum, mean, median range of durations of the rising slopes of energy contours. 55-57: Mean, median, interquartile range of the energy values within the rising slopes of energy contours 58-60: Maximum, mean, median range of durations of the falling slopes of energy contours. 61-63: Mean, median, interquartile range of the energy values within the falling slopes of energy contours.

Spectral features: The spectral features indexed by 64-90 is the energy content of certain frequency bands divided to the length of the utterance. 64-71: Energy below 250, 600, 1000, 1500, 2100, 2800, 3500, 3950 Hz. 72-78: Energy in the frequency bands 250 - 600, 600 - 1000, 1000 - 1500, 1500 - 2100, 2100 - 2800, 2800 - 3500, 3500 - 3950 Hz. 79-83: Energy in the frequency bands 250 - 1000, 600 - 1500, 1000 - 2100, 1500 - 2800, 2800 - 3950 Hz. 84-88: Energy in the frequency bands 250 - 1500, 600 - 2100, 1000 - 2800, 1500 - 3500, 2100 - 3950 Hz. 89-90: Energy ratio between the frequency bands (3950 - 2100) and (2100 - 0) and between the frequency bands (2100 - 1000) and (1000 - 0).

We have run classical and adaptive GAs to investigate the possibility of improving speech emotion recognition by excluding the worst performing features, before applying SFFS. Among them, the results for adaptive were found to be promising.

Fig. 1 illustrates how well the adaptive GA controls the diversity of the population along generations in one of the experiments within $N_{\text{iter}} = 50$ iterations, compared with classical GAs.

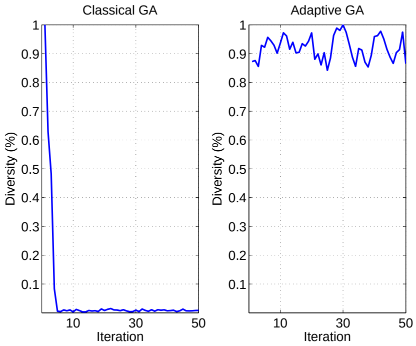

Fig. 1. Normalized diversity through generations for classical (left) and adaptive (right) GAs.

Table I presents the confusion matrix from subjective human evaluation [23]. The utterances are correctly identified with an average rate of 67%. “Surprise” and “Happiness” are often confused as well as “Neutral” and “Sadness”. Table II shows the confusion matrix for speech emotion recognition using the Bayes classifier with SFFS [9] for 30 cross-validation repetitions and when 30% of the utterances are used for testing. Table III demonstrates the confusion matrix for the results provided by the proposed method, when utterances 1132-1135 and feature 2 (i.e. the mean value of the second formant) have been excluded. The cross-validation repetitions are limited to 30 and 30% of the available utterances are used for testing. It is seen that the probability of correct decisions for anger, neutral, sadness, and surprise is slightly increased. Therefore, the first results reported are promising, because the algorithm is able to detect the outliers from features and utterances.

TABLE I

CONFUSION MATRIX FROM SUBJECTIVE HUMAN EVALUATION [23].

Stimuli	Correctly classified responses (%)				
	Anger	Happ.	Neutral	Sadness	Surprise
Anger	**75.1**	4.5	10.2	1.7	8.5
Happiness	3.8	**56.4**	8.3	1.7	29.8
Neutral	4.8	0.1	**60.8**	31.7	2.6
Sadness	0.3	0.1	12.6	**85.2**	1.8
Surprise	1.3	28.7	10.0	1.0	**59.1**
Total rate	67.3%				

TABLE II

CONFUSION MATRIX FOR THE BAYES CLASSIFIER WITH SFFS WHEN CROSS-VALIDATION REPETITIONS ARE LIMITED TO 30 AND 30% OF THE UTTERANCES ARE USED FOR TESTING [9].

Stimuli	Correctly classified responses (%)				
	Anger	Happ.	Neutral	Sadness	Surprise
Anger	**41.65**	19.28	16.20	11.05	11.82
Happiness	19.24	**32.19**	18.29	11.04	19.24
Neutral	7.28	5.88	**47.63**	31.09	8.12
Sadness	2.03	1.52	18.32	**72.79**	5.34
Surprise	22.28	14.40	7.33	14.94	**41.05**
Total rate	47.06%				

TABLE III

CONFUSION MATRIX WHEN THE ADAPTIVE GA REMOVES THE MEAN VALUE OF THE SECOND FORMANT AND UTTERANCES 1132-1135 FROM SUBSEQUENT CLASSIFICATION.

Stimuli	Correctly classified responses (%)				
	Anger	Happ.	Neutral	Sadness	Surprise
Anger	**44.40**	17.43	14.39	10.09	13.69
Happiness	18.86	**37.73**	11.79	12.34	19.28
Neutral	4.79	5.75	**47.81**	36.17	5.48
Sadness	2.40	2.40	19.63	**71.57**	4.00
Surprise	14.62	18.90	10.76	12.69	**43.03**
Total rate	48.91%				

V. CONCLUSION AND FUTURE WORK

We have applied an adaptive GA scheme to further optimize the results of feature subset selection algorithms. Adaptive GAs yield an improvement in correct classification rate. Our future work would employ more efficient pre-processing tasks for extracting features, fuse some new features provided by morphological filtering, and analyze how features affect the rate of classification of a given emotion.

VI. ACKNOWLEDGEMENT

This work was carried out during the tenure of a MUSCLE Internal fellowship.

REFERENCES

[1] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[2] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, pp. 227–256, 2003.

[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Comunnication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[4] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, Eds., pp. 301–320. J. Wiley, N.Y., 1999.

[5] J. M. Susskind, G. Littlewort, M. S. Bartlett, J. Movellan, and A. K. Anderson, "Human and computer recognition of facial expressions of emotion," *Neuropsychologia*, vol. 45, pp. 152–162, 2007.

[6] J. S. Morris, S. K. Scott, and R. J. Dolan, "Saying it with feeling: Neural responses to emotional vocalization," *Neuropsychologia*, vol. 37, pp. 1155–1163, 1999.

[7] D. Wildgruber, A. Riecker, I. Hertrich, M. Erb, W. Grodd, T. Ethofer, and H. Ackermann, "Identification of emotional intonation evaluated by fMRI," *NeuroImage*, vol. 24, pp. 1233–1241, 2005.

[8] R. K. Moore, "Spoken language processing: Piecing together the puzzle," *Speech Communication*, vol. 49, pp. 418–435, 2007.

[9] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proc. XII European Signal Processing Conf.* Vienna, Austria, September 2004, vol. 1, pp. 341–344.

[10] B. Schuller, D. Arsić, F. Wallhoff, M. Land, and G. Rigoll, "Bioanalog acoustic emotion recognition by genetic feature generation based on low-level-descriptors," in *Proc. Int. Conf. Computer as Tool (EUROCON)*, 2005, pp. 1292–1295.

[11] B. Schuller, R. Jiménez, G. Rigoll, and M. Lang, "Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2005, vol. 1, pp. 325–328.

[12] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," in *Proc. XIV European Signal Processing Conf.*, 2006.

[13] Z. Hammal, B. Bozkurt, L. Couvreur, U. Unay, A. Caplier, and T. Dutoit, "Passive versus active: Vocal classification system," in *Proc. XIII European Signal Processing Conf.* Antalya, Turkey, September 2005.

[14] J. Kim, E. André, M. Rehm, T. Vogt, and J. Wagner, "Integrating information from speech and physiological signals to achieve emotional sensitivity," in *Proc. 9th. European Conf. Speech Communication and Technology*, 2005.

[15] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large scale feature selection," in *Pattern Recognition in Practice IV*, J. E. Moody, S. J. Hanson, and R. L. Lippmann, Eds., 1994, pp. 403–413.

[16] A. K. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal., Machine Intell.*, vol. 19, no. 2, pp. 153–158, 1997.

[17] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *IEEE Trans. Pattern Anal., Machine Intell.*, vol. 26, no. 7, pp. 900–912, July 2004.

[18] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, 1994.

[19] D. Goldberg, Ed., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley, Reading, MA, 1989.

[20] J. Wagner, J. Kim, and E. André, "From physiological signals to emotion: Implementing & comparing selected methods for feature extraction & classification," in *Proc. IEEE Int. Conf. Multimedia & Expo*, 2005.

[21] B. A. Julstrom, "Adaptive operator probabilities in a genetic algorithm that applies three operators," in *Proc. ACM Symp. Applied Computing*. San Jose, CA, 1997, pp. 233–238.

[22] P. J. Angeline, "Adaptive and self-adaptive evolutionary computations," in *Computational Intelligence: A Dynamic Systems Perspective*, M. Palaniswami and Y. Attikiouzel, Eds., pp. 152–163. IEEE Press, 1995.

[23] I. S. Engberg and A. V. Hansen, "Documentation of the danish emotional speech database (des)," 1996.

[24] P. Burman, "A comparative study of ordinary cross-validation, ψ-fold cross-validation and the repeated learning-testing methods," *Biometrika*, vol. 76, no. 3, pp. 503–514, 1989.

[25] J. Joines, "The genetic algorithm optimization toolbox for MATLAB".