Emotion Recognition from Children's Speech

Yasmine Maximos[1] and David Suendermann-Oeft[2]

[1] German University in Cairo, Egypt, yasmine.maximos@student.guc.edu.eg
[2] DHBW, Stuttgart, Germany, david@suendermann.com

Abstract. This paper focuses on the recognition of emotions in speech through classification of the FAU Aibo Corpus for the two-class task (negative vs. idle) previously introduced at the Speaker Emotion Challenge at Interspeech 2009. The corpus contains natural, emotional German speech recordings of 51 children. In order to improve emotion recognition, different approches to training predictive models on labeled feature vectors have been examined. The best result was achieved by classifying feature vectors of 110 features (consisting of LLDs, their deltas and combined with functionals) after dimension reduction and SMOTE filtering using the parameter optimized sequential minimal optimization (SMO) algorithm with an unweighted average recall (UAR) of 69.39% and an accuracy of 71.5%.

1 Introduction

Speech is the most natural form of communication between human beings. Through speech individuals can express their feelings, and their emotional state can be detected by others. With the ever-growing presence of spoken dialog technology (e.g. in phone hotlines or conversational agents on smartphones), automatic emotion recognition can be a handy tool in order to reduce the gap between humans and computers [5]. Imagine Siri, the conversational agent on Apple's iOS, detecting one's emotion and responding accordingly. There are several emotional cues (known as features) carried within a speech signal. These features when extracted and grouped, form feature vectors that are later on modeled with different pattern recognition classifiers. Working on the FAU Aibo corpus [6] allowed us to simulate a real life scenario as it includes natural emotional speech and well defined testing and training partitions. To encourage research teams to work on classifying emotions on this corpus, the Interspeech Speaker Emotion Challenge 2009 was the first platform to compare performance of competing systems publicly providing baseline results [5]. Hence, there already exist a number of publications covering this topic, though to the best of our knowledge all of them including the open-performance subchallenge winners from Interspeech [2] only focused on improving UAR (the primary measure) rather than accuracy (secondary measure).

2 System Description

2.1 Basic Setup

Feature vectors were extracted from the FAU Aibo Corpus using the open-source OpenSMILE toolkit [3] after few alterations to the Interspeech 2009 Emotion Challenge configuration file that comes along with OpenSMILE. For each audio file, one feature vector was extracted (for 16 low-level descriptors (LLDs) and their deltas, 12 functionals were computed resulting in 384 features altogether). Details of the features are shown in Table 1. All classification experiments were done using the open-source WEKA toolkit in order to allow reproducibility of results [7].

LLDs	Functionals (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel. position, range
(Δ) MFCC 1-12	linear regression: offset, slope, MSE

Table 1: Features used in this study.

In the beginning, classification tests were executed using all available classifiers provided by WEKA. The top eight performing classifiers RBFNetwork, PART, MultiLayerPerceptron, SimpleLogistic, Logistic, NaiveBayes, ConjunctiveRule and SMO respectively were then chosen for further tests. All experiments have been executed using the test, development, and training sets provided at the Speaker Emotion Challenge. For comparabilty of results with other publications, the unweighted average recall (UAR) was used as primary performance metric followed by the accuracy, calculated as follows:

$$\text{UAR} = \frac{\{\text{recall (I)}\} + \{\text{recall (N)}\}}{2} = \frac{\frac{tp}{tp+fn} + \frac{tn}{tn+fp}}{2}$$

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

3 Experiments

3.1 Corpus Description

The FAU Aibo Corpus includes 18,216 WAV files along with their labels (based on majority vote of linguistics students) for both two and five class emotion tasks. Segmentations and transliterations are also provided. The corpus consists

of emotional German speech of 51 children aged between 10 and 13 from two different schools located in Erlangen, a Montessori school (8 male, 17 female) whose data was used for testing and a high school (13 male, 13 female) for training. Interacting with Sony's Aibo robot, the children were led to believe that the robot will obey to all their commands, however Aibo was controlled by an operator that made it disobedient sometimes which triggered the children's emotion both positively and negatively. WEKA's majority vote classifier (known as ZeroR) always detecting the most frequent class produced a UAR of 50% and an accuracy of 70.1%.

3.2 Experiment I - Filtering

The Synthetic Minority Oversampling TEchnique (SMOTE) is a supervised instance based filter implemented in WEKA. It resamples the class with the least number of instances so that almost all classes would be balanced. It was applied to the top eight performing classifiers in order to improve their results by oversampling the negative class and obtain a more adequate balance between classes. SMOTE's effect on the performance is shown in Figure 1.

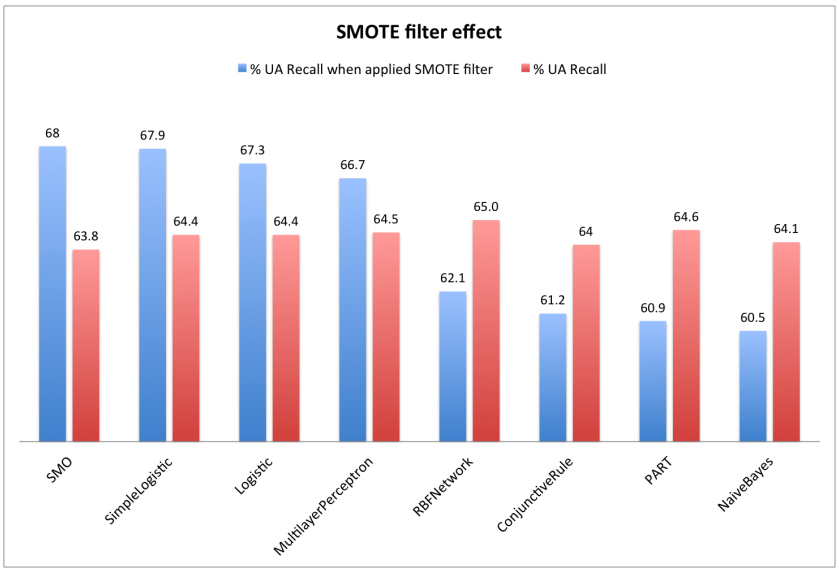

Fig. 1: Classification results before and after applying SMOTE.

3.3 Experiment II - Feature Selection

Dimensionality reduction by way of feature selection (also known as attribute selection) reduces dimensionality by discarding features according to certain criteria. In our study, we ordered features by their information gain and only kept

the top n of them. Classification done in the reduced space can be more accurate than in the original space. Attribute selection was applied twice to the top eight performing classifiers over the interval $n \in [5, 200]$ with a step size of 5, once with the SMOTE filter applied and the other without it, in order to show the impact of SMOTE along with feature selection on the performance. The best results were all produced through attribute selection with the application of SMOTE filter, shown in Table 2, where the SMO classifier was clearly ahead of the others.

Classifier	Number of Features	UAR (%)	Accuracy (%)
RBFNetwork	25	65.80	61.79
NaiveBayes	115	66.34	64.45
SimpleLogistic	110	69.08	70.87
Logistic	105	68.94	70.81
MultiLayerPerceptron	200	68.13	68.89
SMO	**110**	**69.10**	**71.54**
PART	25	65.49	68.19
ConjunctiveRule	15	65.09	69.83

Table 2: Results after feature selection.

3.4 Experiment III - Parameter Optimization

In order to improve the classification results furthermore, parameter optimization for both the SMO classifier and SMOTE filter was carried out. The optimization of SMOTE's NearestNeighbors K parameter was done over the interval [1,100] with a step size of 1, applied to the default SMO classifier. SMOTE's parameter optimization achieved a result of 69.28% UAR and an accuracy of 71.47% with $K = 89$. Finding the optimal complexity parameter value for the SMO classifier was done over the interval of [1,30] with a step size of 1. However, it was very clear that the optimal value lies in the interval [0,2]. In Figure 2 SMOTE's optimal parameter result is combined with SMO's complexity value within the interval of]0,2] with a step size of 0.01 to find the optimal parameter value (1.32).

4 Conclusion and Outlook

The sequential minimal optimization (SMO) was the top performing classifier. Its original performance was 63.8% UAR but after filtering, attribute selection and parameter optimization, an absolute improvement of 5.6% was achieved (69.39% UAR and 71.5% accuracy). The obtained results outperform the Interspeech 2009 Emotion Challenge baseline results of 67.7% UAR and 65.5% accuracy and

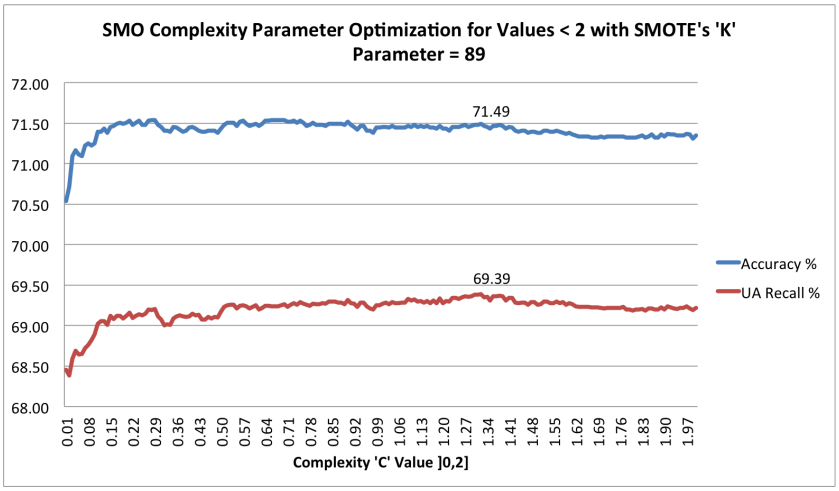

Fig. 2: Optimizing SMO complexity parameter.

are close to the ones achieved by Dumouchel, winner of the open-performance sub-challenge. He and his team achieved a 70.29% UAR and a 68.68% accuracy that was obtained by linear regression fusion of 3 systems.

This study showcases that recognizing real-life non-prototypical emotions is very difficult, however the problem remains essential to solve in order to improve human-computer interactive applications. To improve performance of emotion recognition, it will be essential to investigate the appropriateness of the underlying features. E.g. the use of prosodic features or ones exploring the presence of the Lombard effect [1] could be useful enhancements. On the other hand, further research into more advanced classification techniques could also be beneficial. Examples include maximum entropy, deep neural networks, conditional random fields or the taking of contextual information into account (as done by hidden Markov models and the like). Also more sophisticated dimensionality reduction techniques (such as linear discriminant analysis) as well as speaker adaptation and normalization techniques are worth looking at. It would also be interesting whether the inclusion of textual content can lead to improvements as suggested by [4].

References

1. H Bořil, P Fousek, D Sündermann, P Červa, and J Žďánskỳ. Lombard speech recognition: A comparative study. In *Proceedings of the 16th Czech-German Workshop, Prague*, pages 141–148, 2006.
2. Pierre Dumouchel, Najim Dehak, Yazid Attabi, Reda Dehak, and Narjes Boufaden. Cepstral and long-term features for emotion recognition. In *Proceedings of the Interspeech*, pages 344–347, 2009.

3. Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the International Conference on Multimedia*, pages 1459–1462, 2010.
4. Andreas Jauch, Paul Jaehne, and David Suendermann. Using text classification to detect alcohol intoxication in speech. *In Proceedings of the 7th Workshop on Emotion and Computing at the 36th German Conference on Artificial Intelligence. 2013.*
5. *Björn Schuller, Stefan Steidl, and Anton Batliner. The Interspeech 2009 emotion challenge. In* Proceedings of the Interspeech*, pages 312–315, 2009.*
6. *Stefan Steidl.* Automatic classification of emotion-related user states in spontaneous children's speech. *University of Erlangen-Nuremberg, 2009.*
7. *Ian H Witten and Eibe Frank.* Data Mining: Practical machine learning tools and techniques. *Morgan Kaufmann, 2005.*