Application of Feature Subset Selection based on Evolutionary Algorithms for Automatic Emotion Recognition in Speech

Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, Nestor Garay

Computer Science Faculty (University of the Basque Country)
Manuel Lardizabal 1, E-20018 Donostia (Gipuzkoa), Spain
aalvarez031@ikasle.ehu.es

Abstract

The study of emotions in human-computer interaction is a growing research area. Focusing on automatic emotion recognition, work is being performed in order to achieve good results particularly in speech and facial gesture recognition. In this paper we present a study performed to analyze different machine learning techniques validity in automatic speech emotion recognition area. Using a bilingual affective database, different speech parameters have been calculated for each audio recording. Then, several machine learning techniques have been applied to evaluate their usefulness in speech emotion recognition. In this particular case, techniques based on evolutive algorithms (EDA) have been used to select speech feature subsets that optimize automatic emotion recognition success rate. Achieved experimental results show a representative increase in the abovementioned success rate.

1. Introduction

Human beings are eminently emotional, as their social interaction is based on the ability to communicate their emotions and perceive the emotional states of others [1]. *Affective computing*, a discipline that develops devices for detecting and responding to users' emotions [2], is a growing research area [3]. The main objective of affective computation is to capture and process affective information with the aim of enhancing the communication between the human and the computer.

Within the scope of affective computing, the development of affective applications is a challenge that involves analyzing different multimodal data sources. In order to develop such applications, a large amount of data is needed in order to include a wide range of emotionally significant material. Affective databases are a good chance for developing affective recognizers or affective synthesizers.

In this paper different speech paralinguistic parameters have been calculated for the analysis of the human emotional voice, using several audio recordings. These recordings are stored in a bilingual and multimodal affective database. Several works have already been done in which the use of Machine Learning paradigms takes a principal role.

2. Related work

As previously mentioned affective databases provide a good opportunity for training affective applications. This type of databases usually record information such as images, sounds, psychophysiological values, etc. There are some references in the literature that present affective databases and their characteristics. In [4], the authors carried out a wide review of affective databases. Other interesting reviews are the ones provided in [5] and [6].

Many studies have been focused on the different features used in human emotional speech analysis [7, 8]. The number of voice features analysed varies among the studies, but basically most of these are based in fundamental frequency, energy and timing parameters, such as speech rate or mean phone duration.

Works where the use of Machine Learning paradigms take a principal role can also be found in the literature. [9] presented a good reference paper. The Neural Networks Journal devoted a special issue to emotion treatment from a Neural Networks perspective [10]. The work by [4] is related with this paper in the sense of using a Feature Selection method in order to apply a Neural Network to emotion recognition in speech, although both, the methods to perform the FSS and the paradigms used, are different. In this line it has to be pointed out the work by [11] which uses a reduced number of emotions and a greedy approach to select the features.

3. Study of automatic emotion recognition relevant parameters using Machine Learning paradigms

3.1. RekEmozio Database

The RekEmozio bilingual database was created with the aim of serving as an information repository for performing research on user emotion. The aim when building the RekEmozio resource was to add descriptive information about the performed recordings, so that processes such as extracting speech parameters and video features could be carried out on them. Members of different work groups involved in research projects related to RekEmozio have performed several processes for extracting speech and video features; this information was subsequently added to the database. The emotions used were chosen based on [12], and the neutral emotion was added. The characteristics of the RekEmozio database are described in [13]. The languages that are considered in RekEmozio database are Spanish and Basque.

3.2. Emotional feature extraction

For emotion recognition in speech, one of the most important questions is which features should be extracted from the voice signal. Previous studies show us that it is difficult to find

specific voice features that could be used as reliable indicators of the emotion present in the speech [14].

In this work, RekEmozio database audio recordings (stereo wave files, sampled at 44100 Hz) have been processed using standard signal processing techniques (windowing, Fast Fourier Transform, auto-correlation …) to extract a wide group of 32 features which are described below. Supposing that each recording in the database corresponds to one single emotion, only one global vector of features has been obtained for each recording by using some statistical operations. Parameters used are global parameters calculated over entire recordings. Selected features are detailed next (in italics):

- **Fundamental Frequency (F0):** It is the most common feature analyzed in several studies [7, 8]. For F0 estimation we used Sun algorithm [15] and statistics are computed: *Maximum*, *Minimum*, *Mean*, *Range*, *Variance*, *Standard deviation* and *Maximum positive slope in F0 contour*.
- **RMS Energy:** The mean energy of speech quantified by calculating root mean square (RMS) value and 6 statistics *Maximum*, *Minimum*, *Mean*, *Range*, *Variance* and *Standard Deviation*.
- **Loudness:** *Absolute loudness* based on Zwicker's model [16].
- **Spectral distribution of energy:** Each emotion requires a different effort in the speech and it is known that the spectral distribution of energy varies with speech effort [7]. We have computed energy in *Low band*, between 0 and 1300 Hz, *Medium band*, between 1300 and 2600 Hz and *High band* from 2600 to 4000 Hz [17].
- **Mean Formants and Bandwidth:** Energy from the sound source (vocal folds) is modified by the resonance characteristics of the vocal tract (formants). Acoustic variations dues to emotion are reflected in formants [18]. The *first three mean Formants*, and their corresponding *mean Bandwidths*.
- **Jitter:** *Perturbation in vibration of vocal chords*. It is estimated based on the model presented by [19].
- **Shimmer:** *Perturbation cycle to cycle of the energy*. Its estimation is based on the previously calculated absolute loudness.
- **Speaking Rate:** Rhythm is known to be an important aspect in recognition of emotion in speech. Progress has been made on a simple aspect of rhythm, the alternation between speech and silence [7]. The speaking rate estimation has been divided in 6 values based on their duration with respect to the whole elocution: *Duration of voice* part, *Silence* part, *Maximum voice* part, *Minimum voice* part, *Maximum silence* part and *Minimum silence* part.

3.3. Machine Learning standard paradigms used

In the supervised learning task, a classification problem has been defined where the main goal is to construct a model or a classifier able to manage the classification itself with acceptable accuracy. With this aim, some variables are to be used in order to identify different elements, the so called predictor variables. For the current problem, each sample is composed by the set of 32 speech related values, while the label value is one of the seven emotions identified. The single paradigms used in our experiments that come from the family of *Machine Learning* (ML) are briefly introduced

3.3.1. Decision trees

A decision tree consists of nodes and branches to partition a set of samples into a set of covering decision rules. In each node, a single test or decision is made to obtain a partition. The starting node is usually referred as the root node. In each node, the goal is to select an attribute that makes the best partition between the classes of the samples in the training set [20], [21]. In our experiments, two well-known decision tree induction algorithms are used, ID3 [22] and C4.5 [23].

3.3.2. Instance-Based Learning

Instance-Based Learning (IBL) has its root in the study of nearest neighbor algorithm [24] in the field of machine learning. The simplest form of nearest neighbor (NN) or k-nearest neighbor (k-NN) algorithms simply stores the training instances and classifies a new instance by predicting the same class its nearest stored instance has or the majority class of its k nearest stored instances have, respectively, according to some distance measure as described in [25]. The core of this non-parametric paradigm is the form of the similarity function that computes the distances from the new instance to the training instances, to find the nearest or k-nearest training instances to the new case. In our experiments the IB paradigm is used, an inducer developed in the MLC++ project [26] and based on the works of [27] and [28].

3.3.3. Naive Bayes classifiers

The Naive-Bayes (NB) rule [29] uses the Bayes theorem to predict the class for each case, assuming that the predictive genes are independent given the category. To classify a new sample characterized by d genes X = (X_1,X_2, ...,X_d), the NB classifier applies the following rule:

$$c_{N-B} = \arg\max_{cj \in C} p(cj) \prod_{i=1}^{d} p(xi \mid cj) \quad (1)$$

where C_{N-B} denotes the class label predicted by the Naive-Bayes classifier and the possible classes of the problem are grouped in $C = \{c_1, \ldots, c_l\}$. A normal distribution is assumed to estimate the class conditional densities for predictive genes. Despite its simplicity, the NB rule obtains better results than more complex algorithms in many domains.

3.3.4. Naive Bayesian Tree learner

The naive Bayesian tree learner, NBTree [30], combines naive Bayesian classification and decision tree learning. It uses a tree structure to split the instance space into sub-spaces defined by the paths of the tree, and generates one naive Bayesian classifier in each sub-space.

3.3.5. Feature Subset Selection by Estimation of Distribution Algorithms

The basic problem of ML is concerned with the induction of a model that classifies a given object into one of several known

classes. In order to induce the classification model, each object is described by a pattern of *d* features. Here, the ML community has formulated the following question: *are all of these d descriptive features useful for learning the 'classification rule'?* On trying to respond to this question, we come up with the Feature Subset Selection (FSS) [31] approach which can be reformulated as follows: *given a set of candidate features, select the 'best' subset in a classification problem.* In our case, the 'best' subset will be the one with the best predictive accuracy.

Most of the supervised learning algorithms perform rather poorly when faced with many irrelevant or redundant (depending on the specific characteristics of the classifier) features. In this way, the FSS proposes additional methods to reduce the number of features so as to improve the performance of the supervised classification algorithm.

FSS can be viewed as a search problem [32], with each state in the search space specifying a subset of the possible features of the task. Exhaustive evaluation of possible feature subsets is usually unfeasible in practice due to the large amount of computational effort required. In this way, any feature selection method must determine the nature of the search process. In the experiments performed, an Estimation of Distribution Algorithm (EDA) has been used which has the model accuracy as fitness function.

To assess the goodness of each proposed gene subset for a specific classifier, a wrapper approach is applied. In the same way as supervised classifiers when no gene selection is applied, this wrapper approach estimates, by using the 10-fold crossvalidation [33] procedure, the goodness of the classifier using only the variable subset found by the search algorithm.

4. Experimental Results

The above mentioned methods have been applied over the crossvalidated data sets using the MLC++ library [26]. Each dataset corresponds to a single actor. Experiments were carried out with and without FSS in order to extract the accuracy improvement introduced by the feature selection process. Tables 1 and 2 show the classification results obtained using the whole set of variables, for Basque and Spanish languages respectively. Each column represents a female (Fi) of male (Mi) actor, and mean values corresponding to each classifier/gender is also included. Last column presents the total average for each classifier.

Table 1: 10-fold crossvalidation accuracy for Basque Language using the whole variable set

	Female				*Male*					*Total*
	F1	F2	F3	**mean**	M1	M2	M3	M4	**mean**	
IB	35.4	48.8	35.2	39.8	44.2	49.3	36.9	40.9	42.8	41.5
ID3	38.7	45.5	44.7	**42.9**	46.7	46.9	43.3	51.1	47.0	45.3
C4.5	41.5	52.2	35.0	42.9	60.4	53.3	45.1	49.5	**52.0**	**48.1**
NB	42.9	45.8	37.7	42.1	52.2	44.1	36.2	41.4	43.5	42.9
NBT	42.3	39.8	35.2	39.1	53.1	46.2	45.2	43.3	46.9	43.6

Table 2: 10-fold crossvalidation accuracy for Spanish Language using the whole variable set

	Female						*Male*						*Total*
	F1	F2	F3	F4	F5	**mean**	M1	M2	M3	M4	M5	**mean**	
IB	34.6	43.6	54.6	54.6	38.2	45.1	25.5	33.6	51.8	47.7	33.6	**38.4**	**41.8**
ID3	36.4	52.7	49.1	47.3	42.7	**45.6**	20.9	30.9	40.9	47.3	40.0	36.0	40.8
C4.5	30.9	50.0	46.4	43.6	42.7	42.7	29.1	31.8	46.4	42.7	35.5	37.1	39.9
NB	38.2	42.7	49.1	40.0	42.7	42.5	24.6	30.9	49.1	45.5	34.6	36.9	39.7
NBT	42.7	43.6	49.1	50.0	39.1	44.9	18.2	27.3	40.9	48.2	42.7	35.5	40.2

Results don't seem very impressive; ID3 best classifies the emotions for female actresses, for both Basque and Spanish languages, while C4.5 outstands for Basque male actors and IB for Spanish male actors.

Results obtained after applying FSS are more appealing, as can be seen in Tables 3 and 4. There, classifier IB appears as the best paradigm for all the categories, female and male, and Basque and Spanish languages. Moreover, the accuracies outperform the previous ones in more than 15%. It must also be highlighted that FSS improves the well classified rate for all the ML paradigms, as it can be seen in Figure 1.

Table 3: 10-fold crossvalidation accuracy for Basque Language using FSS

	Female				*Male*					*Total*
	F1	F2	F3	**mean**	M1	M2	M3	M4	**mean**	
IB	63.0	68.0	59.3	**63.5**	72.7	67.4	61.0	62.8	**65.9**	**64.9**
ID3	62.7	60.5	65.5	62.9	72.7	62.0	56.5	62.7	63.4	63.2
C4.5	60.2	66.0	60.0	62.1	71.8	62.8	60.1	63.6	64.6	63.5
NB	64.5	64.6	48.9	59.3	74.6	62.5	62.7	60.0	64.9	62.5
NBT	58.6	61.1	54.8	58.1	74.4	59.9	62.7	59.4	64.1	61.6

Table 4: 10-fold crossvalidation accuracy for Spanish Language using FSS

	Female						***Male***						*Total*
	F1	F2	F3	F4	F5	**mean**	M1	M2	M3	M4	M5	**mean**	
IB	61.8	66.4	75.5	71.8	68.2	**68.7**	42.7	57.3	69.1	63.6	60.9	**58.7**	**63.7**
ID3	59.1	66.4	66.4	60.0	61.8	62.7	42.7	51.8	66.4	61.8	60.0	56.5	59.6
C4.5	57.3	62.7	64.6	65.5	63.6	62.7	43.6	56.4	65.5	64.6	56.4	57.3	60.0
NB	54.6	59.1	68.2	65.5	60.0	61.5	40.9	48.2	64.6	59.1	51.8	52.9	57.2
NBT	53.6	66.4	63.6	58.2	60.0	60.4	38.2	47.3	60.0	63.6	59.1	53.6	57.0

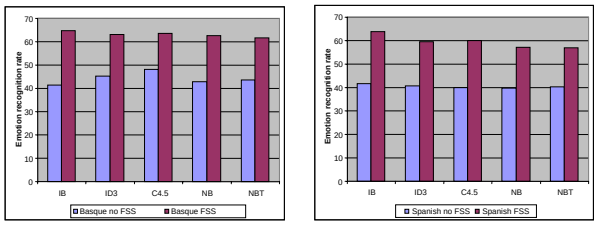

Figure 1: The improvement in Basque and Spanish languages using FSS in all the classifiers.

5. Conclusions and future work

RekEmozio database has been used to training some automatic recognition systems. In this paper we have shown that applying FSS enhances classification rates for the ML paradigms that we have used (IB, ID3, C4.5, NB and NBTree).

An analysis of the selected features by FSS is required. Also, the speech data should be combined with visual information. This combination could be performed by means of a multiclassifier model [34].

6. References

[1] Casacuberta, D., *La mente humana: Diez Enigmas y 100 preguntas (The human mind: Ten Enigmas and 100 questions)*, Océano (Ed), Barcelona, Spain, ISBN: 84-7556-122-5, 2001.

[2] Picard, R. W., *Affective Computing*, MIT Press, Cambridge, MA, 1997.

[3] Tao, J., Tan, T., "Affective computing: A review", *In: J. Tao, T. Tan, R. W. Picard (eds.): Lecture Notes in Computer Science, Vol. 3784 - Proceedings of The First International Conference on Affective Computing & Intelligent Interaction (ACII'05)*, Beijing, China, 981-995, 2005.

[4] Cowie, R., Douglas-Cowie, E., Cox, C., "Beyond emotion archetypes: Databases for emotion modelling using neural networks", *Neural Networks, Vol. 18, 2005, p 371-388.*

[5] Humaine, "Retrieved January 10, 2007", from [http://emotion-research.net/wiki/Databases], (n.d.).

[6] López, J.M., Cearreta, I., Fajardo, I., Garay, N., "Validating a multimodal and multilingual affective database", *To be published in Proceedings of HCI International*, 2007.

[7] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine, Vol. 18(1), 2001, p 32-80.*

[8] Schröder, M., *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*, Ph.D. thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University, 2004.

[9] Dellaert, F., Polzin, T., Waibel, A., "Recognizing Emotion in Speech", *In Proc. of ICSLP'96*, 1996.

[10] Taylor, J. G., Scherer, K., Cowie, R., "Neural Networks, special issue on Emotion and Brain", *Vol. 18, Issue 4, 2005, p 313-455.*

[11] Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann, H., "Recognition of emotion in a realistic dialogue scenario", *In Proc. ICSLP'00*, p 665-668, 2000.

[12] Ekman, P., Friesen, W., *Pictures of facial affect*, Consulting Psychologist Press, Palo Alto, CA, 1976.

[13] López, J.M., Cearreta, I., Garay, N., López de Ipiña, K., Beristain, A., "Creación de una base de datos emocional bilingüe y multimodal", *In Redondo, M.A., Bravo C., Ortega M. (Eds). Proceeding of the 7th Spanish Human Computer Interaction Conference, Interacción-06*, Puertollano, p 55-66, 2006.

[14] Laukka, P., *Vocal Expression of Emotion. Discrete-emotions and Dimensional Accounts*, Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences, pp 141, 80, ISBN 91-554-6091-7, Uppsala, 2004.

[15] Sun, X., "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio", *To appear in the Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, 2002

[16] Fernandez, R., *A Computational Model for the Automatic Recognition of Affect in Speech*, Ph.D. thesis, Massachusetts Institute of Technology, 2004.

[17] Kazemzadeh, A., Lee, S., Narayanan, S., "Acoustic correlates of user response to errors in human-computer dialogues", *Proc. IEEE ASRU, (St. Thomas, U.S. Virgin Islands)*, 2003 (December).

[18] Bachorowski, J.A., Owren, M. J., "Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context" *Psychological Science, Vol. 6,1995, p 219-224.*

[19] Rothkrantz, L.J.M., Wiggers, P., van Wees, J.W.A., van Vark, R.J., "Voice stress analysis", *Proceedings of Text, Speech and Dialogues 2004*, 2004.

[20] Martin, J.K., "An exact probability metric for Decision Tree splitting and stopping, Machine Learning", 1997, p 28(2/3).

[21] Mingers, J., "A comparison of methods of pruning induced Rule Trees, Technical Report", Coventry, England: University of Warwick, School of Indutrial and Business Studies, 1988.

[22] Quinlan, J.R., "Induction of Decision Trees", *Machine Learning, Vol 1, 1986, p 81-106.*

[23] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, Inc. Los Altos, California, 1993.

[24] Dasarathy, B.V., "Nearest Neighbor (NN) Norms: NN Pattern Recognition Classification Techniques", *IEEE Computer Society Press*, 1991.

[25] Ting, K.M., *Common issues in Instance-Based and Naive-Bayesian classifiers*, Ph.D. Thesis, Basser Department of Computer Science, The University of Sidney, Australia, 1995.

[26] Kohavi, R., Sommerfield, D., Dougherty, J., "Data mining using MLC++, a Machine Learning Library in C++", *International Journal of Artificial Intelligence Tools, Vol. 6 (4), 1997, p 537-566*, [http://www.sgi.com/Technology/mlc/].

[27] Aha, D., Kibler, D., Albert, M.K., "Instance-Based learning algorithms", *Machine Learning, Vol. 6*, 37-66, 1991.

[28] Wettschereck, D., *A study of distance-based Machine Learning Algorithms*, Ph.D. Thesis, Oregon State University, 1994.

[29] Minsky, M. "Steps towards artificial intelligence", *Proceedings of the IRE*, 49, 8–30, 1961.

[30] Kohavi, R., "Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid", *In: Simoudis, E., Han, J.-W., Fayyad, U. M. (eds.): Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press*, p 202-207, 1996.

[31] Liu, H., Motoda, H., *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers, 1998.

[32] Inza, I., Larrañaga, P., Etxeberria, R., Sierra, B., "Feature subsetselection by Bayesian network-based optimization", *Artificial Intelligence, Vol. 123, 2000, p 157-184.*

[33] Stone, M., "Cross-validation choice and assessment of statistical procedures", *Journal Royal of Statistical Society, Vol. 36, 1974, p 111-147.*

[34] Gunes, V., Menard, M., Loonis, P., Petit-Renaud, S., "Combination, cooperation and selection of classiers", *A state of the art. International Journal of Pattern Recognition, Vol. 17, 2003, p 1303-1324.*