INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Linear Discriminant Differential Evolution for Feature Selection in Emotional Speech Recognition

Soumaya Gharsellaoui[1]*, Sid Ahmed Selouani*[1]*, Mohammed Sidi Yakoub*[1]

[1]Université de Moncton, 128 Boul. J. - D. Gauthier, Shippagan (NB), E8S 1P6, Canada

soumaya.gars@gmail.com, sidi-ahmed.selouani@umoncton.ca, mohammed.sidi.yakoub@umoncton.ca

Abstract

In this paper, an evolutionary algorithm is used to select an optimal set of acoustic features for emotional speech recognition. A new algorithm that combines differential evolution (DE) optimization and linear discriminant analysis (LDA) is proposed to design an effective feature selection and classification model. An original acoustic feature framework based on auditory modeling is also presented. The auditory-based features are provided as inputs to the DE-LDA based emotional speech recognition system. To evaluate the effectiveness of the DE-LDA approach, a subset of the Emotion Prosody Speech and Transcript corpus covering five emotional states (happiness, anger, panic, sadness, and interest) is used throughout the experiments. The results show that the proposed DE-LDA model performs significantly better than the baseline systems. It achieves a classification rate of 91.6% using only 50 input parameters that are optimally selected from 128 original acoustic features.

Index Terms: auditory modeling, differential evolution, emotion, linear discriminant analysis

1. Introduction

Emotional speech recognition (ESR) applications are characterized by the higher dimensionality of the features, which are not necessarily all relevant. The problem of selecting an optimal subset from the original features can be resolved either by feature selection or transformation to a low-dimensional space. Many efforts have been made to improve the performance of emotional state recognition from speech. Most efforts have been concentrated on the feature selection side of the solution, as investigated in [1] and [2]. For instance, according to the work published in [3], discrimination among seven emotions needs only 26 relevant features of 133 original speech features composed of formants, mel-frequency coefficients, and prosodic parameters.

Many approaches treat feature selection as an optimization problem. Therefore, the population-based search strategy approaches such as ant colony, artificial bee colony, particle swarm optimization, differential evolution (DE), and all other algorithms belonging to the evolutionary-based approach can be used to successfully find the optimal subset of representation features in various domains. In [4], an exhaustive study was carried out on feature selection using particle swarm optimization for maintaining higher classification rates with only the relevant features in various applications such as image processing or bioinformatics. In [5], the artificial bee colony algorithm was used to perform a feature selection on various benchmark datasets related to health and imagery domains.

Among the evolutionary-based approaches, DE has numerous advantages and has been proven effective for dealing with the problem of feature selection and/or classification. It is characterized by a simple structure and a high performance that allows it to deal with high-dimensional problems. Using DE requires the definition of an objective function that is optimized throughout the generations and thus covers more search space than conventional approaches [6]. DE is designed to find the globally optimum solution using a stochastic parallel direct search method to explore the solution space, but can also perform feature selection, as shown in [7]. In [8], the approach was developed to reduce the computational costs and memory requirement in statistical learning from brain–computer interfaces by performing a DE-based feature selection. Mallipeddi and Lee in [9] found a way to reduce the huge number of features using PCA combined with the DE algorithm to generate a relevant set of features for face recognition. In [10], DE was assessed as a feature selection and classifier on a wide range of datasets provided in [11]. The hybridization of evolutionary-based methods with machine learning and classification techniques proven to be effective in various applications such as in [12], where good performance was achieved in the classification of electroencephalography motor imagery despite the high dimensionality of the data.

In the light of related work, our contribution consists of proposing a new framework designed to improve emotion recognition from speech, performing an optimal feature selection after incorporating a new type of input feature based on the human auditory system. This is performed to increase the effectiveness of the automatic disambiguation of human emotional states from speech flow. The proposed feature selection framework, based on DE and linear discriminant analysis (LDA), allows us to formalize the acoustic feature selection as an optimization problem. Hence, the goal of the new DE-LDA algorithm is to provide the optimal subset of features that achieves the highest ESR score.

The remainder of the paper is organized as follows. Section 2 introduces the ear model, which provides the auditory-based features used as input features. Section 3 presents the optimal feature selection approach and the proposed DE-LDA model. The experiments and results of the proposed ESR method are presented in Section 4. The paper is concluded in Section 5.

2. Auditory modeling for ESR

The auditory model used in our work simulates the coding of speech signals performed by the three parts of the human ear that enable sounds to be decoded by the brain: the external, middle, and inner parts. To extract relevant phonetic features, Caelen's auditory model is used [13]. The external and middle parts of the ear are represented by a band-pass filter that can be adjusted to take into account the various adaptive movements of the ossicles. The expression of these two parts is represented by the following equation:

$$S^{'}(k) = S(k) - S(k-1) + \alpha_1 S^{'}(k-1) + \alpha_2 S^{'}(k-2), \quad (1)$$

http://dx.doi.org/10.21437/Interspeech.2019-1218

where $S(k)$ is the speech signal, $S^{'}(k)$ is the filtered signal, $k = 1, ..., K$ is the time index, and K is the number of frame samples. The coefficients α_1 and α_2 depend on the sampling frequency F_s, the central frequency of the filter, and its Q-factor. The values of 1,500 Hz as a central frequency and 1.5 as a Q-factor were found to be optimal [13].

One important feature used in our ESR model is the log-energy of the signal received after passing the external and middle parts of the ear. This parameter, W_{om}, can be calculated as follows:

$$W_{om} = 20 \log \sum_{k=1}^{K} \left| S^{'}(k) \right|. \quad (2)$$

The most important part of the inner ear is the basilar membrane (BM), which is located in the cochlea. Different areas along the BM have a variable stiffness and are responsive to sounds with distinct spectral content. Each position on the BM has a distinctive vibration frequency that occurs for a certain input sound. This is simulated in the ear model by a filter bank structure. More filters lead to a more precise model. Before being processed by the BM, the model mimics the effects of the outer and middle ear using band-pass filters. The final part of the model deals with the electro-mechanical transduction of hair-cells and afferent fibers as well as the encoding at the level of the synaptic endings. In the present study, the BM is designed using 24 overlapping digital filters. Our model simulates the ear mechanism and functioning of the BM. The output y_i of each filter is simulated by the following equation:

$$y_i(k) = \beta_{1,i} y_i(k-1) - \beta_{2,i} y_i(k-2) + G_i[S^{'}(k) - S^{'}(k-2)], \quad (3)$$

where $y_i(k)$ represents the BM response to the mid-external wave $S^{'}(k)$; parameters G_i, $\beta_{1,i}$, and $\beta_{2,i}$ represent the gain and two coefficients, respectively, of filter i. The auditory-based parameter is calculated using a combination of the log-linear energy outputs of the channels. The log-energy of each channel output is defined as follows:

$$W_i^t(k) = 20 \log \sum_{k=1}^{k} \left| y_i^{'}(k) \right|, \quad (4)$$

where T refers to the frame index and i refers to the channels, where $i = 1, ..., 24$. To reduce the energy variations, a smoothing function is applied:

$$W_i(T) = c_0 W_i(T-1) + c_1 W_i(T), \quad (5)$$

where $W_i(T)$ is the smoothed energy and c_0 and c_1 are coefficients for averaging $W_i(T-1)$ and $W_i(T)$ such that their sum is unity. In our case, c_0 and c_1 are equal to 0.5. More details about our quantitative auditory model can be found in [14]. The following eight auditory-based features are extracted from the speech signal to perform ESR: grave/acute (G/A), open/closed (O/C), diffuse/compact (D/C), flat/sharp (F/S), mellow/strident (M/S), continuous/discontinuous (C/D), tense/lax (T/L), and mid-external energy (W_{om}). Table 1 describes the eight auditory cues and the formula used for their calculation.

Table 1: *Eight ear model-based parameters used to ESR.*

Model description
(G/A) : measures the difference of energy between low frequencies (50-400 Hz) and high frequencies (3800-6000 Hz): $\left(W_1 + ... + W_5\right) - \left(W_{20} + ... + W_{24}\right)$
(O/C) : a phoneme is considered closed if the energy of low frequencies (230-350 Hz) is greater than that of the middle frequencies (600-800 Hz). Hence, the O/C cue is calculated by: $W_8 + W_9 - W_3 - W_4$
(D/C) : compactness reflects the prominence of the central formant region (800-1050 Hz) compared with the surrounding regions (300-700 Hz) and (1450-2550 Hz): $W_{10} + W_{11} - \left(W_4 + ... + W_8 + W_{13} + ... + W_{17}\right)/5$
(F/S) : a phoneme is considered sharp if the energy in (2200-3300 Hz) is more important than the energy in (1900-2900 Hz): $W_{17} + W_{18} + W_{19} - W_{11} - W_{12} - W_{13}$
(M/C) : strident phonemes are characterized by a presence of noise because of a turbulence at their articulation point which leads to more energy in (3800-5300 Hz) than in (1900-2900 Hz): $W_{21} + W_{22} + W_{23} - W_{16} - W_{17} - W_{18}$
(C/D) : quantifies the variation of the spectrum magnitude by comparing the energy of current and preceding frames. $\sum_{i=1}^{N_c} \left\| W_i(T) - Wa(T) - W_i(T-1) + Wa(T-1) \right\|$ $W_i(T)$ is the energy of channel *i* $Wa(T)$ is the energy average over all channels of current frame *T*.
(T/L) : measures the difference of energy between middle frequencies (900-2000 Hz) and relative high frequencies (2650-5000 Hz): $\left(W_{11} + ... + W_{16}\right) + \left(W_{18} + ... + W_{23}\right)$
(W_{om}) : measures the log-energy of the signal received from the external and middle ear : $W_{om} = 20 \log \sum_{k=1}^{K} \left\| S^{'}(k) \right\|$

3. Evolutionary-based feature selection and classification

The proposed approach treats the feature selection problem of emotion recognition from speech as an optimization problem. The novelty of this approach consists of combining DE with LDA to perform optimal feature selection for ESR. The following subsections give a brief background for both LDA and DE. The proposed approach is also presented.

3.1. Linear Discriminant Analysis

LDA is a multivariate statistical method that is used to build a model of a given dependent categorical variable based on its relationship with one or more predictors [15], [16]. The discriminant analysis builds optimal linear combinations of the variables (features) for each class. These combinations are known as discriminant classification functions and can be used to determine to which class each observation should be assigned [17]. The number of classification functions is equal to the number of groups. Each classification function allows us to calculate the classification scores for each observation for each group by applying the following formula:

$$Sc_i = C_i + w_{k,1} X_1 + ... + w_{p,n} X_n. \quad (6)$$

In this formula, i represents the class, n denotes the variables, C_i is a constant assigned to the i^{th} class, $w_{i,j}$ is the coefficient assigned to the j^{th} variable in the computation of the classification score for the i^{th} class, and X is the observed value for the respective case for the j^{th} variable. Score Sc_i is the resultant classification score. Once the classification functions have been derived by LDA, they can be used to directly compute the classification scores for new observations.

3.2. Differential Evolution

DE is an optimization method originally developed by Storn and Price [6] for nondifferentiable and multimodal optimization problems, providing a near-optimal solution for the objective function. DE uses three main operators in each generation: mutation, crossover, and selection. It uses mutation and crossover to calculate the trial vector of each target vector. The selection operator chooses the next generation's population vectors from the trial and target vectors that fit best. The implementation of DE requires the determination of a set parameters such as the crossover parameter CR, scaling factor F, population size, and number of generations. According to the literature, F is always between 0.4 and 1, and in our work, it is set to 0.6. Moreover, CR is set to 0.9. The population size is relative to the problem dimension d and is usually set between $5\times d$ and $10\times d$. In our case, the population size is 50. The number of generations was set to 100.

3.2.1. Mutation

Mutation is the key operator that allows DE to generate new parameter vectors by adding the weighted difference between two population vectors to a third type of vector within each generation. The following equation governs the creation of the mutant vector:

$$V_g = X_g^{r_1} + F(X_g^{r_2} - X_g^{r_3}), \tag{7}$$

where F is the differential weight, which is a scale factor that controls the rate at which the population evolves. Moreover, r_1, r_2, and r_3 are random indexes and V_g is the mutant vector in generation g, while $X_g^{r_1}$, $X_g^{r_2}$, and $X_g^{r_3}$ are the agents (candidates) selected randomly by the random indexes.

3.2.2. Crossover

DE also uses a discrete recombination, known as crossover, to develop a trial vector from the elements of the mutant and target vectors. The new vector is the outcome of a binary crossover of an agent with the mutant vector. Hence, the trial vector (outcome) is represented by the following equation:

$$U_g^i = \begin{cases} V_g^i & if \quad rand(0,1) \leq CR \\ X_g^i & otherwise, \end{cases} \tag{8}$$

where U_g^i is the trial vector generated from the mutant vector V_g^i and target vector X_g^i; CR is a crossover probability used to control the fraction of the mutant parameter value.

3.2.3. Selection

In this a step, the newly generated vector and the target vector are compared. The vector with the best fitness is chosen to replace the target vector. In our case the objective function is the classification rate. Therefore, agents with the highest classification rates are kept for the next step. The selection process is formulated by the following equation:

$$X_g^i = \begin{cases} U_g^i & if \quad f(U_g^i) \leq f(X_g^i) \\ X_g^i & otherwise. \end{cases} \tag{9}$$

3.3. DE-LDA algorithm

The proposed method combines the LDA and DE techniques described above to design a feature selection model. The LDA can be used two ways depending on the manner in which the features in an entry are considered. In the first method, the LDA performs the classification step by simultaneously considering all features as one block. Then, a set of linear classification functions is provided at the output. The set of functions is composed of a linear combination of the different input features. Thus, a weighting coefficient is attributed to each input feature. The role of the DE is to find the optimal set of weighting coefficients, using the LDA classification rate as an objective function. The algorithm used to perform DE-LDA-based optimal feature selection starts with an initialization of different variables such as the population size, subset dimension, and number of generations. The first population is randomly generated. Then, the three DE operations are evaluated by Equations (6), (7), and (8). Finally, the set of optimal features with the highest classification rate is retained for the testing stage.

The second method is a stepwise method in which LDA plays the role of the selection method and classifier simultaneously. This approach consists of the following steps: i) selecting a subset from the original set of features; ii) returning the classification rate with a set of weighting coefficients for the corresponding subset of features; and iii) repeating these steps until all training data has been processed. DE-LDA can be applied using the stepwise method. However, we have noticed that the classification rate obtained by the first method is better than it is with the second method. This is the reason why we use the first LDA method.

4. Experiments and results

This section presents the experimental protocol as well as the data used throughout the experiments. The validation and assessment of the proposed DE-LDA technique using the original set of auditory-based features was carried out by comparing its performance against the widely used baseline approaches PCA, LDA, and ANOVA. Our goal is to demonstrate the effectiveness of the proposed approach on a relevant and internationally referenced corpus compared with well-known feature selection techniques.

4.1. Data

The Emotional Prosody Speech and Transcripts corpus is one of the most important emotion corpora provided by the Linguistic Data Consortium [18]. It is used throughout this study. In this corpus, the speakers were advised to avoid exaggerated expressions, as it would have affected the production of these kinds of contextual variations. A total of seven speakers (3 males and 4 females) were chosen, and each speaker acted out the different emotional states. In this work, two sets of experiments were carried out. In the first set, five emotional states were targeted: Happiness, Hot Anger, Panic, Sadness, and Interest. Each speaker pronounced eight sentences for each emotion. The total number of sentences used in the experiments is 280. The training set is composed of 188 sentences and the test set is composed of 92 sentences. In the second set of experiments, the same number of speakers and sentences were used for the emotions of Cold Anger, Despair, Pride, Shame, and Disgust. By using two different sets of emotions, our intention is to assess the proposed approach on more data and investigate its robustness for different emotions. The raw number of features of the original input vector (before reduction) for the two experiences is 128.

The input vector is composed of eight auditory-based cues that are represented by 16 Gaussians. Each utterance is transformed to be compatible with the LDA input, which is com-

posed of 128 parameters (8×16) that constitute the original vector.

Algorithm 1 DE-LDA feature selection

1. Initialization
   - **Set** the number and dimension of agents in a population
   - **Set** the feature boundaries
   - **Generate** the initial population
   - **Set** the number of generations Gen_{max}
   - **Set** the mutation factor F and the crossover rate CR
   - **Make** the feature coefficient result of LDA training

   For Gen_{max} generation **Do**
2. **Calculate** the mutant vectors using the mutation operator
3. **if** any element of the mutant vectors is below/above of the minimum/maximum feature accepted boundary **then** it is replaced by the minimum/maximum feature boundary
4. **Generate** the trial vectors by using the crossover operator
5. **Calculate** the objective function (classification accuracy) of each trial vector using the LDA classifier
6. **Perform** the selection of the population vectors to be retained for the next generation
   End For
7. **Terminate** the process and provide the set of optimal features

4.2. Optimization of the feature space dimension

The first step of our experimental protocol consists of finding the best feature dimension for each of the two baseline methods. The first baseline method is PCA, which performs an optimal projection of the original features onto a new feature space with a lower dimension. PCA reduces the feature dimensionality by removing the redundant information and constructs the new space by considering, as new representation axes, only the eigenvectors that keep the largest percentage of the original data variance. The second method is ANOVA, which is applied to the original feature vector and performs a statistical test to evaluate the significance of each feature. After performing ANOVA, features with a p-value of 0.05 or less are considered significant and then kept for the classification step. Before doing the classification tasks that involve PCA and ANOVA, it is important to determine the number of optimal features (optimal dimension) to consider in the next step of ESR. For PCA, the first twelve components that have the highest eigenvalues are kept as the optimal set of features. This optimal number of components was obtained after performing cross-validation experiments that showed no improvement in the performance when more than twelve eigenvectors were used. The ANOVA cross-validation tests allow us to return 83 of the 128 features as the most relevant features.

Table 2: *Classification accuracy rates and feature space dimensions of the PCA-LDA, ANOVA-LDA, LDA without feature selection, and DE-LDA systems for the emotion states: Happiness, Hot Anger, Panic, Sadness, and Interest.*

Classification systems	**#features**	**Accuracy(%)**
PCA-LDA	12	57.5
ANOVA-LDA	83	72.5
LDA without feature selection	128	85.7
DE-LDA	50	91.6

Table 3: *Classification accuracy rates and feature space dimensions for the PCA-LDA, ANOVA-LDA, LDA without feature selection, and DE-LDA systems for the emotion states: Cold Anger, Despair, Pride, Shame, and Disgust.*

Classification systems	**#features**	**Accuracy(%)**
PCA-LDA	6	63.2
ANOVA-LDA	23	67.0
LDA without feature selection	128	80.7
DE-LDA	14	89.4

In addition to the statistical approaches (PCA and ANOVA), a cross-validation experiment using a reference system based only on DE was performed to determine the optimal number of features (dimensions) for DE. Of the 128 features, 50 were shown to be effective and were retained for DE-based ESR.

4.3. ESR results

The second step of the experimental protocol was to perform ESR by applying LDA as a classification technique to the baseline techniques and DE. The two set of features obtained by PCA and ANOVA were passed to LDA to evaluate their discrimination abilities. The emotion recognition system is based on the LDA classifier that is trained many times to improve the classification rate. When LDA is applied without any combination with another feature selection method, it achieves an 85.7% correct recognition rate when using all the original features. The proposed DE-LDA algorithm achieves the best emotion classification rate of 91.6%. This rate decreases to 72.5% and 57.5% when LDA is combined with ANOVA and PCA, respectively. Tables 2 and 3 summarize the results obtained by each method as well as their respective optimal feature space dimension.

5. Conclusion

In this paper, a new algorithm that combines DE optimization and LDA was proposed to perform effective human emotion classification from speech signals. An original and relevant acoustic feature framework based on auditory modeling was also presented. The auditory-based features were used as inputs by the proposed DE-LDA system, leading to a very satisfactory emotion classification rate of 91.6%. We note that the highest score was obtained with only 50 of the 128 original features. In the context of these very promising results, we are currently working on hybridizing various feature selection methods with other evolutionary-based algorithms to advance the research in the field of ESR.

6. References

[1] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, N. Amir,

Whodunnit-searching for the most important feature types signalling emotion-related user states in speech, Computer Speech and Language (2011) 4–28.

[2] M. Tahon, L. Devillers, Towards a small set of robust acoustic features for emotion recognition: Challenges, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (1) (2016) 16–28.

[3] T. Iliou, C. N. Anagnostopoulos, Classification on speech emotion recognition–a comparative study, International Journal On Advances in Life Sciences 2 (2010) 18–28.

[4] Y.Zhang, D. Gong, X.Y.Sun, Y. Guo, A PSO-based multi-objective multi-label feature selection method in classification, Scientific Reports 7 (1) (2017) 376–388.

[5] E.Hancer, B.Xue, D. Karaboga, M. Zhang, A binary ABC algorithm based on advanced similarity scheme for feature selection, IEEE Transactions on Evolutionary Computation 36 (2015) 334–348.

[6] R. Storn, K. Price, Differential evolution - A simple and efficient adaptive scheme for global optimization over continuous spaces, Journal of Global Optimization 1 (11) (1997) 341–359.

[7] B.Xue, M.Zhang, W. Browne, X.Yao, A survey on evolutionary computation approaches to feature selection, IEEE Transactions on Evolutionary Computation 20 (4) (2015) 606–626.

[8] E. Hancer, B. Xue, M. Zhang, Differential evolution for filter feature selection based on information theory and feature ranking, Knowledge-Based Systems 140 (C) (2018) 103–119.

[9] R. Mallipeddi, M. Lee, Ensemble based face recognition using discriminant PCA features, WCCI 2012 IEEE World Congress on Computational Intelligence 1 (1) (2012) 1–7.

[10] J. Wang, B. Xue, X. Gao, M. Zhang, A differential evolution approach to feature selection and instance selection, Artificial Intelligence (2016) 588–602.

[11] M. Lichman, UCI machine learning repository (2013).
URL http://archive.ics.uci.edu/ml

[12] M.Z.Baig, N. Aslam, H. P. H.Shum, L.Zhang, Differential evolution algorithm as a tool for optimal feature subset selection in motor imagery EEG, Expert Systems with Applications 90 (2017) 184–195.

[13] S. A. Selouani, D. O'Shaughnessy, J. Caelen, Incorporating phonetic knowledge into an evolutionary subspace approach for robust speech, International Journal of Computers and Applications (2007) 143–154.

[14] S. A. Selouani, Y. Alotaibi, W. Cichocki, S. Gharsellaoui, K. Kadi, Native and non-native class discrimination using speech rhythm- and auditory-based cues, Computer Speech and Language 31 (2015) 28–48.

[15] A. Khan, H. Farooq, Principal component analysis-linear discriminant analysis feature extractor for pattern recognition, International Journal of Computer Science Issues 8 (6) (2011) 267–270.

[16] S. Robert, V. Myrtille, R. Christelle, A new proposal, multiway discriminant analysis:STATIS-LDA, Journal de la Société Française de Statistique 154 (3) (2013) 31–43.

[17] T. Li, S. Zhu, M. Ogihara, Using discriminant analysis for multi-class classification: an experimental investigation, Knowledge and Information Systems (2006) 453–472.

[18] M. Liberman, K. Davis, M. Grossman, N. Martey, J. Bell, Emotional prosody speech and transcripts, Linguistic Data Consortium.