Machine Learning Approach for Emotion Recognition in Speech

Martin Gjoreski and Hristijan Gjoreski
Department of Intelligent Systems, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: {martin.gjoreski, hristijan.gjoreski}@ijs.si

Andrea Kulakov
Faculty of Computer Science and Engineering, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
E-mail: andrea.kulakov@finki.ukim.mk

Keywords: machine learning, emotions, speech, recognition, Auto-WEKA

Received: December 11, 2014

This paper presents a machine learning approach to automatic recognition of human emotions from speech. The approach consists of three steps. First, numerical features are extracted from the sound database by using audio feature extractor. Then, feature selection method is used to select the most relevant features. Finally, a machine learning model is trained to recognize seven universal emotions: anger, fear, sadness, happiness, boredom, disgust and neutral. A thorough ML experimental analysis is performed for each step. The results showed that 300 (out of 1582) features, as ranked by the gain ratio, are sufficient for achieving 86% accuracy when evaluated with 10 fold cross-validation. SVM achieved the highest accuracy when compared to KNN and Naive Bayes. We additionally compared the accuracy of the standard SVM (with default parameters) and the one enhanced by Auto-WEKA (optimized algorithm parameters) using the leave-one-speaker-out technique. The results showed that the SVM enhanced with Auto-WEKA achieved significantly better accuracy than the standard SVM, i.e., 73% and 77% respectively. Finally, the results achieved with the 10 fold cross-validation are comparable and similar to the ones achieved by a human, i.e., 86% accuracy in both cases. Even more, low energy emotions (boredom, sadness and disgust) are better recognized by our machine learning approach compared to the human.

Povzetek: Predstavljeno je prepoznavanje čustev iz govora s pomočjo strojnega učeanja.

1 Introduction and related work

Human capabilities for perception, adaptation and learning about the surroundings are often three main compounds of the definition about what intelligent behaviour is. In the last few decades there are many studies suggesting that one very important compound is left out of this definition about intelligent behaviour. That compound is emotional intelligence. Emotional intelligence is the ability of one to feel, express, regulate his own, to recognize and handle the emotional state of others. In psychology the emotional state is defined as complex state that results in psychological and physiological changes that influence our behaving and thinking [1] .

With the recent advancements of the technology and the growing research areas like machine learning (ML), audio processing and speech processing, the emotional states will be inevitable part of the human-computer interaction. There are more and more studies that are working on providing the computers with abilities like recognizing, interpretation and simulation of emotional states.

Automatic emotion recognition is part of growing research areas such as industry for robots [22], automobile industry, entertainment industry, marketing industry, and similar. The automatic emotion recognition also can be also used for improving the accuracy in speech recognition. It is expected that automatic emotion recognition will change the human-computer communication [23].

The goal of the emotion recognition systems is to recognize the emotional state that is experiencing the speaker. The focus is usually on how something is said, and not what is said. Besides the approaches where only the speaker's voice is analysed, there are several different approaches for recognizing the emotional state. In some approaches the voice and the spoken words are analysed [2] . Some are focused only on the facial expressions [3] . Some are analysing the reactions in the human brain for different emotional states [4] . Also there are combined approaches where combination of the mentioned approaches is used [5] .

In general, there are two approaches in human emotions analysis. In the first approach the emotions are represented as discrete and distinct recognition classes [6] . The other approach represents the emotional states in 2D or 3D space, where parameters like emotional distance, level of activeness, level of dominance and level of pleasure are be observed [7] .

In this research we present a ML approach for automatic recognition of emotions from speech. Our

approach uses the discrete type of approach; therefore the emotional states are represented by seven classes: anger, fear, sadness, happiness, boredom, disgust and neutral. Even though ML approaches have been proposed in the literature, our approach improves upon them by performing a thorough ML analysis, including methods for: feature extraction, feature standardization, feature selection, algorithm selection, and algorithm parameters optimization. With this analysis, we try to find the optimal ML configuration of: features, algorithms and parameters, for the task of emotion recognition in speech.

The remainder of this paper is organized as follows. Section 2 is a brief overview of speech emotion analysis. In Section 3 our ML approach for emotion recognition is presented. Section 4 presents the experimental setup and the experimental results. Finally, the conclusion and a brief discussion about the results is given.

2 Speech emotion analysis

Speech emotion analysis refers to usage of methods to extract vocal cues from speech as a marker for emotional state, mood or stress. The main assumption is that there are objectively measureable cues that can be used for predicting the emotional state of the speaker. This assumption is quite reasonable since the emotional states arouse physiological reactions that affect the process of speech production. For example, the emotional state of fear usually initiates rapid heartbeat, rapid breathing, sweating and muscle tension. As a result of these physiological activities there are changes in the vibration of the vocal folds and the shape of the vocal tract. All of this affects the vocal characteristics of the speech which allows to the listener to recognize the emotional state that the speaker is experiencing [8] .

The basic speech audio features that are used for speech emotion recognition are: fundamental frequency (human perception for fundamental frequency is pith), power, intensity (human perception for intensity is loudness), duration features (ex. rate of speaking) and vocal perturbations. The main question is: Are there any objective feature profiles of the voice that can be used for speaker emotion recognition? A lot of studies are done for the sake of providing such feature profiles that can be used for representation of the emotions, but results are not always consistent. For some basic problems like distinguishing normal speech from angry speech or distinguishing normal speech from bored speech the experimental results converge [9] . For example such converging results are showing that compared to normal speech, when expressing fear or happiness human speak with higher pitch (fundamental frequency).

Figure 1 shows an example of audio wave (top graphs) and pitch (bottom graphs) of normal speech (left) and angry speech (right). The missing parts of the pitch graphs are parts of the speech signals which would not have foundation in human perception. They relate to parameters (ex. silence threshold, voicing threshold) of the pitch analysis algorithms. The graphs show that by using the pitch as a feature, one can note the different characteristics of speech under different emotional states. On the left we have normal speech and on the right we have angry speech of the same words by the same person. On the left lower graph (normal speech) we can see that the pitch is around 120Hz and it is monotone. On the other hand the right lower graph (angry speech) we can see higher pitch (there are parts where the pitch goes up to 500Hz) and there is noticeable variability (there are parts where the pitch goes from 500Hz to 100Hz and vice versa). This simple analysis is just an example of how we can compare speech signals by using their physical characteristics. This simple approach cannot be used for speech emotion recognition. The problem arises when we have to distinguish emotional states like anger from happiness or fear from happiness. By using the basic speech audio features for describing these emotional states, the feature profiles are quite similar so distinguishing them is hard.

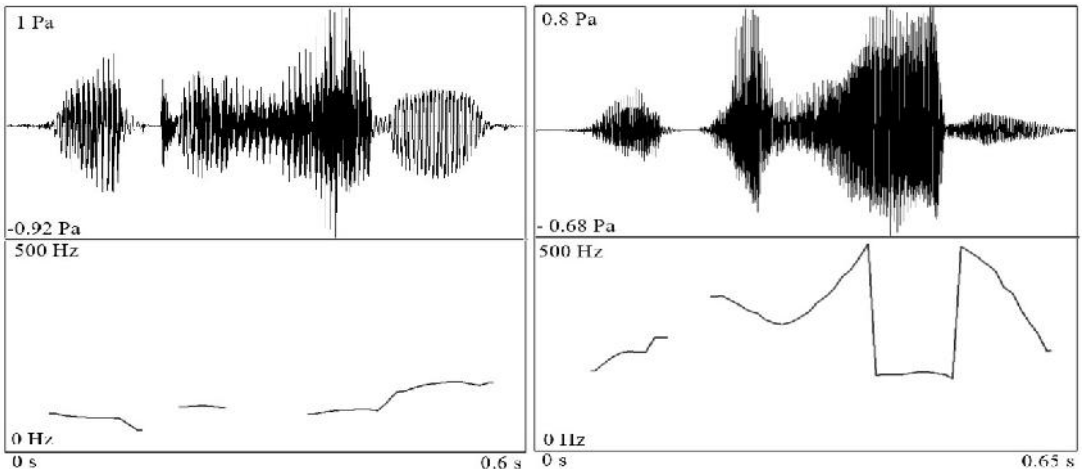

Figure 1. Audio wave (top graphs) and pitch (bottom graphs) of normal speech (left) and angry speech (right). The missing parts of the bottom graphs are parts of the speech signals which would not have foundation in human perception.

In the last few years, new method is introduced where static feature vectors are obtained by using so called acoustic Low-Level Descriptors (LLDs) and descriptive statistical functionals [10] . By using this approach a big number of large feature vectors is obtained. The downside is that not all of the feature vectors are of good value, especially not for emotion recognition. For that reason a feature selection method is often used.

3 ML Approach

Figure 2 shows the whole process of the ML speech emotion recognition used in this study. First, an emotional speech database is used, which consists of simulated and annotated utterances. Next, feature extraction is performed by using open source feature extractor. Then, feature selection method is used for decreasing the number of features and selecting only the most relevant ones. Finally, the emotion recognition is performed by a classification algorithm.

3.1 Emotional speech database

There are several emotional speech databases that are extensively used in the literature [11] : German, English, Japanese, Spanish, Chinese, Russian, Dutch etc. One main characteristic of an emotional speech database is the type of the emotions expressed in the speech: whether they are simulated or they are extracted from real life situations. The advantage of having a simulated speech is that the researcher has a complete control over the emotion that it is expressed and complete control over the quality of the audio. However, the disadvantage is that there is loss in the level of naturalness and spontaneity. On the other hand, the non-simulated emotional databases consist of a speech that is extracted from real life scenarios like call-centers, interviews, meetings, movies, short videos and similar situations where the naturalness and spontaneity is kept. The disadvantage is that in these databases there is not a complete control over the expressed emotions. Also the low quality of the audio can be problem.

For this research the Berlin emotional speech database [12] is used, which is one of the most exploited databases for speech emotion analysis. It consists of 535 audio files, where 10 actors (5 male and 5 female) are pronouncing 10 sentences (5 short and 5 long). The sentences are chosen so that all 7 emotions that we are analyzing can be expressed. The database is additionally checked for naturalness by testing it with 20 human volunteers. The volunteers were supposed to recognize and rate the naturalness of the expressed emotion by listening to random utterance. The utterances that were rated with more than 60% naturalness and from which the expressed emotion was recognized with more than 80%, were included in the final database. In Figure 3 statistics for the Berlin emotional speech database is shown. We can see information about the number of instances per class and information about the human recognition rate obtained from the tests.

Table 1. Statistics for the Berlin emotional speech database, including the human recognition rate.

Emotions	Number of instances	Human recognition rate (%)
Anger	127	96.2
Neutral	79	88.2
Fear	69	87.3
Boredom	81	86.2
Happiness	71	83.7
Sadness	62	80.7
Disgust	46	79.6

3.2 Feature Extraction

The feature extraction tool used in this research is OpenSmile (Open Speech and Music Interpretation by Large Space Extraction) [13] . It is a commonly used tool for signal processing and feature extraction when ML approach is applied on sound data. OpenSmile provides configuration files that can be used for extracting predefined features. For this research the configuration file ‘emobase2010’ is used. By using the ‘emobase2010’ configuration file in total 1582 features are extracted [14] . OpenSmile computes LLDs from basic speech features (pitch, loudness, voice quality) or representations of the speech signal (cepstrum, linear predictive coding). On these LLDs functionals are applied and static feature vectors are computed, therefore static classifiers can be used. The functionals that are applied are: extremes (position of mix/min value), statistical moments (first to forth), percentiles (ex. the first quartile), duration (ex. percentage of time the signal is above threshold) and regression (ex. the offset of a linear approximation of the contour)

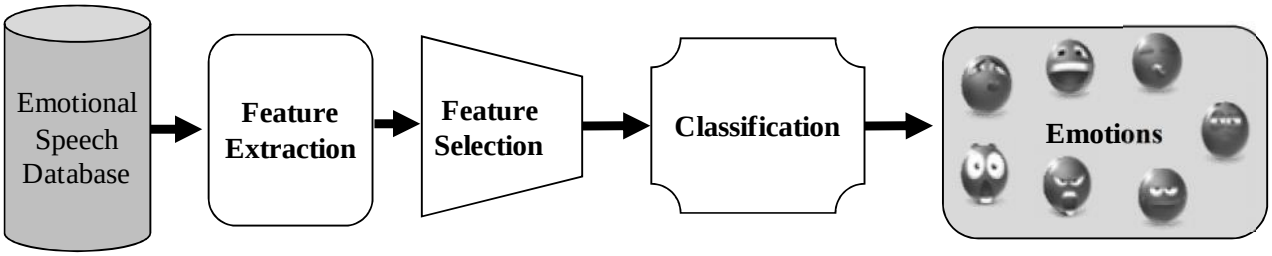

Figure 2. ML approach for emotion recognition.

After the feature extraction the feature vectors are standardized so the distribution of the values of each feature is with mean equal to 0 and standard deviation equal to 1. This way, the values for each feature are on the same scale from -1 to 1, preventing some features (with bigger values) to have more influence when creating the ML model. This is an important step in ML, especially for classification algorithms that do not have mechanism for feature standardization.

3.3 Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context [15] .

To deal with this issue, we used a method for feature selection. Features were ranked with an algorithm for feature ranking and experiments were performed with varying number of top ranked features. For ranking the features the well-known gain ratio [16] algorithm is used. Gain ratio is the ratio of information gain and the entropy of one feature. It is used to avoid overestimation of multi-valued features (the drawback of information gain).The algorithm is used as it is implemented in Orange ML toolkit [18] .

3.4 Classification

Once the features are extracted, standardized and selected, they are used to form the feature vector database. Each data sample in the data base is an instance, i.e., feature vector, used for classification. Because each instance is labeled with the appropriate emotion, supervised classification algorithms are used. In our experiments three commonly used algorithms for classification were tested, K-Nearest Neighbors (KNN) [19] , Naïve Bayes [21] and Support Vector Machine (SVM) [20] . KNN is an instance-based classifier (lazy) that does not learn a model, but it uses similarity metrics (e.g. Euclidian distance) to find the K most similar training instances and apply the majority class value (emotion) of these K instances. Naïve Bayes is a probability-based algorithm. It applies a probability theory to the feature values in order to create a model that divides the instances according to the class values. The SVM is the most complex of the three algorithms. Its goal is to find hyperplanes in the attribute's space in order to maximize the margin between instances that belong to distinct classes. It uses a kernel function in order to create non-linear classifiers.

We performed thorough experiments with each of the classification models, and once we selected the one with the highest recognition accuracy, we further enhanced its accuracy with Auto-WEKA [25] . Auto-WEKA is a ML tool that is using approach for parameter optimization of classification algorithms. It searches to the huge space of algorithm parameters and by using an intelligent optimization functions finds the near optimal parameter setting, which should increase the accuracy of the chosen algorithm. The problem of parameter optimization is viewed as a single hierarchical parameter optimization in which even the classification algorithm is considered as a parameter. The root-level parameter is the learning algorithm and the rest of the searching space is depending on the chosen parameter (algorithm) in the previous level. The main idea of Auto-WEKA is that search in the combined space of algorithms and parameters results with better-performing models than standard algorithm selection and parameter optimization methods. For searching the huge space of parameters Auto-WEKA is using Bayesian optimization methods TPE [23] and SMAC [24] .

For the model selection problem Auto-WEKA splits the train data in k folds so that each of the pre-selected learning algorithms is tested with k-fold cross validation. The ultimate goal of the model selection approach is to find an algorithm with optimal generalization performance. The selection criteria is minimizing the misclassification rate. For the parameter optimization problem Auto-WEKA is using hierarchical parameter search and again the goal of the optimization problem is minimization of the misclassification rate.

4 Experiments

4.1 Experimental setup

The experiments were conducted in the following order. First, the OpenSmile feature extraction tool extracted 1582 feature. Next, all of the extracted features were ranked using the gain ratio algorithm [16] . Then, we tested the accuracy of the three algorithms by using different number of features as ranked by the ranking algorithm. In particular we used 50, 100, 200, 400, 500, 600, 750, 1000, and 1582. As an evaluation metrics, we used the 10 fold cross-validation, which is a gold-standard technique for evaluating datasets when the instances are not structured or time dependent (e.g., time series). It usually gives a good estimate of the accuracy of an algorithm.

In the next step, the algorithm with the highest accuracy is further evaluated with Leave-One-Speaker-Out (LOSO) technique. Because in our case the data are collected by 10 individuals (speakers) the model was trained on the data recorded for nine people and tested on the remaining person. This procedure was repeated for each person (10 times). The LOSO evaluation approach is more reliable than using the same person’s data for training and testing if the model is intended to be used by unknown people (not included in the training dataset).

Finally, we applied Auto-WEKA toolkit in order to optimize the parameters of the chosen algorithm, i.e., parameters, and therefore to improve the accuracy. For each Auto-WEKA experiment the SMAC optimization method was chosen. The optimization timeout for Auto-WEKA was set to 24h. The training memory per experiment was set to 1000MB and the training run

timeout was set to 150 min. For these experiments the Auto-WEKA's feature selection module was turned off.

For each comparison, tests to confirm the statistical significance of the results were performed using paired Student's T-test with a significance level of 5%.

Three, commonly used in ML, evaluation metrics were analyzed: the recall, precision and accuracy. The following formulas define each of the metrics, where Q can be any emotion that we are trying to recognize (happiness, neutral, etc.):

$$recall = \frac{\text{No. of correctly recognized emotions labeled as Q}}{\text{No. of all the emotions labeled as Q}} \quad (1)$$

$$precision = \frac{\text{No. of correctly recognized emotions labeled as Q}}{\text{No. of all the emotions recognized as Q}} \quad (2)$$

$$accuracy = \frac{\text{No. of correctly recognized emotions of all types}}{\text{No. of all the emotions}} \quad (3)$$

4.2 Experimental Results

Once the features were extracted and ranked by the gain ratio algorithm, we compared the accuracy of the three ML algorithms (KNN, SVM and Naïve Bayes) for different number of top-ranked features (50, 100, 200, 400, 500, 600, 750, 1000, and 1582). The results presented in Figure 3 show that, as the number of features increases up to 400, also the accuracy increases for each of the three algorithms. After that, the accuracy drops for the SVM, and small (statistically insignificant) improvements are noticed for the other two algorithms. The decrease in performance as the number of features increases is due to overfitting. This is especially notable for the SVM, which was in a way expected because its model is more complex compared to the KNN and Naive Bayes and this complexity usually increases as the number of features increases. If too many not relevant features are used, it will overfit on the training data and the accuracy on the test data will drop (which is the case in our experiments).

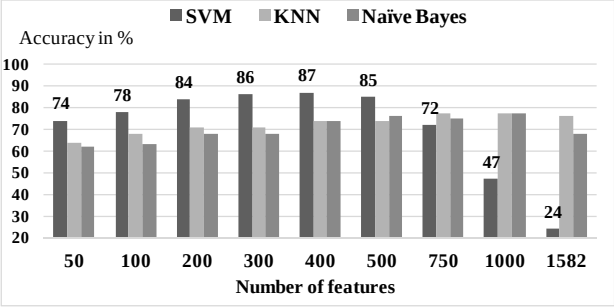

Figure 3. SVM, KNN and Naïve Bayes accuracy for 10 fold-cross validation with varying number of features.

We additionally analysed the results obtained by the SVM, which achieved the highest accuracy, i.e., 87% when the top ranked 400 features are used. However, there was no statistical difference between the accuracy achieved for 300 and 400 features. Therefore, for further analysis we used the top ranked 300 features, which was a good tradeoff between accuracy and number of features.

Figure 4 shows the confusion matrix of the SVM for the top ranked 300 features. Additionally we present the recall and the precision for each class (emotion), and the overall accuracy. The highest precision and recall are achieved for the class "sadness" and the lowest are achieved for the class "happiness". Also we can see that the classes "anger" and "happiness" are often mixed by the classifier. The class "fear" is mixed with all other 6 classes.

10 fold cross-valid. Conf. Matrix SVM		PREDICTED CLASS							Recall (%)
		A	B	D	F	H	N	S	
REAL CLASS	Anger (A)	**116**	0	1	0	10	0	0	91
	Boredom (B)	0	**72**	3	1	0	3	2	89
	Disgust (D)	2	2	**40**	1	0	1	0	87
	Fear (F)	3	1	1	**53**	4	6	1	77
	Happiness (H)	17	0	0	2	**51**	1	0	72
	Neutral (N)	0	6	0	2	1	**69**	1	87
	Sadness (S)	0	0	0	0	0	1	**61**	98
Precision (%)		84	89	89	90	77	85	94	**Acc = 86%**

Figure 4. Confusion matrix for SVM obtained with 10 fold cross-validation with the top ranked 300 features.

In the next step, we evaluated the SVM algorithm with LOSO technique. We compared the accuracy of the standard SVM (with default parameters) and the one enhanced by the Auto-WEKA. The results are shown for each test subject (speaker) individually in Figure 5. The results show that the SVM enhanced with Auto-WEKA achieved significantly better accuracy than the standard SVM, except for the first two speakers (S1 and S2).

Also we can see that the accuracy depends on test subject. For example by using Auto-WEKA the lowest average accuracy (64%) is obtained when the speaker S2 is used as a test speaker and the highest average accuracy (91%) is obtained when the user S8 is used as test speaker.

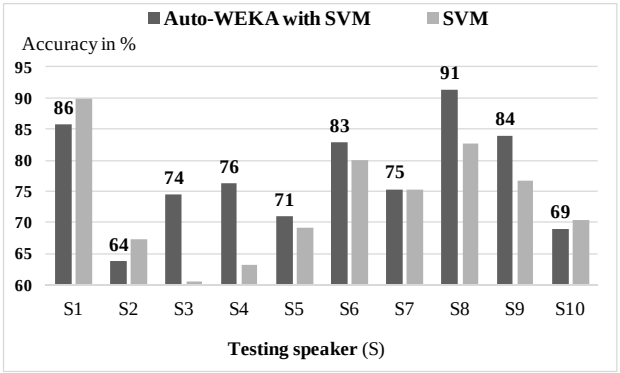

Figure 5. Auto-WEKA and SVM classification accuracy for LOSO with top ranked 300 features

Figure 6 shows more detailed analysis, i.e., confusion matrix of the results achieved by the SVM enhanced with Auto-WEKA. The highest precision and recall are achieved for the class "sadness" and the lowest are achieved for the class "happiness". Also we can see that the class "anger" is mixed with the class "happiness" and vice versa. Also the class "boredom" is mixed with the class "disgust". For the classes "fear" and "neutral"

there is no single class that can be pointed as mixing class. These classes are mixed with several others.

LOSO Conf. Matrix Auto-weka		PREDICTED CLASS							Recall (%)
		A	B	D	F	H	N	S	
REAL CLASS	Anger (A)	**105**	0	1	1	20	0	0	83
	Boredom (B)	0	**63**	11	0	0	4	3	78
	Disgust (D)	2	1	**38**	1	1	3	0	83
	Fear (F)	7	1	3	**46**	3	8	1	67
	Happiness (H)	22	0	0	3	**44**	2	0	62
	Neutral (N)	0	7	4	4	4	**59**	1	75
	Sadness (S)	0	3	0	0	0	1	**58**	94
Precision (%)		77	84	67	84	61	77	92	Acc = 77%

Figure 6. Confusion matrix for Auto-WEKA obtained with LOSO cross-validation with top ranked 300 features.

Finally, we compared the recognition accuracy achieved by a human (manual recognition) and the two ML techniques: SVM trained and tested with 10 fold cross-validation and SVM trained and tested with LOSO (the results are shown in Figure 7). The human recognition rate is obtained from the tests for checking the naturalness of the database. That is, 20 volunteers were asked to recognize the emotions from the audio files. The results show that the SVM trained and tested with 10 fold cross-validation achieved similar results as the human; on average, both achieve 86% accuracy. However, low energy emotions (boredom, sadness and disgust) are better recognized by ML compared to human. This means that for these emotions, the human ear requires additional information, which can be easily extracted using a software tools. The comparison to the LOSO shows that a model trained on subjects different from the ones used for testing should achieve significantly lower accuracy, i.e., 77%.

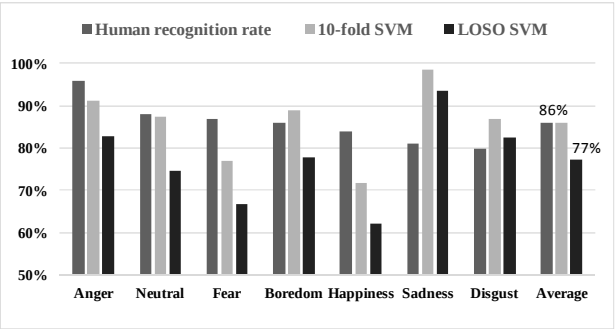

Figure 7. Comparison between the recognition accuracy achieved by a human, and the two ML techniques: 10 fold cross-validation and LOSO.

5 Conclusion

Even though ML approaches have been proposed in the emotion recognition literature, our approach improves upon them by performing a thorough ML analysis, including methods for: feature extraction and standardization, feature selection analysis, algorithm selection analysis, and algorithm parameters optimization. With this whole analysis, we tried to find the optimal ML configuration of: features, algorithms and parameters, for the task of emotion recognition in speech.

The results showed that the top 300 features, as ranked by the gain ratio, are sufficient for achieving 86% accuracy in emotion recognition. Adding more would just cause overfitting. SVM achieved the highest accuracy and significantly outperformed the KNN and Naive Bayes.

When trained and tested with 10 fold cross-validation, SVM achieved 86% over all the emotions. The per-emotion analysis shows that the highest precision and recall were achieved for the “sadness” and the lowest were achieved for the “happiness”. Also the “anger” and “happiness” were often mixed by the classifier, and the “fear” was mixed with all other 6 emotions.

In the next step, we evaluated the SVM algorithm with LOSO technique. We compared the accuracy of the standard SVM (with default parameters) and the one enhanced by the Auto-WEKA. The results showed that the SVM enhanced with Auto-WEKA achieved significantly better accuracy than the standard SVM. The overall accuracy achieved was 73% and 77% for the standard SVM and the one enhanced with Auto-WEKA. The highest accuracy (94%) was achieved for the “sadness” emotion and the lowest accuracy (62%) for the “happiness”. The confusion matrix in Figure 6 shows that the “anger” is mixed with the “happiness” and vice versa. Also the “boredom” is mixed with “disgust”.

SVM trained and tested with 10 fold cross-validation achieves better accuracy compared to LOSO, 86% and 77% accuracy respectively. The reason for this is that with the 10 fold cross-validation the training and the testing data usually contain data samples of the same speaker. On the other hand, the LOSO technique gives better estimate if the system for speech emotion recognition is supposed to work in an environment where it does not have any information about the speaker. A hybrid approach that includes a calibration phase at the beginning of the usage of the system (for example asking the user to record several data samples) is considered for future work.

The recognition accuracy achieved by the SVM trained and tested with 10 fold cross-validation is similar to the one achieved by human (manual recognition); in both cases the accuracy is 86%. Even more, low energy emotions (boredom, sadness and disgust) are better recognized by ML compared to the human. This means that for these emotions, the human ear requires additional information, which can be easily extracted using a software tools.

Auto-WEKA is state of the art approach for parameter optimization in ML. This is the first time that it is used in the field of speech analysis especially in speech emotion recognition where there is not yet gold standard approach that is widely accepted by the research community.

For future work we plan to test our approach on other languages and to provide language independent model for emotion recognition. This is possible since the emotions that we are trying to recognize are proven to be

universal and the features that we are using are language-independent. The ultimate goal would be real time language independent emotion recognition service that can be used as a part of a human affect tracking system which promotes wellbeing.

References

[1] D. G. Myers. Theories of Emotion. Psychology: Seventh Edition. New York NY: Worth Publishers. 2004.

[2] V. Perez-Rosas, R. Mihalcea. Sentiment Analysis of Online Spoken Reviews. Interspeech, 2013.

[3] A. Halder, A. Konar, R. Mandal, A. Chakraborty. General and Interval Type-2 Fuzzy Face-Space Approach to Emotion Recognition. IEEE Transactions on Systems, Man, and Cybernetics,43 (3), 2013.

[4] R. Horlings, D. Datcu, L. J. M. Rothkrantz. Emotion recognition using brain activity. Proceeding CompSysTech '08 Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing, 2008.

[5] A. Metallinou, S. Lee, S. Narayanan. Audio-Visual Emotion Recognition Using Gaussian Mixture Models for Face and Voice. Multimedia. 2008. ISM 2008. IEEE International Symposium on Multimedia, 2008.

[6] P. Ekman. Emotions in the Human Faces. 1982.

[7] James A. Russell. A circumplex model of affect. 1980.

[8] P. N. Juslin, K. R. Scherer. Vocal expression of affect. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.). The new handbook of methods in nonverbal behavior research, pp. 65-135, 2004.

[9] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. Speech Communication 40: 227–256. 2003

[10] M. E. Mena. Emotion Recognition From Speech Signals, 2012.

[11] D. Ververidis, C. Kotropoulos. A review of emotional speech databases. In: PCI 2003. 9th Panhellenic Conference on Informatics., pp. 560–574, 2003.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss. A Database of German Emotional Speech. 2005. In: Proc. Interspeech. pp. 1517–1520.

[13] F. Eyben, M. Wöllmer, B. Schuller. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. 2010.

[14] F. Eyben, F. Weninger, M. Wollmer, Bjorn Schuller. openSmile Documentation. Version 2.0.0., 2013.

[15] G. Isabellem E. André. "An Introduction to Variable and Feature Selection". JMLR, 2003.

[16] H. Deng, G. Runger, E. Tuv. Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN2011). 2011

[17] I. Kononenko, E. Simec, M. Robnik-Sikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. Applied Intelligence, Forthcoming.

[18] J. Demšar, B. Zupan. Orange: From experimental machine learning to interactive data mining. White Paper (http://www.ailab.si/orange). Faculty of Computer and Information Science. University of Ljubljana.

[19] D. Aha, D. Kibler. Instance-based learning algorithms. 1991, Machine Learning. 6:37-66.

[20] N. Cristianini, J. Shawe-Taylor. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.

[21] R. Stuart, N. Peter. Artificial Intelligence: A Modern Approach. Second Edition, Prentice Hall.

[22] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn, Morgan Kaufmann, 2005.

[23] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for Hyper-Parameter Optimization. In Proc. Of NIPS-11, 2011.

[24] F. Hutter, H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In Proc. of LION-5, pages 507–523, 2011.

[25] Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classifiaction Algorithms. Chris Thornton, Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. In Proc. of KDD 2013, 2013.